ВЫСШЕЕ
ОБРАЗОВАНИЕ

Б. Г. Миркин

НИУ
«ВЫСШАЯ ШКОЛА
ЭКОНОМИКИ»

БАЗОВЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебник и практикум
3-е издание

Курс с онлайн-оцениванием

УМО ВО
РЕКОМЕНДУЕТ

Юрайт
издательство

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Б. Г. Миркин

БАЗОВЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

УЧЕБНИК И ПРАКТИКУМ ДЛЯ ВУЗОВ

3-е издание, переработанное и дополненное

Рекомендовано Учебно-методическим отделом высшего образования в качестве учебника и практикума для студентов высших учебных заведений, обучающихся по экономическим, инженерно-техническим, естественнонаучным направлениям

Курс с практическими заданиями и дополнительными материалами доступен на образовательной платформе «Юрайт», а также в мобильном приложении «Юрайт.Библиотека»

Москва ▪ Юрайт ▪ 2024

УДК 51(075.8)
ББК 22.161я73
М63

Автор:

Миркин Борис Григорьевич — доктор технических наук, доцент, старший научный сотрудник.

Рецензенты:

Моттль В. В. — доктор технических наук, профессор Московского физико-технического института, ведущий научный сотрудник вычислительного центра РАН;

Алескеров Ф. Т. — доктор технических наук, профессор Московского физико-технического института; ведущий научный сотрудник вычислительного центра РАН.

Миркин, Б. Г.

М63 Базовые методы анализа данных : учебник и практикум для вузов / Б. Г. Миркин. — 3-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2024. — 297 с. — (Высшее образование). — Текст : непосредственный.

ISBN 978-5-534-19709-9

Анализ данных — предмет, порожденный компьютерной революцией, приведшей к накоплению огромного количества конкретных данных о совокупностях объектов, таких как страны или регионы, веб-сайты, работодатели и работники, товары и продавцы. В отличие от классической математической статистики анализ данных не пытается вывести свойства окружающего мира исходя из специально собранных данных, а ориентирован на отыскание каких-либо паттернов, закономерностей, структуры в имеющихся данных.

В данном учебнике, подготовленном на основе большого международного опыта исследований и преподавания, излагаются основные методы анализа данных, относящихся прежде всего к одному или двум изучаемым признакам. Подробно рассмотрены вопросы анализа и интерпретации связей между двумя количественными, двумя качественными, а также качественным и количественным признаками. Из многомерных методов рассмотрены наивный Бэйесовский классификатор и метод К-средних для кластерного анализа, включая «интеллектуальную» версию с автоматическим определением числа кластеров и их начального местоположения. Изложение ориентировано на людей, предпочитающих не формулы, а вычисления, и содержит большое количество иллюстративных примеров применения рассматриваемых понятий к анализу реальных данных.

Для студентов бакалавриата и магистратуры инженерно-технических специальностей, также может использоваться для самостоятельного изучения.

УДК 51(075.8)
ББК 22.161я73

ISBN 978-5-534-19709-9

Оглавление

Но ты, художник, твердо веруй
В начала и концы. Ты знай,
Где стерегут нас ад и рай.
Тебе дано бесстрастной мерой
Измерить всё, что видишь ты.
Твой взгляд — да будет тверд и ясен.
Сотри случайные черты —
И ты увидишь: мир прекрасен.

Александр Блок

Предисловие

Анализ данных как инженерная дисциплина имеет дело с такими данными, которые оказались в распоряжении исследователя более или менее случайно, не как результат целенаправленного изучения, а как результат чьих-то наблюдений или просто статистической сводки. Это могут быть, например, данные о социально-экономическом состоянии регионов России или стран Европы в таком-то году. Или это может быть совокупность сообщений, отправленных членами какой-либо социальной сети в течение определенного промежутка времени. В подобных ситуациях типичные вопросы таковы. Какой смысл можно извлечь из этих данных? Есть ли какая-нибудь структура в данных о рассматриваемом множестве объектов? Могут ли эти признаки помочь в прогнозировании каких-то других? Такая ситуация характерна скорее для путешественника, чем ученого. Ученый сидит за столом и применяет научный метод, т. е. получает воспроизводимые данные об окружающем мире и старается включить их в грандиозную научную модель этого мира. Путешественник же должен сориентироваться, как ему лучше себя вести здесь и сейчас.

Анализ данных в настоящее время не очень занят проблематикой изучения механизмов формирования данных. Всё, что нас интересует — это наличие в данных каких-либо общих паттернов. Если удастся такой паттерн обнаружить; если удастся потом убедиться, что он — не артефакт применения метода, а действительно существует в данных; если удастся понять, исходя из наших знаний о явлении, к которому относятся данные, возможную причину возникновения паттерна; и если, наконец, на этой основе удастся внести что-то новое в изучение или использование явления — вот тогда можно говорить о том, что анализ данных работает. Впрочем, вопросы использования результатов анализа данных обычно остаются вне поля зрения специалистов по анализу данных: считается достаточным, чтобы нашелся паттерн, проливающий некий новый свет на явление, к которому относятся данные, чтобы объявить об успешности проведенного анализа.

Согласно точке зрения, подробно описанной в более полном международном учебнике автора [32], имеется два основных способа анализа данных: суммаризация и коррелирование. Такое понимание исходит из того, что теоретическое знание выражается,

прежде всего, через понятия и утверждения об связи понятий. А понятия в данных выражаются признаками. Это приводит к необходимости различения двух базовых задач анализа данных: (а) формирование признаков и (б) исследование связей между признаками. Первая задача выражается через суммаризацию данных, а вторая — через коррелирование признаков. Суммаризация, как и английский оригинал, означает подытоживание, агрегирование, представление в сжатом виде. Коррелирование — это отыскание связей между различными признаками, описывающими объекты.

Оба термина понимаются в самом широком смысле. Так, к задачам агрегации/суммаризации относятся: вычисление среднего значения ряда чисел, количественная оценка уровня интеллекта школьников по результатам тестирования и школьным оценкам, выявление кластера школьников со сходными оценками. Коррелирование — это отыскание взаимосвязи между разными признаками (совокупностями) признаков в таблице данных, будь это в виде аналитических соотношений, связывающих признаки, или концептуальных утверждений. Пример первого — утверждение, что «вес мужчин в килограммах примерно равен их росту в сантиметрах минус 100». Примеры второго — утверждения, что «люди пожилого возраста в среднем проводят у телевизора в день на 2 часа больше времени, чем люди среднего возраста» и что «новорожденные дети начинают говорить раньше, если их матери ели много рыбы во время беременности». Задержимся на этом последнем примере.

Наблюдения показали следующую корреляцию: новорожденные дети более активны и восприимчивы у тех матерей, которые ели много рыбы во время беременности. Значит ли это, что именно рыбоедение дает эффект? Да, говорят одни: в рыбе много фосфора, а фосфор — строительный материал мозга. Нет, говорят другие: рыба тут ни при чем. Просто эти женщины — богатые, ведь рыба дорого стоит, особенно в пересчете на калории. А у богатых уход за ребенком лучше, вот он и более активен. Кто же прав? Полученная нами информация не позволяет прийти к однозначному выводу. Все, что анализ данных может дать — это паттерн, а для выяснения причины паттерна нужны дополнительные данные, в данном случае надо изучать приток фосфора в мозг новорожденного в процессе беременности (очень сложно!) и (или) уровни благосостояния рожениц (значительно проще).

Начальные темы данного курса раскрывают проблематику суммаризации и коррелирования на уровне одно- и двумерных распределений. Первая тема содержит некоторое количество наборов данных и типичных проблем их анализа. Случай одного признака рассмотрен в теме 2. Тема 3 трактует случай, когда в анализ включаются два признака. При этом отдельно проанализированы задачи коррелирования для ситуаций, в которых (а) оба признака ко-

личественные, или (б) оба признака — категоризованные, или (в) один — категоризованный, а другой — количественный. Во всех трех случаях идея коррелирования проводится, исходя из основной цели — улучшения предсказания значений одного признака по значениям другого. Почему-то эта довольно популярная идея не нашла своего отражения в существующих учебниках. Поэтому изложение даже таких довольно традиционных тем как линейная регрессия (ситуация (а)) и табличная регрессия (ситуация (в)) получается довольно свежим и прагматически ориентированным[1]. Что касается ситуации (б) категоризованных признаков, то здесь и вовсе разработана нетрадиционная идея. За счет применения так называемых индексов Кетле удается представить коэффициент хи-квадрат, введенный К. Пирсоном для проверки гипотезы о статистической независимости категоризованных признаков, как меру их корреляции, и на этой основе визуализировать структуру связи между значениями признаков. В других учебниках читателя специально предупреждают: величина хи-квадрат не характеризует уровень связи и не может использоваться для ее оценки. Автор данного текста показывает, что это не так, что на самом деле коэффициент Пирсона имеет четкий операциональный смысл. В теме 4 обсуждается проблема коррелирования в многомерных данных. Приводятся два наиболее популярных метода изучения связи входных и выходных признаков, один для категоризованного выходного признака (наивный Бейесов[2] классификатор), второй для количественных выходных признаков (искусственные нейронные сети). Тема 5 посвящена методам суммаризации, как количественной (метод главных компонент), так и неколичественной (метод *K*-средних кластер-анализа). Выбор методов определяется не только популярностью, но и наличием некоторого внутреннего единства, определяемого «сквозным» использованием принципа наименьших квадратов. Всё изложение иллюстрируется на примерах конкретных данных из вводной темы 1. Следует специально отметить, что эти примеры играют важную роль в разъяснении материала книги. Их не следует пропускать. Они не только иллюстрируют описанные методы, но часто

---

[1] Здесь хочется сослаться на мнение рецензента исходной англоязычной версии учебника (Mirkin 2011): «Выделю только одно из многих успешных мест учебника: я сомневаюсь, что читатель когда-либо снова встретит такое детальное и превосходное описание корреляционных понятий», Computing Reviews of ACM, June 2011.

[2] Бейес (*Bayes*, 1702—1761) — английский «непрофессиональный» математик, чья работа стала известна после его смерти. Написание «Бейес» ближе к английскому произношению фамилии, «Бейиз», чем укоренившаяся в России форма «Байес». Мы предпочитаем это более корректное написание, имея в виду, что читатели данного текста — люди международных контактов, в которых произношение «Байес» не совсем уместно, так как воспроизводит произношение английского слова «bias», означающего «предвзятость».

служат площадкой для введения некоторых понятий анализа данных, которые обсуждаются только в них.

Основные особенности учебника заключаются в следующем.

(а) Четкое отделение дисциплины «Анализ данных» от дисциплины «Машинное обучение». Машинное обучение включает в себя все машинные алгоритмы, основанные на использовании данных, делая упор на использование значений одних признаков для предсказания значений других признаков. Анализ данных ограничивается исключительно методами, направленными на обогащение теоретических представлений о явлении/процессе, к которому относятся данные — грубо говоря, на производство знаний. При этом возникает ряд специфических для этих подходов моментов. Например, в машинном обучении в качестве истины признаются сведения, полученные из модели, тогда как в анализе данных во главу угла ставятся именно данные. Другой пример: в машинном обучении валидность метода определяется по тому, насколько аккуратны предсказания на «отложенных», не участвовавших в процессе обучения, данных. В анализе данных во главу угла ставится интерпретируемость знания, добытого с помощью метода.

(б) Крен в сторону методов «суммаризации» по сравнению с методами «коррелирования». Разработанные к настоящему времени методы коррелирования в основном сводятся к тематике прогнозирования значений «целевых», «выходных» признаков по значениям «входных» признаков и поэтому занимают основное место в машинном обучении. Напротив, аспект суммаризации в машинном обучении представлен относительно слабо. Достаточно сказать, что основополагающий в анализе данных метод главных компонент до недавнего времени вообще не включался в основной корпус машинного обучения (см., например, [26]). Поэтому я счел возможным уделить суммаризации данных больше внимания. При этом оба излагаемых подхода, метод главных компонент для количественной и метод к-средних для неколичественной суммаризации, оказываются основанными на одной и той же модели факторизации матриц по критерию наименьших квадратов.

(в) Распределение основного изложения по трем относительно независимым линиям: «представление», «формулировки» и «вычисление». «Представление» не содержит математических формул. Оно на конкретных данных показывает задачу, метод ее решения, а также комментарии к результатам, когда это необходимо. Напротив, в «формулировке» сосредоточены все математические детали постановки задачи и метода ее решения. В «вычислении» объясняется, как провести вычисление, зачастую с приведением псевдокода. Для удобства псевдокод дается на языке вычислительной среды MATLAB, у которой имеется бесплатная версия Октав/GNU Octave

(octave.org)[1]. Такое представление материала связано с желанием автора оградить читателя-нематематика от менее интуитивных математических понятий, связанных с многомерными данными, таких как векторы и матрицы, и, тем более, их свойств. Каждый из этих трех «потоков»: «представление», «формулирование», «вычисление» — может изучаться относительно независимо от других, так что те читатели, которые избегают формулы, могут вообще обойтись без них.

Обратим внимание на то, что указанные три потока в некотором смысле соответствуют трем типичным ролям, необходимым для успеха инженерной группы. Одна роль — это общий взгляд на вещи, роль «визионера». Вторая — роль конструктора, который переводит общую картину в технически корректный проект. Третья — это роль наладчика, который может перевести проект в работающий прототип изделия. Читатель может выбрать ту роль, которая ему ближе, или даже совместить все три, как это нередко бывает в жизни.

(г) Многоуровневая структура самостоятельных заданий, предназначенных для активизации работы читателя. А именно, среди заданий выделяются:

— «рабочие примеры», которые просто иллюстрируют работу того или иного метода; в начале каждого из них на конкретном примере показывается, как провести расчет и интерпретацию решения, когда это уместно, а затем дается задание для «самостоятельной работы» — повторить то же на других данных;

— «задания», более сложные задачи, в которых имеется определенный неформальный элемент, например, необходимость создания нового множества данных (по определенному правилу) или же неформальный способ интерпретации;

— «проекты», еще более сложные проблемы, в какой-то мере имитирующие научные проекты и требующие проведения небольшого научного исследования;

— «вопросы», математические или вычислительные — они, как правило, снабжены ответами, либо в явном виде, либо в самой формулировке вопроса. Это не значит, что их не надо решать самостоятельно. Надо. Ответы приводятся лишь для проверки;

— «самостоятельные работы», задания, полностью аналогичные каким-то «заданиям» или «рабочим примерам» — в них речь идет о том, чтобы проделать те же вычисления, но на других данных.

Всего в курсе содержатся 32 рабочих примера, 15 заданий, 6 проектов, 64 вопроса, а также 36 самостоятельных работ.

---

[1] Желательно, чтобы читатель имел доступ к этой или подобной среде. Использование MATLAB в контексте рассматриваемых понятий и методов не требует программистских навыков. Азы работы на MATLAB объясняются в Приложении к данной книге.

Важно понимать, что выделение какого-то материала в самостоятельное задание или вопрос не обязательно свидетельствует о том, что данный материал является дополнительным и может быть освоен в другом месте учебника. Напротив, большая часть этих заданий содержит уникальный материал, который следует усваивать именно в рамках задания!

(д) Вводятся самые современные методы вычислительной науки, такие как бутстрэп (*bootstrap*) для оценки доверия к результатам и эволюционные, инспирированные природой, алгоритмы для оптимизации нелинейных критериев.

(е) Книгу предваряет освещение определенных моментов истории анализа данных (на примере биографий отцов-основателей), а заключает обсуждение связей между анализом данных, практикой его применения и другими разделами искусственного интеллекта — машинным обучением и классической вероятностной статистикой.

Кроме того, с учетом современной тенденции уделять и делу время, и потехе час, в учебнике представлено несколько картинок и десятки шуток из современного фольклора, с юмористической стороны иллюстрирующих обсуждаемые понятия. В конце каждой темы имеется небольшой раздел «Кстати говоря», в котором размещено некоторое количество анекдотов, связанных с содержанием темы.

Данный курс использует авторские курсы для студентов бакалавриата и магистратуры в Биркбек колледже Лондонского университета (2004—2010), для слушателей Школы Анализа Данных ШАД при Яндексе (2008—2010), а также для студентов бакалавриата и магистратуры факультета компьютерных наук Национального Исследовательского Университета Высшей школы экономики (2008—2023). В значительной мере его содержание следует моему более полному англоязычному учебнику [32], а также русскоязычному учебнику «Введение в анализ данных» [14]. По сравнению с последним, здесь значительно расширен материал, посвященный методам анализа многомерных данных.

Хотя основной текст написан так, чтобы его мог освоить человек, не изучавший высшую математику, некоторое знакомство с ней, конечно, полезно. Речь идет, прежде всего, об азах математического анализа (понятия функции, ее производной, точек минимума), теории вероятностей (частота и условная вероятность, функция плотности), алгебры матриц и векторов, а также теории множеств (понятия включения множеств и принадлежности элемента данному множеству).

Таким образом, в результате изучения представленного материала студент должен ***быть компетентным*** в понимании основных понятий и методов, связанных с анализом данных, а также умении их применять для анализа реальных данных с использованием вычислений на современных вычислительных устройствах. Более подробно компетенции описаны в аннотациях к отдельным темам.

Учебник ориентирован на использование в курсах анализа данных, математической статистики и машинного обучения в бакалавриате инженерных специальностей — прикладная математика, информатика, программная инженерия, а также и курсах количественных методов для не инженерных специальностей — экономика, социология, менеджмент, география, филология и пр. Для не инженерных специальностей учебник может быть рекомендован к использованию и в магистерских программах. Кроме того, данное пособие может быть использовано для самостоятельного изучения теми, кто по характеру своей деятельности хотел бы использовать данные и методы их анализа.

В заключение хочу выразить благодарность моим коллегам по работе в НИУ ВШЭ, сделавшими возможной и приятной работу над данным учебником, за внимание и поддержку. Речь идет, прежде всего, о Департаменте анализа данных и искусственного интеллекта (руководитель С. О. Кузнецов) и международном Центре анализа и выбора решений (руководитель Ф. Т. Алескеров). Конечно, все остающиеся ошибки — всецело на моей ответственности.

Тема 1
ЧТО ТАКОЕ АНАЛИЗ ДАННЫХ

В этой вводной главе рассказывается, что такое анализ данных и чем он отличается от математической статистики. Приводятся примеры задач анализа данных. Вводятся два основных типа задач анализа данных: суммаризация и коррелирование. Дается представление о сходных дисциплинах, возникших в связи с развитием вычислительной техники.

1.1. Понятие о таблице данных

Объектом анализа данных являются таблицы данных типа той, что представлена в табл. 1.1.

Таблица 1.1

Компании*

Компания	Доход, млрд руб.	Доля рынка, %	ОП: Число основных потребителей	Интернет (да/нет)	Сектор экономики
Авер	19.0	21.85	2	–	Химия
Ант	29.4	18.00	3	–	Химия
Астон	23.9	19.00	3	–	Металлургия
Бмарт	18.4	13.95	2	+	Химия
Брек	25.7	11.15	3	+	Металлургия
Бумо	12.1	8.45	2	+	Металлургия
Виж	23.9	15.10	4	+	Торговля
Вурд	27.2	29.00	5	+	Торговля

* Совокупность 8 компаний охарактеризована пятью разнотипными признаками. Имена компаний отражают основные группы производимой продукции (либо А, либо Б, либо В)

В этой таблице приводятся данные о 8 компаниях в разрезе следующих пяти признаков:

1) Доход — годовой доход в млрд рублей;
2) ДоляР — доля рынка, %;
3) ОП — количество основных потребителей;

4) Инт — есть ли ведение бизнеса по Интернету (+) или нет (–);

5) Сектор — превалирующая отрасль народного хозяйства: (a) Химия, (b) Металлургия, (c) Торговля.

В табл. 1.1 строки соответствуют компаниям, столбцы — признакам, так что элементами таблицы являются значения признаков для конкретных компаний. В общем случае такая таблица содержит *N* строк, соответствующих рассматриваемым *объектам*, *V* столбцов, соответствующих *признакам*, а в самой таблице находятся *значения* признаков. Эти-то значения и образуют то, что называется *данные*. Названия признаков и столбцов образуют то, что принято называть *метаданные*. Метаданные задают тот элемент знаний о реальном мире, который связан с таблицей данных. Это, прежде всего, названия и методы измерения признаков. В зависимости от прикладной области, признаки могут называться свойствами, переменными, атрибутами и даже состояниями. В зависимости от того, насколько структурировано множество значений в предметной области, признак может иметь разный тип шкалы. Чаще всего различают так называемые количественные, бинарные, номинальные и порядковые типы шкал. Эти типы шкал определяются следующим образом:

— *количественный* тип шкалы имеет место для признаков, для которых осмысленна операция усреднения их значений. В табл. 1.1 таковы признаки Доход, ДоляР и ОП. Иногда считают, что признаки с дискретными количественными значениями, такие как ОП, не допускают операцию усреднения. «Разве осмысленно говорить о 4,5 основных потребителях в группе Торговли в табл. 1.1?» Автор считает, что осмысленно, если допустить возможность умножения объектов. Если, например, строки Виж и Вурд в таблице соответствуют сотне объектов каждая, то среднее количество потребителей, согласно таблице, будет не 4.5, а 450 — вполне осмысленная величина;

— *бинарный* тип шкалы соответствует признакам, которые допускают только два значения, «Да» и «Нет»; таков признак Инт в табл. 1.1;

— *номинальный* тип шкалы характеризует признаки, допускающие несколько непересекающихся категорий, никак не связанные друг с другом,такие как Сектор в табл. 1.1;

— *порядковый* тип шкалы, как и номинальный, относится к признакам, допускающим несколько непересекающихся категорий. В отличие от номинального типа шкалы, значения (категории) здесь линейно упорядочены друг относительно друга. Таковы, например, признаки, характеризующие отношения людей к тем или иным событиям или персонажам: значения «очень нравится», «нравится», «безразлично», «не нравится» упорядочены очевидным образом. В отличие от других упомянутых типов шкал, математиче-

ская структура порядковой шкалы (конус) не допускает простого способа ортогонального проецирования; поэтому этот тип шкалы рассматривается отдельно от остальных. Обычно при совместном анализе с признаками других типов шкал значениям порядкового признака приписываются числовые ранги. Например, значению «очень нравится» могут сопоставить ранг 1, значению «нравится» — ранг 2, значению «безразлично» — ранг 3, значению «не нравится» — ранг 4, а могут — ранги 2, 1, 0, –1, соответственно. После этого признак обрабатывается как количественный.

Обычно бинарные и номинальные признаки рассматривают как неколичественные, т. е. «категоризованные». На самом деле их можно перевести в «количественный» формат, приписывая 1 ответу «Да» и 0 — ответу «Нет». Для бинарных признаков эта перекодировка тривиальна. Номинальный признак должен быть сначала «расщеплен» в систему бинарных признаков, соответствующих отдельным категориям. Например, признак «Сектор» в табл. 1.1 будет заменен на три бинарных признака, соответствующих его категориям (см. табл. 1.2):

— сектор Химии?
— сектор Металлургии?
— сектор Торговли?

Такие бинарные признаки часто называют «дамми» (от английского «dummy», т. е. чурбан).

Таблица 1.2

Данные о компаниях из табл. 1.1, преобразованные в количественный формат

Номер	Доход	ДоляР	ОП	Инт	Хим	Мета	Торг
1	19.0	21.85	2	0	1	0	0
2	29.4	18.00	3	0	1	0	0
3	23.9	19.00	3	0	0	1	0
4	18.4	13.95	2	1	1	0	0
5	25.7	11.15	3	1	0	1	0
6	12.1	8.45	2	1	0	1	0
7	23.9	15.10	4	1	0	0	1
8	27.2	29.00	5	1	0	0	1
Среднее	22.4	17.06	3.0	0.625	0.375	0.375	0.25

Названия объектов также относятся к метаданным и могут нести важную информацию, как, например, о соперничающих компаниях или лидерах общественного мнения. В табл. 1.1, например, первые буквы названий компаний соответствуют одному из трех главных секторов рынка (А, Б и В), в которых они наиболее представлены.

Взятие среднего значения у бинарного 1/0 признака вполне осмысленно: среднее значение бинарного признака, соответствующе-

го данной категории, есть не что иное, как доля категории в множестве объектов! Этот факт, а также ему подобные, приводят к тому, что в данном тексте 1/0 признак считается количественным[1].

Может показаться, что любой столбец в таблице данных можно рассматривать в качестве признака, лишь бы он был заполнен сопоставимыми значениями. Но это не так. Например, столбец 1 табл. 1.1, содержащий названия компаний, признаком не является. Почему? Потому что он не несет никакой информации об отношениях между объектами. То есть он позволяет отличать компании друг от друга — и всё! Между тем, количественный признак ОП информирует, что компания Виж имеет вдвое больше основных потребителей, чем компания Авер (4 и 2). Даже номинальный признак «Сектор» сообщает об отношениях между объектами информацией о своих категориях: компании Авер и Ант — в одном секторе экономики, а Астон — в другом.

1.2. Преобразование файла данных в таблицу данных

Не всякий файл данных является таблицей данных, потому что не всякий столбец таблицы имеет смысл признака. Рассмотрим, например, табл. 1.3.

Столбцы этой таблицы характеризуют те или иные аспекты городских метросистем. Попробуем разобраться, какие из столбцов соответствуют признакам, а какие — нет, и что делать с «не признаковыми» столбцами.

Из 9 столбцов табл. 1.3 три последних — обычные количественные признаки:

— пассажиров в год;
— длина;
— количество станций.

Предшествующий им столбец, Год открытия, тоже может рассматриваться как количественный признак, особенно если вычесть его из года, к которому относятся данные, 2020. Тогда этот столбец превратится в признак «Возраст», измеряемый количеством лет, прошедших с момента начала работы метросистемы.

Не вызывает особых сомнений и предыдущий столбец, Континент, отвечающий номинальному признаку с двумя непересекающимися значениями, «Азия» и «Европа». Возникает вопрос — почему этот признак не бинарный? Ведь у него всего два значения.

---

[1] При этом игнорируется другое важное свойство количественных шкал: возможность непрерывного изменения их значений, очевидным образом не применимое к бинарным шкалам.

Таблица 1.3[1]

Характеристики городских метросистем бывшего Советского Союза

№	Городская система метро	Город	Страна	Континент	Год открытия	Пассажиров, млн в год	Длина, км	Количество станций
1	Метро Баку	Баку	Азербайджан	Азия	1967	72.1	36.7	26
2	Метро Еревана	Ереван	Армения	Азия	1981	10.75	13.4	10
3	Метро Минска	Минск	Беларусь	Европа	1984	219.3	40.8	33
4	Метро Тбилиси	Тбилиси	Грузия	Азия	1966	69.8	27.1	23
5	Метро Ташкента	Ташкент	Узбекистан	Азия	1977	38.8	57.1	39
6	Метро Алма-Аты	Алма-Ата	Казахстан	Азия	2011	7.0	13.4	11
7	Метро Киева	Киев	Украина	Европа	1960	279.5	67.6	52
8	Метро Москвы	Москва	Россия	Европа	1935	1618.2	435.7	203
9	Метро Петербурга	С-Петербург	Россия	Европа	1955	495.0	124.8	64
10	Метро Н. Новгорода	Н. Новгород	Россия	Европа	1985	20.4	21.6	15
11	Метро Новосибирска	Новосибирск	Россия	Азия	1985	58.1	15.9	13
12	Метро Самары	Самара	Россия	Европа	1987	8.8	11.6	10
13	Метро Казани	Казань	Россия	Европа	2005	21.5	16.8	11

[1] По данным сайта https://en.wikipedia.org/wiki/List_of_metro_systems (дата посещения — 5 августа 2022 г.).

Потому что с этими значениями нельзя связать ответы «Да» и «Нет». Признаку «Континент» отвечают два бинарных признака, по количеству его категорий. Вот они:

— Это метро в Азии? Для Баку — Да, для Самары — Нет.

— Это метро в Европе? Для Баку — Нет, для Самары — Да.

Столь же очевидно, что первые 3 столбца — номер, Городская система метро, Город — признаками не являются. Каждый из них по-разному кодирует каждую строку, не выявляя никаких отношений между ними.

Остается столбец Страна. Он объединяет последние 6 метросистем как находящиеся в России и присваивает индивидуальную метку каждой из остальных 7 систем. Таким образом, он ведет себя наполовину как номинальный признак, а наполовину — как индивидуальная метка. Поэтому можно рекомендовать преобразование этого столбца в бинарный признак «Российская?», отвечающий на вопрос — российская ли это система или нет.

Проведенный анализ позволяет рекомендовать преобразование наблюденной табл. 1.3 в табл. 1.4. Каждый значащий столбец табл. 1.4 задает признак. Поэтому табл. 1.4 — таблица данных, в отличие от табл. 1.3, представляющей собой просто файл данных или, как иногда говорят, «датасет».

Таблица 1.4

Таблица данных о городских метросистемах бывшего Советского Союза, полученная из табл. 1.3

№	В России?	Континент	Возраст	Пассажиров, млн в год	Длина, км	Количество станций
1	Нет	Азия	53	72.1	36.7	26
2	Нет	Азия	39	10.75	13.4	10
3	Нет	Европа	36	219.3	40.8	33
4	Нет	Азия	54	69.8	27.1	23
5	Нет	Азия	43	38.8	57.1	39
6	Нет	Азия	9	7.0	13.4	11
7	Нет	Европа	60	279.5	67.6	52
8	Да	Европа	85	1618.2	435.7	203
9	Да	Европа	65	495.0	124.8	64
10	Да	Европа	35	20.4	21.6	15
11	Да	Азия	35	58.1	15.9	13
12	Да	Европа	33	8.8	11.6	10
13	Да	Европа	15	21.5	16.8	11

Обратим внимание на то, что преобразование файла табл. 1.3 в таблицу данных (табл. 1.4) — не совсем механическая работа. Она потребовала от нас неформального решения — перевода столбца «Страна» в бинарный признак «В России?». Это обычное дело. Прежде чем анализировать файл данных, надо внимательно проверить каждый столбец на соответствие понятию «признак» и, при необходимости, либо его удалить, либо преобразовать таким образом, чтобы он стал признаком.

Самостоятельная работа 1.1

Преобразование файла «Титаник» в таблицу данных.

Задача: преобразовать в таблицу данных табл. 1.5 — фрагмент файла Titanic.csv[1], содержащего сведения о пассажирах печально-знаменитого британского пассажирского парохода Титаник, затонувшего в 1912 г. в результате столкновения с айсбергом.

Столбец «Выжил(а)?» информирует о том, удалось ли спастись пассажиру; «Класс» информирует о классе каюты, занимаемой пассажиром. Столбцы «Имя», «Пол», «Возраст» самоочевидны. Разве что стоит отметить, что возраст пассажира № 6 неизвестен. В столбце «БрСесСуп» стоит количество братьев, сестер, супругов данного пассажира на борту вместе взятых, а в столбце РодД размещается суммарное число детей и родителей пассажира на корабле. Столбец «Пункт посадки» содержит названия городов, в которых пароход принимал пассажиров (С — Саутгемптон (Southampton, Англия), Ш — Шербур (Cherbourg, Франция), К — Квинстаун (Qweenstown, Ирландия)).

Предлагаемое решение. Не будем трогать первые три столбца. Оставим нумерацию (первый столбец) как есть. Столбцы «Выжил(а)?» и «Класс» трактуем как количественные признаки. «Имя» индивидуально и признаком не является — этот столбец удаляем. «Пол» будем трактовать как номинальный признак с двумя значениями и в соответствии с вышеописанной процедурой квантизации заменим его двумя бинарными признаками «Ж» и «М» соответствующими женскому и мужскому полу. Признак «Возраст» оставим как есть — вопрос о заполнении пробела у пассажира № 6 рассмотрим позже. Столбцы «БрСесСуп» и «РодД» проводят непонятное российскому читателю различение между горизонтальным и вертикальным родством. Их объединяем в суммарный признак «Семья» — число членов семьи, путешествующих вместе с пассажиром. В цене билета отражаются мелкие различия, несущественные с современной точки зрения. Поэтому предлагается сделать ценник соответствующим классу, установив единую цену для каждого класса на уровне медианы значений цены в классе (напомним, что медиана ряда чисел — это значение его середины после упорядочения значений в порядке возрастания).

---

[1] URL: https://github.com/datasciencedojo/datasets/blob/master/titanic.csv, дата загрузки 4 июля 2022 г.

Таблица 1.5

Фрагмент файла Titanic.csv

№	Выжил(а)?	Класс	Имя	Пол	Возраст	БрСесСуп	РодД	Цена	Пункт посадки*
1	0	3	Braund, Mr. Owen Harris	М	22	1	0	7.25	С
2	1	1	Cumings, Mrs. John Bradley	Ж	38	1	0	71.28	Ш
3	1	3	Heikkinen, Miss. Laina	Ж	26	0	0	7.92	С
4	1	1	Futrelle, Mrs. Jacques Heath	Ж	35	1	0	53.1	С
5	0	3	Allen, Mr. William Henry	М	35	0	0	8.05	С
6	0	3	Moran, Mr. James	М	—	0	0	8.46	К
7	0	1	McCarthy, Mr. Timothy J	М	54	0	0	51.86	С
8	0	3	Palsson, Mr. Gosta Leonard	М	2	3	1	21.08	С
9	1	3	Johnson, Mrs. Oscar W	Ж	27	0	2	11.13	С
10	1	2	Nasser, Mrs. Nicholas	Ж	14	1	0	30.07	Ш
11	1	3	Sandstrom, Miss. Marguerite	Ж	4	1	1	16.70	С

Для 3-го класса упорядоченный ряд цен — 7.25, 7.92, 8.05, 8.46, 11.13, 16.70, 21.08. Середина ряда 8.46, после округления становится 8. Для 2-го класса — в ряду всего одно число, 30.07. После округления получаем 30. Упорядоченный ряд цен класса 1 — 51.86, 53.1, 71.28. Его середина 53.1, после округления становится 53. Пункт посадки — номинальный признак с тремя значениями; квантифицируем его той же процедурой — сопоставим каждой категории бинарный признак. Получим бинарные признаки «Посадка в С?», «Посадка в Ш?», «Посадка в К?». Окончательная таблица данных — в табл. 1.6. Обратим внимание, что при ее создании использовались неформальные соображения, прежде всего связанные с созданием признака «Количество членов семьи (Семья)».

Для решения поставленной задачи можно использовать материал очерка С. Мухия [46]. Ниже — решение, предлагаемое автором.

Для заполнения пропущенного значения возраста на объекте № 6 используем *метод ближних соседей*. Согласно этому методу, отыскиваются ближайшие соседи объекта № 6 в пространстве признаков, значения которых на объекте № 6 известны. В качестве расстояния используется квадрат Евклидова расстояния между объектами (определение и примеры см. в подпараграфе 5.2.1). В качестве априорного ограничения может выступать количество ближайших объектов или уровень расстояния. В данном случае ближайший к № 6 объект, объект № 5, находится от него на расстоянии 2, отличаясь от цели только признаками «Посадка в С» и «Посадка в К». Следующий ближайший объект, № 1, дополнительно отличается признаком «Семья». Расстояние между № 6 и любым другим объектом табл. 1.6 превышает 3. Если принять 3 в качестве границы понятия «ближнее соседство», то объект № 6 имеет только двух ближних соседей, объекты № 5 и № 1 (см. табл. 1.7).

Таблица 1.6

Таблица данных, полученная из файла сведений о пассажирах Титаника в табл. 1.5

№	Выжил(а)?	Класс	Ж?	М?	Возраст	Семья	Цена	Посадка в С	Посадка в Ш	Посадка в К
1	0	3	0	1	22	1	8	1	0	0
2	1	1	1	0	38	1	53	0	1	0
3	1	3	1	0	26	0	8	1	0	0
4	1	1	1	0	35	1	53	1	0	0
5	0	3	0	1	35	0	8	1	0	0
6	0	3	0	1	28	0	8	0	0	1
7	0	1	0	1	54	0	53	1	0	0

Окончание табл. 1.5

№	Выжил(а)?	Класс	Ж?	М?	Возраст	Семья	Цена	Посадка в С	Посадка в Ш	Посадка в К
8	0	3	0	1	2	4	8	1	0	0
9	1	3	1	0	27	2	8	1	0	0
10	1	2	1	0	14	1	30	0	1	0
11	1	3	1	0	4	2	8	1	0	0

Таблица 1.7

Объект № 6 табл. 1.6 и его ближние соседи

№	Выжил(а)?	Класс	Ж?	М?	Возраст	Семья	Цена	Посадка в С	Посадка в Ш	Посадка в К
1	0	3	0	1	22	**1**	8	**1**	0	**0**
5	0	3	0	1	35	0	8	**1**	0	**0**
6	0	3	0	1		0	8	0	0	1

Примечание. Жирным шрифтом выделены отличающиеся значения.

Значения возраста у ближних соседей — 22 и 35. Полусумма, 28.5 (после округления 28), представляет собой одновременно и среднее значение, и медиану этого ряда. Это число и использовано для заполнения пробела.

1.3. Иллюстративные проблемы анализа данных

Рассмотрим примеры данных и задач их анализа. Чтобы проиллюстрировать мысль, что анализ данных может применяться не только к большим множествам данных, но и малым, первый пример взят намеренно очень небольшим. Это также полезно с точки зрения того, что в данном случае данные можно обозреть как они есть, просто глядя на таблицу.

Пример 1

Компании

Табл. 1.8 — копия табл. 1.2, рассматривавшейся в параграфе 1.1. Хотя сами данные носят чисто иллюстративный характер, следующие вычислительные проблемы довольно типичны в анализе данных.

Таблица 1.8

Данные о компаниях из табл. 1.1, преобразованные в количественный формат

Номер	Доход	ДоляР	ОП	Инт	Хим	Мета	Торг
1	19.0	21.85	2	0	1	0	0
2	29.4	18.00	3	0	1	0	0
3	23.9	19.00	3	0	0	1	0
4	18.4	13.95	2	1	1	0	0
5	25.7	11.15	3	1	0	1	0
6	12.1	8.45	2	1	0	1	0
7	23.9	15.10	4	1	0	0	1
8	27.2	29.00	5	1	0	0	1
Среднее	22.4	17.06	3.0	0.625	0.375	0.375	0.25

— **Однородность.** Есть ли в данных аномально большие или аномально малые компании? Можно ли утверждать, что в данных механически объединены разные типы компаний?

— **Визуализация.** Как отобразить компании на экране так, чтобы расстояния между их позициями отражали сходство между ними: чем более похожи компании, тем ближе позиции?

— **Роль признаков.** Если кластеризовать компании в группы по степени сходства, будут ли кластеры соответствовать основным группам продукции, А, Б и В? И если да, то какие признаки окажутся наиболее весомыми?

— **Правила.** Можно ли вывести какие-либо точные правила для атрибуции вида продукции, исходя из данных признаков? (Эти правила затем можно применить и для компаний, не вошедших в данную таблицу (прогноз)).

— **Связь между аспектами.** Можно ли обнаружить какую-либо взаимосвязь между структурными признаками компаний (три признака справа в таблице) и признаками рыночной активности (доход и доля рынка)?

Пример 2

Ирисы

Ирисы — пожалуй, самая популярная таблица данных (см. табл. 1.9).

Таблица 1.9

Ирисы (в строках — информация о 150 экземпляров ириса, измеренных по 4 признакам в разрезе трех видов)

#	I *Iris setosa*	II *Iris versicolor*	III *Iris virginica*
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
1	5.1 3.5 1.4 0.3	6.4 3.2 4.5 1.5	6.3 3.3 6.0 2.5
2	4.4 3.2 1.3 0.2	5.5 2.4 3.8 1.1	6.7 3.3 5.7 2.1
3	4.4 3.0 1.3 0.2	5.7 2.9 4.2 1.3	7.2 3.6 6.1 2.5
4	5.0 3.5 1.6 0.6	5.7 3.0 4.2 1.2	7.7 3.8 6.7 2.2
5	5.1 3.8 1.6 0.2	5.6 2.9 3.6 1.3	7.2 3.0 5.8 1.6
6	4.9 3.1 1.5 0.2	7.0 3.2 4.7 1.4	7.4 2.8 6.1 1.9

Окончание табл. 1.9

#	I *Iris setosa*	II *Iris versicolor*	III *Iris virginica*
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
7	5.0 3.2 1.2 0.2	6.8 2.8 4.8 1.4	7.6 3.0 6.6 2.1
8	4.6 3.2 1.4 0.2	6.1 2.8 4.7 1.2	7.7 2.8 6.7 2.0
9	5.0 3.3 1.4 0.2	4.9 2.4 3.3 1.0	6.2 3.4 5.4 2.3
10	4.8 3.4 1.9 0.2	5.8 2.7 3.9 1.2	7.7 3.0 6.1 2.3
11	4.8 3.0 1.4 0.1	5.8 2.6 4.0 1.2	6.8 3.0 5.5 2.1
12	5.0 3.5 1.3 0.3	5.5 2.4 3.7 1.0	6.4 2.7 5.3 1.9
13	5.1 3.3 1.7 0.5	6.7 3.0 5.0 1.7	5.7 2.5 5.0 2.0
14	5.0 3.4 1.5 0.2	5.7 2.8 4.1 1.3	6.9 3.1 5.1 2.3
15	5.1 3.8 1.9 0.4	6.7 3.1 4.4 1.4	5.9 3.0 5.1 1.8
16	4.9 3.0 1.4 0.2	5.5 2.3 4.0 1.3	6.3 3.4 5.6 2.4
17	5.3 3.7 1.5 0.2	5.1 2.5 3.0 1.1	5.8 2.7 5.1 1.9
18	4.3 3.0 1.1 0.1	6.6 2.9 4.6 1.3	6.3 2.7 4.9 1.8
19	5.5 3.5 1.3 0.2	5.0 2.3 3.3 1.0	6.0 3.0 4.8 1.8
20	4.8 3.4 1.6 0.2	6.9 3.1 4.9 1.5	7.2 3.2 6.0 1.8
21	5.2 3.4 1.4 0.2	5.0 2.0 3.5 1.0	6.2 2.8 4.8 1.8
22	4.8 3.1 1.6 0.2	5.6 3.0 4.5 1.5	6.9 3.1 5.4 2.1
23	4.9 3.6 1.4 0.1	5.6 3.0 4.1 1.3	6.7 3.1 5.6 2.4
24	4.6 3.1 1.5 0.2	5.8 2.7 4.1 1.0	6.4 3.1 5.5 1.8
25	5.7 4.4 1.5 0.4	6.3 2.3 4.4 1.3	5.8 2.7 5.1 1.9
26	5.7 3.8 1.7 0.3	6.1 3.0 4.6 1.4	6.1 3.0 4.9 1.8
27	4.8 3.0 1.4 0.3	5.9 3.0 4.2 1.5	6.0 2.2 5.0 1.5
28	5.2 4.1 1.5 0.1	6.0 2.7 5.1 1.6	6.4 3.2 5.3 2.3
29	4.7 3.2 1.6 0.2	5.6 2.5 3.9 1.1	5.8 2.8 5.1 2.4
30	4.5 2.3 1.3 0.3	6.7 3.1 4.7 1.5	6.9 3.2 5.7 2.3
31	5.4 3.4 1.7 0.2	6.2 2.2 4.5 1.5	6.7 3.0 5.2 2.3
32	5.0 3.0 1.6 0.2	5.9 3.2 4.8 1.8	7.7 2.6 6.9 2.3
33	4.6 3.4 1.4 0.3	6.3 2.5 4.9 1.5	6.3 2.8 5.1 1.5
34	5.4 3.9 1.3 0.4	6.0 2.9 4.5 1.5	6.5 3.0 5.2 2.0
35	5.0 3.6 1.4 0.2	5.6 2.7 4.2 1.3	7.9 3.8 6.4 2.0
36	5.4 3.9 1.7 0.4	6.2 2.9 4.3 1.3	6.1 2.6 5.6 1.4
37	4.6 3.6 1.0 0.2	6.0 3.4 4.5 1.6	6.4 2.8 5.6 2.1
38	5.1 3.8 1.5 0.3	6.5 2.8 4.6 1.5	6.3 2.5 5.0 1.9
39	5.8 4.0 1.2 0.2	5.7 2.8 4.5 1.3	4.9 2.5 4.5 1.7
40	5.4 3.7 1.5 0.2	6.1 2.9 4.7 1.4	6.8 3.2 5.9 2.3
41	5.0 3.4 1.6 0.4	5.5 2.5 4.0 1.3	7.1 3.0 5.9 2.1
42	5.4 3.4 1.5 0.4	5.5 2.6 4.4 1.2	6.7 3.3 5.7 2.5
43	5.1 3.7 1.5 0.4	5.4 3.0 4.5 1.5	6.3 2.9 5.6 1.8
44	4.4 2.9 1.4 0.2	6.3 3.3 4.7 1.6	6.5 3.0 5.5 1.8
45	5.5 4.2 1.4 0.2	5.2 2.7 3.9 1.4	6.5 3.0 5.8 2.2
46	5.1 3.4 1.5 0.2	6.4 2.9 4.3 1.3	7.3 2.9 6.3 1.8
47	4.7 3.2 1.3 0.2	6.6 3.0 4.4 1.4	6.7 2.5 5.8 1.8
48	4.9 3.1 1.5 0.1	5.7 2.6 3.5 1.0	5.6 2.8 4.9 2.0
49	5.2 3.5 1.5 0.2	6.1 2.8 4.0 1.3	6.4 2.8 5.6 2.2
50	5.1 3.5 1.4 0.2	6.0 2.2 4.0 1.0	6.5 3.2 5.1 2.0

Она характеризует коллекцию, собранную ботаником Э. Андерсоном и использованную Р. Фишером в его основополагающей статье о дискриминантном анализе (1936). В ней представлены 150 цветков ириса, относящихся к трем видам: I. *Iris setosa* (диплоид), II. *Iris versicolor* (тетраплоид)

и III. *Iris virginica* (гексаплоид), по 50 экземпляров из каждого. Признаки табл. относятся к измерениям длины и ширины чашелистика (w1, w2), а также лепестков (w3, w4), см. рис. 1.1.

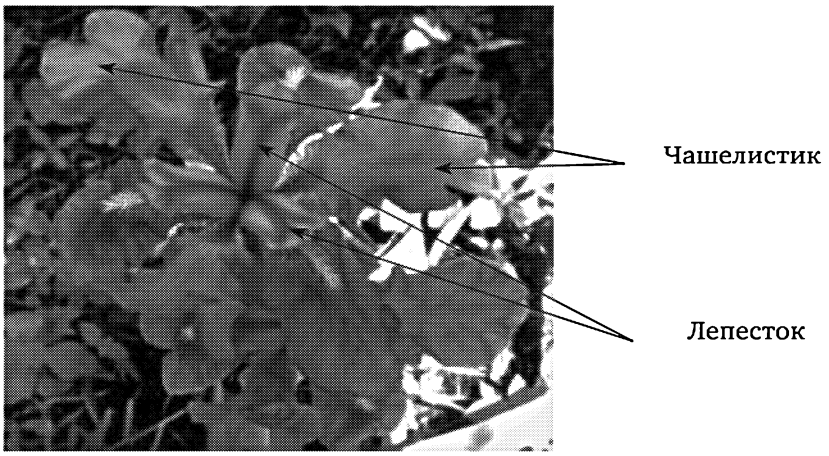


Рис. 1.1. **Чашелистик (*sepal*) и лепесток (*petal*) в цветке ириса**

Виды определяются генотипом, а признаки — фенотипом. Возникает вопрос, можно ли описать виды в терминах измеренных признаков. Хорошо известно, что вид I довольно легко отделяется от остальных, тогда как виды II и III перемешаны (возможно, из-за ошибок Андерсона в определении видов). С другой стороны, ботаники понимают, что виды можно неплохо отличить по площади лепестка, которая в определенной степени отражается площадью прямоугольника, т. е. произведением $w3 \cdot w4$.

Из проблем анализа данных, относящихся к этой таблице, укажем следующие:

- существует ли признак или пара признаков, распределение которых информативно с точки зрения описания трех видов ириса?
- визуализация данных на экране так, чтобы похожие объекты отображались близкими друг к другу точками;
- анализ связи между признаками, включая возможность предсказания, скажем, размеров лепестка по размерам чашелистика;
- возможность сведения всех признаков в единый непосредственно неизмеримый признак «размер цветка».

Пример 2

Компьютерные атаки

С учетом растущего значения компьютерных сетей возрастает опасность атак на них, приводящих к нарушению их функционирования. Простейший вид атаки — отказ от обслуживания [*denial of service*, DoS]. Такая атака причиняется командами, которые приводят к тому, что какой-либо ресурс — процессор, память, входное устройство — оказывается перегружен и не может обслуживать нормальные запросы. Две такие атаки в нижеследующей табл. 1.10 помечены как «apache2» и «smurf».

Таблица 1.10

Данные о компьютерных атаках

Pr	BySD	SH	SS	SE	RE	A	Pr	ByS	SH	SS	SE	RE	A
Tcp	62344	16	16	0	0.94	Ap	Tcp	287	14	14	0	0	no
Tcp	60884	17	17	0.06	0.88	Ap	Tcp	308	1	1	0	0	no
Tcp	59424	18	18	0.06	0.89	Ap	Tcp	284	5	5	0	0	no
Tcp	59424	19	19	0.05	0.89	Ap	Udp	105	2	2	0	0	no
Tcp	59424	20	20	0.05	0.9	Ap	Udp	105	2	2	0	0	no
Tcp	75484	21	21	0.05	0.9	Ap	Udp	105	2	2	0	0	no
Tcp	76944	22	22	0.05	0.91	Ap	Udp	105	2	2	0	0	no
Tcp	59424	23	23	0.04	0.91	Ap	Udp	105	2	2	0	0	no
Tcp	57964	24	24	0.04	0.92	Ap	Udp	44	3	8	0	0	no
Tcp	59424	25	25	0.04	0.92	Ap	Udp	44	6	11	0	0	no
Tcp	0	40	40	1	0	Ap	Udp	42	5	8	0	0	no
Tcp	0	41	41	1	0	Ap	Udp	105	2	2	0	0	no
Tcp	0	42	42	1	0	Ap	Udp	105	2	2	0	0	no
Tcp	0	43	43	1	0	Ap	Udp	42	2	3	0	0	no
Tcp	0	44	44	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	45	45	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	46	46	1	0	Ap	Udp	44	2	4	0	0	no

Продолжение табл. 1.10

Pr	BySD	SH	SS	SE	RE	A	Pr	ByS	SH	SS	SE	RE	A
Tcp	0	47	47	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	48	48	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	49	49	1	0	Ap	Udp	44	3	14	0	0	no
Tcp	0	40	40	0.62	0.35	Ap	Udp	105	1	1	0	0	no
Tcp	0	41	41	0.63	0.34	Ap	Udp	105	1	1	0	0	no
Tcp	0	42	42	0.64	0.33	Ap	Udp	45	3	6	0	0	no
Tcp	258	5	5	0	0	No	Udp	45	3	6	0	0	no
Tcp	316	13	14	0	0	No	Udp	105	1	1	0	0	no
Tcp	287	7	7	0	0	No	Udp	34	5	9	0	0	no
Tcp	380	3	3	0	0	No	Udp	105	1	1	0	0	no
Tcp	298	2	2	0	0	No	Udp	105	1	1	0	0	no
Tcp	285	10	10	0	0	No	Udp	105	1	1	0	0	no
Tcp	284	20	20	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	314	8	8	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	303	18	18	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	325	28	28	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	232	1	1	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	295	4	4	0	0	No	Tcp	0	482	1	0.05	.95	sa

Окончание табл. 1.10

Pr	BySD	SH	SS	SE	RE	A	Pr	ByS	SH	SS	SE	RE	A
Tcp	293	13	14	0	0	No	Tcp	0	482	1	0.06	.94	sa
Tcp	305	1	8	0	0	No	Tcp	0	482	1	0.06	.94	sa
Tcp	348	4	4	0	0	No	Tcp	0	482	1	0.06	.94	sa
Tcp	309	6	6	0	0	No	Tcp	0	483	1	0.06	.94	sa
Tcp	293	8	8	0	0	No	Tcp	0	510	1	0.04	.96	sa
Tcp	277	1	8	0	0	no	Icmp	1032	509	509	0	0	sm
Tcp	296	13	14	0	0	no	Icmp	1032	510	510	0	0	sm
Tcp	286	3	6	0	0	no	Icmp	1032	510	510	0	0	sm
Tcp	311	5	5	0	0	no	Icmp	1032	511	511	0	0	sm
Tcp	305	9	15	0	0	no	Icmp	1032	511	511	0	0	sm
Tcp	295	11	25	0	0	no	Icmp	1032	494	494	0	0	sm
Tcp	511	1	4	0	0	no	Icmp	1032	509	509	0	0	sm
Tcp	239	12	14	0	0	no	Icmp	1032	509	509	0	0	sm
Tcp	5	1	1	0	0	no	Icmp	1032	510	510	0	0	sm
Tcp	288	4	4	0	0	no	Icmp	1032	511	511	0	0	sm

Атака apache2 нацелена на популярный веб-сервер открытого пользоваиия, Apache HTTP Server, заставляя его посылать клиенту огромное количество «пустых» запросов и переполняя буферные каналы. Атака smurf посылает фальшивые ответные послания, «пинги», по различным адресам, заставляя соответствующие компьютеры отвечать «пингами» же по обратным адресам. Если главный адрес «подделан» нарушителем, возникает шквал «пингов», направляемых какому-либо серверу из группы компьютеров сети, переполняя его входные каналы. Такая атака может начинаться с «разведки», отыскивающей проблемы в сети. Популярное программное обеспечение (ПО) для проведения разведки называется SAINT [Security Administrator's Integrated Network Tool].

Таблица 1.10 — случайная выборка из данных, синтезированных в одной из лабораторий Массачусетского технологического института (Бостон, США, http://www.ll.mit.edu/mission/communications/ist/ corpora/ ideval/data/ intex.html). Признаки характеризуют пакет и его источник:

1) Pr — тип протокола, в данном случае один из трех: tcp, icmp, udp (этот признак — номинальный);

2) BySD — число байтов в пакете;

3) SH — количество соединений источника за последние две секунды;

4) SS — число соединений с тем же сервером за последние две секунды;

5) SE — процент ошибочных соединений;

6) RE — процент соединений с отказом обслуживания;

7) A — тип атаки (ap — apache2, sa — saint, sm — smurf, и отсутствие атаки — no).

Из сотни объектов в табл. 1.10 первые 23 — атаки сервера apache2, пакеты 24—69 — нормальные, следующие одиннадцать, 80—90, соответствуют разведке SAINT, и последние десять, 91—100, — атаки smurf.

Примеры проблем анализа данных, которые можно исследовать с использованием табл. 1.10:

— найти признаки пакетов, которые можно использовать, чтобы определить, нормально функционирует сеть или же она атакована;

— выяснить, есть ли связь между используемым протоколом и типом атаки;

— визуализировать данные так, чтобы близость точек соответствовала похожести значений признаков соответствующих объектов.

Пример 4

Малые города английского побережья

В табл. 1.11 приведены иллюстративные данные о 45 малых городах юго-запада Англии. Для целей социального планирования имеет смысл выделить сравнительно небольшую их группу так, чтобы каждый попавший в нее город представлял весь «кластер» похожих на него городов. В таблице города упорядочены по числу жителей. Например, 21 самых малых городов имеют меньше 4000 жителей каждый. Число 4000 взято в качестве разделителя не случайно. Во-первых, оно круглое. Во-вторых, оно помечает значительный разрыв в более чем 1300 между Кингскерсвеллом (3672 жителя) и следующим по численности городом Луе (5022 жителя). Следующий большой разрыв — после Лискярда (7044), отделяю-

щего 9 городов среднего размера от двух групп больших городов, насчитывающих соответственно 6 и 9 городов соответственно. Разделитель между двумя последними группами — между Тавистоком (10 222) и Бодмином (12 553). Так мы получим три или четыре группы городков, которые можно использовать укрупненно при социальном мониторинге. А достаточно ли однородны эти группы с точки зрения других имеющихся признаков? В монографии [33] показано, что более однородны в этом множестве не 4, а 7 кластеров: большие города порядка 17—20 тыс. жителей, два кластера городков средних размеров (8—10 тыс. жителей), три кластера малых городков (порядка 5 тыс. жителей), а также кластер совсем небольших поселений (порядка 2.5 тыс. жителей). Каждый из трех кластеров маленьких городков выделяется наличием некоего объекта, отсутствующего в двух других кластерах — фермерский рынок в одном, больница в другом и спортцентр с плавательным бассейном в третьем. Наличие этих объектов предполагает соответствующую направленность образа жизни.

Таблица 1.11

Данные о малых городах юго-запада Англии по переписи 1991 г.

Город	Нас	Нш	Тер	Бол	Ба	Ун	Ав	Ст	Бас	По	Юр	Фр
Муллион	2040	1	0	0	2	0	1	0	0	1	0	0
Юж. Брент	2087	1	1	0	1	1	0	0	0	1	0	0
Сент-Жюст	2092	1	0	0	2	1	1	0	0	1	0	0
Сент-Колумб	2119	1	0	0	2	1	1	0	0	1	1	0
Нанпин	2230	2	1	0	0	0	0	0	0	2	0	0
Гуннислэйк	2236	2	1	0	1	0	1	0	0	3	0	0
Мевагисси	2272	1	1	0	1	0	0	0	0	1	0	0
Иплепен	2275	1	1	0	0	0	1	0	0	1	0	0
Алстон	2362	1	0	0	1	1	0	0	0	1	0	0
Лоствизел	2452	2	1	0	2	0	1	0	0	1	0	1
С. Кулом	2458	1	0	0	0	1	3	0	0	2	0	0
Падстоу	2460	1	0	0	3	0	0	0	0	1	1	0
Перранпорс	2611	1	1	0	1	1	2	0	0	2	0	0
Бугль	2695	2	0	0	0	0	1	0	0	2	0	0
Бакфастль	2786	2	1	0	1	2	2	0	1	1	1	1
Сент-Агнес	2899	1	1	0	2	1	1	0	0	2	0	0
Порслевен	3123	1	0	0	1	1	0	0	0	1	0	0
Каллингтон	3511	1	1	0	3	1	1	0	1	1	0	0
Хорабридж	3609	1	1	0	2	1	1	0	0	2	0	0
Эшбуртон	3660	1	0	1	2	1	2	0	1	1	1	0

Окончание табл. 1.11

Город	Нас	Нш	Тер	Бол	Ба	Ун	Ав	Ст	Бас	По	Юр	Фр
Кингскерс	3672	1	0	0	0	1	2	0	0	1	0	0
Луе	5022	1	1	0	2	1	1	0	1	3	1	0
Кингсбридж	5258	2	1	1	7	1	2	0	0	1	1	1
Вадбридж	5291	1	1	0	5	3	1	0	1	1	1	0
Дартмаут	5676	2	0	0	4	4	1	0	0	2	1	1
Лонсестон	6466	4	1	0	8	4	4	0	1	3	1	0
Тотнес	6929	2	1	1	7	2	1	0	1	4	0	1
Пенрин	7027	3	1	0	2	4	1	0	0	3	1	0
Хэйль	7034	4	0	1	2	2	2	0	0	2	1	0
Лискярд	7044	2	2	2	6	2	3	0	1	2	2	0
Торпойнт	8238	2	3	0	3	2	1	0	0	2	1	0
Хелстон	8505	3	1	1	7	2	3	0	1	1	1	1
Сент-Блэзи	8837	5	2	0	1	1	4	0	0	4	0	0
Айвибридж	9179	5	1	0	3	1	4	0	0	1	1	0
Сент-Ивс	10 092	4	3	0	7	2	2	0	0	4	1	0
Тависток	10 222	5	3	1	7	3	3	1	2	3	1	1
Бодмин	12 553	5	2	1	6	3	5	1	1	2	1	0
Салташ	14 139	4	2	1	4	2	3	1	1	3	1	0
Бриксхэм	15 865	7	3	1	5	5	3	0	2	5	1	0
Ньюкэй	17 390	4	4	1	12	5	4	0	1	5	1	0
Труро	18 966	9	3	1	19	4	5	2	2	7	1	1
Пензанс	19 709	10	4	1	12	7	5	1	1	7	2	0
Фалмаут	20 297	6	4	1	11	3	2	0	1	9	1	0
Сент-Остелл	21 622	7	4	2	14	6	4	3	1	8	1	1
Эббот-Ньютон	23 801	13	4	1	13	4	7	1	1	7	2	0

В табл. 1.11 приводятся данные по следующим 12 признакам:
1) Нас — число жителей;
2) Нш — число начальных школ;
3) Тер — число терапевтов[1];
4) Бол — число больниц;

[1] В Англии медицинское обслуживание населения организовано не в поликлиниках, а в виде сети «практикующих терапевтов».

5) Ба — число отделений банков;
6) Ун — число продуктовых универмагов;
7) Ав — число автозаправок;
8) Ст — число магазины строительных материалов;
9) Бас — число бассейны;
10) По — число почтамты;
11) Юр — число бесплатных юридических консультаций;
12) Фр — число фермерских рынков..

Таким образом, главная цель анализа таких данных — это формирование кластеров однородных городов. Также интересным могло бы явиться исследование на тему, не являются ли все представленные признаки городов просто поверхностнымит измерителями некоторого «внутреннего», не измеряемого непосредственно, фактора «уровень развития города»? Если ответ не совсем отрицательный, то следующий вопрос — нельзя ли сформировать шкалу измерения этого внутреннего фактора, например, связав ее с шкалами измерения признаков в данных. Определенный интерес, конечно, может представлять анализ распределения городов по численности населения, а также корреляции этого показателя как с сопрягающимися признаками, такими как число начальных школ Нш, так и признаками, не связанными с населением административно, такими, скажем, как число отделений банков.

1.4. Комментарии к истории науки о данных

Наука о данных — безусловно, детище двадцать первого века, конституированное размахом компьютеризации, интернета, методов обработки и анализа данных, включая тексты и «большие данные», машинного обучения и т. п.

Древние жители средиземноморья наукой о данных не занимались. Ясно, что работа по восстановлению земельных наделов после ежегодных обильных разливов Нила в древнем Египте, приведшая к развитию геометрии, как и работа по предсказанию планетарных событий, таких как затмение луны или солнца, в древнем Вавилоне, приведшая к развитию арифметики, не могла быть осуществлена без постоянной записи и анализа соответствующих данных. Но нам об этом ничего не известно, вероятно, из-за того, что такая работа считалась вспомогательной. Ранние письменные сведения о сборе и анализе данных можно найти в Библии: сначала Моисей проводит перепись евреев перед нашествием на Ханаан (Книга чисел, Глава 1), потом царь Давид проводит перепись в Израиле (Первая книга Паралипоменон, Глава 21). Любопытно, что первое мероприятие Бог, согласно Библии, всемерно одобрил, а второе — нет; более того, серьезно наказал царя Давида. Возможно, дело в декларированной цели этой второй переписи — «чтобы знать». Не за это ли первые люди были изгнаны из рая?

Пьер-Симон Лаплас (1749—1827, Франция) и **Карл Фридрих Гаусс** (1777—1855, Германия) — выдающиеся математики, помимо других замечательных математических результатов заложившие основу вероятностно-математического моделирования в анализе данных. Вывели нормальное (Гауссово) распределение и доказали его универсальность, включая несколько версий центральной предельной теоремы. Предложили методы наименьших квадратов и наименьших модулей для оценки правильных значений. Лаплас установил также несколько общих фактов относительно так называемого Байесова (правильнее — Бейесова) подхода в статистике.

Настоящее развитие наука о данных получает в процессе становления капитализма. В XVI—XVII вв. города-республики Северной Италии (Флоренция, Венеция и др.) приходят к необходимости измерения основных параметров своего хозяйства — народонаселения, производства, торговли.

Адольф Кетле (*Lambert Adolphe Jacques Quételet*, 1796—1874) — бельгийский астроном и статистик, установивший, что разные виды массового социального поведения как, например, различные виды преступлений, обладают статистической устойчивостью, что позволило ему создать социальную статистику, а также организовать всемирное статистическое движение. Любопытно, что в настоящее время популярность Кетле связана в основном с изобретенным им индексом массы тела для оценки уровня ожирения человека (вес в килограммах, деленный на квадрат роста в метрах).

Возникает наука статистика. Термин был введен немецким ученым Г. Ахенвалем (*Gottfried Achenwall*) в его книге 1749 года издания как производное итальянских терминов *ragione di stato* (практиче-

ская политика) и *statista* (государственный деятель) для обозначения науки о государствах и государственных ресурсах.

Френсис Гальтон (1822—1911), как и выдающийся биолог Чарлз Дарвин (1809—1882), был внуком известного английского врача, натуралиста и философа Эразма Дарвина (1731—1802). Оба в молодости входили в круг так называемой «золотой молодежи». Длительные путешествия решительно изменили обоих. Ф. Гальтоном овладела идея, что талант достается человеку не случайно, а по наследству. В процессе разработки этой идеи Ф. Гальтон изобрел все основные понятия линейной многомерной статистики, прежде всего, регрессию и корреляцию, а также ряд других, на первый взгляд, не связанных вещей, таких как понятие антициклона в метеорологии, свисток для измерения восприятия колебаний в психофизике, а также и евгенику — псевдонауку о добровольной селекции общества. Его научные занятия никогда не оплачивались, но все-таки были вознаграждены рыцарским званием командора Британской империи (1909). В связи с недавним пересмотром роли выдающихся людей, придерживавшихся расистских взглядов, в США и Великобритании, имя Ф. Гальтона, как и имя К. Пирсона, было убрано в 2020 г. со всех основанных ими подразделений в Университетском Колледже Лондона.

В дальнейшем вплоть до середины XX в. наука о данных развивалась двумя независимыми ветвями: математическая статистика и общая теория статистики. Основные вехи этого развития:

— разработка теории нормального распределения и метода наименьших квадратов в связи, прежде всего, с необходимостью усреднения измерений положения небесных объектов, выполнявшегося в разных странах десятком астрономов (Карл Гаусс в Германии (*K. Gauss*, 1777—1855) и Пьер-Симон Лаплас во Франции (*P-S. Laplace*, 1749—1827)) — первая половина XIX в.;

— установление постоянства некоторых социальных характеристик (таких как процент самоубийств в данной стране) и создание социальной статистики, опирающейся на теорию вероятностей (Адольф Кетле (*A. Quetelet*, 1796—1884)) — XIX в.;

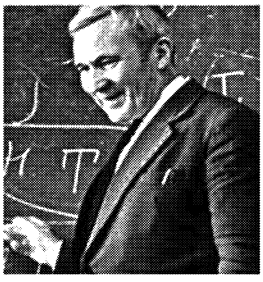
— разработка методов анализа многомерной информации (регрессия, корреляция, дисперсионный анализ, анализ главных компонент, факторный анализ и пр.) в контексте исследования таланта как наследственного дарования (Френсис Гальтон (*F. Galton*, 1822—1911), Карл Пирсон (*K. Pearson*, 1857—1936)) — на стыке XIX и XX вв.;

— формирование теории вероятностей и математической статистики в рамках теории измеримых множеств и функций (А. Н. Кол-

могоров (1903—1987), Харальд Крамер (*H. Kramer*, 1893—1985), Рональд Фишер (*R. Fisher*,1890—1962)) — первая половина XX в.

Карл Пирсон (*Karl* (*Charles*) *Pearson*, 1857—1936) — профессор прикладной математики Университетского Колледжа в Лондоне; автор философского труда «Грамматика науки» (1892), завершившего важный этап в развитии позитивизма как философии науки. Вдохновленный идеями Ф. Гальтона, создал основы математической статистики как науки. В частности, предложил математическую модель двумерного нормального распределения, включающую коэффициент корреляции как теоретическую конструкцию. Предложил современный подход к проверке статистических гипотез. Разработал теорию распределения хи-квадрат, в том числе для гипотезы независимости в таблицах сопряженности. Стойкий последователь идей евгеники, руководил кафедрой евгеники (1909), созданной Ф. Гальтоном. В связи с недавним пересмотром роли выдающихся людей, придерживавшихся расистских взглядов, в США и Великобритании, имя К. Пирсона, как и имя Ф. Гальтона, было убрано в 2020 г. со всех основанных ими подразделений в Университетском Колледже Лондона.

В дальнейшем развитие науки о данных переросло рамки математической статистики. Это произошло в связи, во-первых, с резким расширением круга приложений и, во-вторых, повсеместным внедрением электронной вычислительной техники. До середины двадцатого века математическая статистика развивалась как часть науки и старалась следовать стандарту, сложившемуся в науках о неживой природе.



Академик **Андрей Николаевич Колмогоров** (1903—1987) заложил фундамент современной теории вероятностей как математической теории, основанной на теории измеримых множеств и функций. Эта теория явилась надеж-

ной базой для дальнейшего развития математической статистики. Ему удалось также получить замечательные результаты в различных разделах кибернетики, включая конструктивный подход к теории информации, основанный на понятиях сложности алгоритмов. В России запомнилась предпринятая А. Н. Колмогоровым в 70—80-х гг. XX в. реформа школьной математики, основанная на широком использовании современных понятий, прежде всего, преобразования и эквивалентности. Увы, реформа провалилась (не без существенных промахов со стороны организаторов: например, в учебнике моей дочери по геометрии для 6 класса в 1978 г. я не нашел ни одного корректно сформулированного определения — в каждом отсутствовало какое-либо существенное ограничивающее условие).

В этих науках на основе многократных экспериментов определялись основные параметры/признаки, характеризующие данную систему/явление, а также механизмы их взаимодействия, позволяющие охарактеризовать идеализированную, но достаточно адекватную модель системы. Таковы, например, распределение Больцмана вероятностей того, что система в статистической механике будет находиться в определённом состоянии как функция энергии этого состояния и температуры, или распределение Максвелла для параметров частиц идеального газа, связывающее такие признаки как скорость частицы, ее энергия, импульс и т. д. Эти модели позволяют надежно прогнозировать поведение системы в тех или иных конкретных условиях. Соответственно, господствующая парадигма математической статистики при анализе данных такова. Постулируется то или иное семейство распределений (генеративная модель); имеющиеся данные рассматриваются как случайная выборка отдельных, независимых друг от друга, наблюдений, которые используются: иногда для проверки адекватности модели, иногда — для оценки тех или иных ее параметров или более общих свойств.

При обращении к системам живой природы — в биологии, экологии, психологии, социологии — эта схема рассыпается. Здесь, как правило, никто не может определить ни основные признаки, ни механизмы их взаимодействия, ни, тем более, адекватную модель вероятностного распределения. В этих условиях внимание исследователей концентрируется на отдельных проблемах. Общий прогресс достигается путем продвижения в отдельных направлениях:

— методы распознавания образов и машинного обучения в задаче классификации объектов (Фрэнк Розенблатт (*F. Rosenblatt*, 1928—1971), Эммануил Маркович Браверман (*E. M. Braverman*, 1931—1977), Владимир Наумович Вапник (*V. N. Vapnik*, р. 1936) — вторая половина XX в.;

— методы выявления ассоциаций между множествами категорий в больших компьютерных базах данных (*data mining*, т. е. майнинг данных) — на стыке XX и XXI вв.;

Российский исследователь **Эммануил Маркович Браверман** (1931—1977) первым предложил геометрическую интерпретацию проблемы распознавания образов как проблему отыскания компактных классов точек в пространстве признаков. Вместе с соавторами он предложил использование ядерных функций для так называемого ядерного трюка (*kernel trick*) — линеаризации проблемы разделения «вычурных» образов путем перехода в виртуальное пространство высшей размерности. Одним из первых предложил парадигму машинного обучения. Наметил пути автоматизации формирования интерпретируемых решений в анализе данных.

— методы обучения глубоких (многоуровневых) нейронных сетей решению задач распознавания и аннотации изображений, машинного перевода текстов, распознавания и предсказания речи и пр. (*deep learning*, т. е. глубокое — иногда говорят, глубинное — обучение) — XXI в.;

— широкое внедрение элементов искусственного интеллекта (*Artificial Intelligence*, AI) через тотальную связь между компьютерами, огромные скорости связи и вычислений, а также глубокие нейронные сети — XXI в.

В настоящее время наука о данных — одно из наиболее популярных направлений научного и технологического прогресса. Следует, однако, понимать, что многие кардинальные вопросы этой дисциплины как научного и технологического направления далеки от решения. Это определяется тем, что основные цели и методы науки о данных направлены на обогащение знания об объекте исследования или приложения. Между тем, наши представления о знании как таковом пока что довольно бедны и неточны.

Кстати говоря...

1. Различия в подходах

1.1. Перед учеными поставили задачу: предсказать исход скачек. Всем желающим принять участие в проекте выдали по 100 тыс. долл. Результаты:

Биолог:

— Я провел всесторонний анализ анатомии лошадей, нужно замерить вес, рост, объем мышц, объем легких, длину хвоста и цвет глаз у каждой лошади, и по моим таблицам определить, какая из них добежит первой.

Матстатистик:

— Я собрал данные о забегах, начиная с XVI в., на всех ипподромах мира, и теперь по дате забега и погоде могу предсказать, какая лошадь выиграет.

Физик-теоретик:

— А можно получить еще 200 тыс. долларов для окончания исследований?

— Ну, вы хоть что-нибудь уже сделали?

— Конечно! Построил модель для шарообразной лошади в вакууме.

1.2. Три математика и три физика собираются ехать поездом на конференцию молодых ученых в другой город. Они встречаются перед кассой на вокзале. Первыми покупают физики — по билету на человека. Математики же покупают один билет на троих.

Физики:

— В поезде контролер, двоих без билета оттуда выгонят!

Математики:

— Не выгонят. У нас есть метод.

Перед отправкой поезда математики все набиваются в один туалет. Когда контролер подходит к туалету и стучит, дверь приокрывается, оттуда высовывается рука с билетом. Контролер компостирует билет, после чего все они без проблем доезжают до пункта назначения.

После конференции те же вновь встречаются на вокзале. Физики, по примеру математиков, покупают один билет. Математики не берут ни одного.

Физики:

— А что же вы покажете контролеру?

Математики:

— У нас есть метод.

В поезде физики набиваются в один туалет, математики — в другой. Незадолго до отправления один из математиков подходит к туалету, где прячутся физики. Стучит. Высовывается рука с билетом. Математик забирает билет и возвращается к коллегам.

Мораль: не используй математический метод, если не понимаешь его сути.

1.3. Пессимист видит темный зловещий туннель. Оптимист видит свет в конце туннеля. Реалист видит набирающий скорость состав в туннеле. Машинист поезда видит трех идиотов, стоящих перед ним на рельсах.

1.4. — Когда я был маленький, я ходил с дедушкой в синагогу и просил у бога велосипед.

— И что, получил?

— Нет, но я понял, что бог работает по-другому. Я украл велосипед и стал просить у бога прощения.

1.5. В офисе сломался компьютер. Вызвали специалиста. Тот приехал, все починил. Два рассказа о том, как это происходило.

Сотрудник: «Пришел программист, пристально посмотрел на компьютер, воздел руки к небу, что-то пошептал, повернул мой стул 10 раз против часовой стрелки, пнул компьютер ногой, еще раз что-то пошептал и ушел. Все заработало. Настоящий маг».

Программист: «Очередной вызов — что-то случилось с компьютером. А сотрудник шнур навертел на ножку стула — видать, вертится постоянно на стуле. Я матерюсь, распутываю шнур. Потом передвинул компьютер подальше, вставил выскочивший штекер и ушел».

2. Данные и их заполнение

2.1. — Как у тебя с твоей девушкой?

— Мы расстались.

— А чего?

— Поругались. Она кричит: «Ты не любишь меня!» Я ей: «Оля, да люблю я тебя!»

— А она?

— А она Лена.

2.2. Воскресенье. Птичий рынок. Идет человек — на поводке белый медведь. Прошел один круг по рынку, второй. На третьем останавливает его милиция: — Не положено тут с медведем ходить.

Тот отвечает: — Я только хочу посмотреть в глаза тому парню, который мне в прошлом году продал маленького, беленького, пушистого хомячка.

2.3. Бежит Заяц по лесу. Навстречу ему Медведь.

— Ты куда, косой? — спрашивает Медведь.

— Приказ по лесу вывесили. У кого пять лап, пятую отрезать, чтобы не мешала.

— У тебя что, пять лап?

— Да нет. Но руководит Осел. Он сперва отрезает, а потом считает.

2.4. — Милый, ты где? — Я на охоте...

— А кто там так громко дышит? — Это медведь...

2.5. — Дорогая, где чай? Я никак не могу его найти.

— Ах, какой ты беспомощный! Чай в аптечке, в банке из-под какао с наклейкой «соль»!

2.6. — Девушка, девушка, сколько вам лет?

— Столько, насколько я выгляжу.

— Вай, не морочьте мне голову, люди столько не живут!

2.7. — Когда у вас день рождения?

— Одиннадцатого мая.

— Какого года?

— Любого года.

2.8. — Сколько длился каменный век?

— Пока не кончились камни.

2.9. — Ал-ле... Это пятьдесят... Один... Сорок шесть... Тридцать... Два? — Нет!

— Так зачем... было трубку ... снимать?..

2.10. — Где вы работаете?

— На почте. Штемпелюю письма.

— Должно быть, это очень скучная работа?

— Скучная?! Что вы! Совсем нет! Ведь каждый день — новая дата.

2.11. — Слыхали?! Мужику дали десять лет за то, что он бросил жену.

— Не говорите вздор! За это не дают срок. Я сам бросил двух жен — и ничего!

— А вы с какого этажа бросали?

2.12. Жена собирает мужа в командировку. Чемодан, бритва, полотенце и т. д. — все, что нужно в дороге. Муж смотрит, лежит пачка сливочного масла и гвозди! Он:

— Масло зачем?

— В дороге проголодаешься, намажешь на хлеб и покушаешь.

— А гвозди?

— Так вот же они!

2.13. Разговаривают двое:

— Ты знаешь, что по статистике каждая вторая женщина изменяет своему мужу?

— Да что мне статистика? Мне нужны фамилии, адреса, телефоны.

2.14. — Скажите пожалуйста, кто вас стриг?

— Мастер...

— Я понимаю, что мастер. А по профессии он кто?

2.15. Психиатр показывает пациенту листок бумаги с кляксой:

— Что вы видите?

— Грустного одинокого человека, изнывающего от общения с идиотами, нудной работы задешево и прочей несправедливости.

Врач, всхлипывая:

— А на картинке?

2.16. — Где можно недорого отметить день рождения?

— В календарике, ручкой.

2.17. — Какое у тебя хобби?

— Еда.

— Готовишь?

— Ем.

2.18. — А как правильно: Иран или Ирак?

— Да вроде бы и так, и так говорят.

2.19.

Заболела стюардесса. Вместо нее взяли новенькую. Самолет разгоняется по полосе, капитан говорит:

— Сообщи пассажирам, что взлетаем.

Та в микрофон:

— Уважаемые пассажиры, сейчас наш самолет взлетит на воздух...

2.20. Работодатель: Назовите вашу главную слабость.

Кандидат: Я даю семантически корректные, но практически неприменимые ответы на вопросы.

Работодатель: Могли бы вы привести пример?

Кандидат: Да, мог бы.

3. Уточнение данных

3.1. Дискотека в деревенском клубе. Парень - девушке:

— Ты танцуешь?

— Пока нет.

— Отлично, пойдем — поможешь трактор толкнуть!

3.2. Жена мужу, уходя на работу:

— Водку, сок, мясо пожаришь.

— Водку-то зачем?

— Совсем спятил со своей водкой. Сказала же: «Вот кусок мяса — пожаришь.»

3.3. Телефонный звонок:

— Можно Мойшу к телефону?

— Здесь таких нет.

Через 5 минут:

— Таки можно Мойшу к телефону?

— Такие не проживают.

Третий звонок:

— Будьте любезны, Михаил Борисыча.

— Мойша! Тебя к телефону!

3.4. — Леночка, с днем рождения! А сколько вам стукнуло, можно узнать?

— Конечно. Когда я выходила замуж, мне было 20, а ему 40, то есть я в два раза моложе. Сейчас ему 70, а мне, стало быть, 35!

3.5. — Вы играете на скрипке?

— Нет.

— А ваш брат?

— Да.

— Что «да»?

— Тоже нет.

3.6. Больной доктору:

— Доктор, меня вылечат?

— Мы это увидим.

— А я?

— А вы, возможно, и нет.

3.7. Ученик:

— Я не могу разобрать, что вы написали в моей тетради.

Учитель:

— Я написал: «Пиши разборчиво!»

3.8. — Батя, а сколько стоит свадьба?

— Не знаю, сынок; я все еще расплачиваюсь.

3.9. — Сколько вам лет?

— Не помню.

— А на вид меньше.

Тема 2
ОДНОМЕРНЫЙ АНАЛИЗ

В этой теме рассмотрены задачи суммаризации и визуализации для самого простого вида данных, когда признак всего один.

Рассматриваются два взаимно-дополнительных математических уточнения понятия «признак». Объясняются понятия гистограммы и функции плотности, центральной точки (центра) и разброса признака. Изложены две точки зрения на задачу суммаризации: первая — это классическая вероятностная, а вторая — аппроксимационная, в рамках которой разброс данных разложим на объясненную и необъясненную части.

Разница между количественными и качественными признаками определяется с помощью операции взятия среднего. Для количественных признаков взятие среднего имеет смысл, в то время как для качественных признаков — нет. Это различие стирается на бинарных признаках, представляющих отдельные категории. Они задаются так называемыми фиктивными переменными.

Современные вычислительные подходы, (а) имитирующие природу методы и (б) бутстрэп для оценки доверительных интервалов, объяснены на примерах в Проектах в конце главы.

Одномерные данные (1D данные) представляют собой набор объектов, описанных с помощью какого-либо одного признака, качественного или количественного. В этом случае нет смысла говорить о коррелировании — будут рассмотрены методы суммаризации. Не существует простого универсального критерия, по которому можно было бы определить, является признак качественным или количественным. С практической точки зрения полезен следующий критерий: признак — количественный, если имеет смысл расчет его среднего значения.

2.1. Две математические модели для понятия «признак»

В настоящее время в анализе данных сосуществуют две принципиально разные математические формулировки понятия «признак». Одна исходит из табличного представления данных, вторая — из представления о функции плотности. Согласно первому определению, признак x — это отображение множества объектов в множество его

значений. Например, согласно табл. 1.9 признак w1 (Длина чашелистика) отображает множество 150 цветков ириса в числа, лежащие в интервале [4.4, 5.7], так что w1 на объекте 1 равно 5.1, а на объекте 10 — 4.8. При таком определении упор делается на представление объектов в виде точек одномерного или многомерного пространства, а таблица данных может трактоваться как числовая матрица. Правда, это определение привязывает признак к конкретному множеству объектов, так что универсальность метода измерения (в данном примере длины чашелистика) теряется из виду.

Согласно второму определению, признак — это вероятностное распределение, обычно задаваемое функцией плотности. Конкретные объекты при этом трактуются как случайная выборка из данного распределения. Понятие плотности не привязано ни к какому множеству объектов и в этом плане вполне универсально.

Как видно, данные два определения касаются разных аспектов понятия «признак» и абсолютно не согласованы друг с другом. Подобное встречается в современной науке. Например, такое важное явление как «свет» в физике может пониматься и как электромагнитное поле, и как поток корпускулярных частиц. Эти два понимания тоже не согласованы. В некоторых ситуациях удобно одно из них, а в некоторых — другое.

В определенном смысле в данных двух определениях отражается многовековой спор двух философских течений — реализма и номинализма. Реалисты, начиная с древнегреческого мыслителя Платона, утверждают, что «индивидуальным вещам предшествуют бестелесные идеи и лишь они обладают подлинным бытием», тогда как номиналисты объявляют реальными только наблюдаемые вещи, а идеи — абстрактными отображениями вещей в нашем мозгу. В определенном смысле это давнее расхождение отражается в современной науке о данных. Анализ данных исходит из того, что именно наблюденные данные первичны, а все остальное — построения нашего разума, «идеальные аппроксимации» наблюдений, тогда как машинное обучение олицетворяет научный подход, в котором приоритет — за моделью, а наблюдения — не более чем случайные реализации модельных распределений.

Таблица 2.1 подытоживает особенности рассмотренных определений понятия «признак».

Таблица 2.1

Основные отличия двух математических определений понятия «признак»

№	Аспект	Признак — отображение	Признак — распределение
1	Характер	Эмпирический	Универсальный
2	Объекты	Индивидуальны	Случайная выборка

Окончание табл. 2.1

№	Аспект	Признак — отображение	Признак — распределение
3	Адекватный аппарат	Матричная алгебра	Теория вероятностей
4	Удачные приложения	Управление реальными производствами	Крупномасштабные и долгосрочные системы

2.2. Понятие гистограммы для количественного признака

Самый понятный и исчерпывающий способ агрегирования — это распределение, представленное так называемой *гистограммой*. На оси признака x отмечают границы, в которых изменяется признак, т. е. его минимальное и максимальное значения на имеющихся объектах. Отмеченный интервал, называемый также размахом признака, делят на некоторое число непересекающихся интервалов одинаковой длины, так называемых *бинов* (рис. 2.1). Затем подсчитывают, сколько объектов попадает в каждый отдельный бин, и рисуют столбики высотой, соответствующей числу объектов в бине. В результате получают гистограмму.

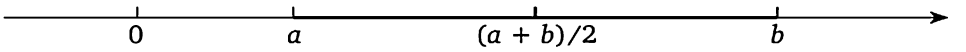


Рис. 2.1. **Если бинов всего два, их разделяет точка полуразмаха**

На рис. 2.2 представлены гистограммы для признаков таблицы ирисов.

Вопрос 2.1. Почему бины не должны пересекаться?

Ответ. Каждый объект попадает только в один бин, если бины не пересекаются, так что сумма количеств объектов в бинах совпадает с общим числом объектов. Если же бины пересекаются, то один объект может принадлежать двум разным бинам, так что нарушается принцип «один объект — один голос».

Вопрос 2.2. Правда ли, что в случае всего двух бинов, их разделяет точка полуразмаха?

Ответ. Да, потому что размеры бинов совпадают, а их всего два в этом случае (см. рис. 2.1).

Вопрос 2.3. Почему на рис. 2.2 прямоугольники на гистограммах слева выше прямоугольников на гистограммах справа?

Ответ. Потому что на гистограммах справа бины в два раза короче, чем на гистограммах слева. Следовательно, число объектов, в них попадающих, в среднем в два раза меньше.

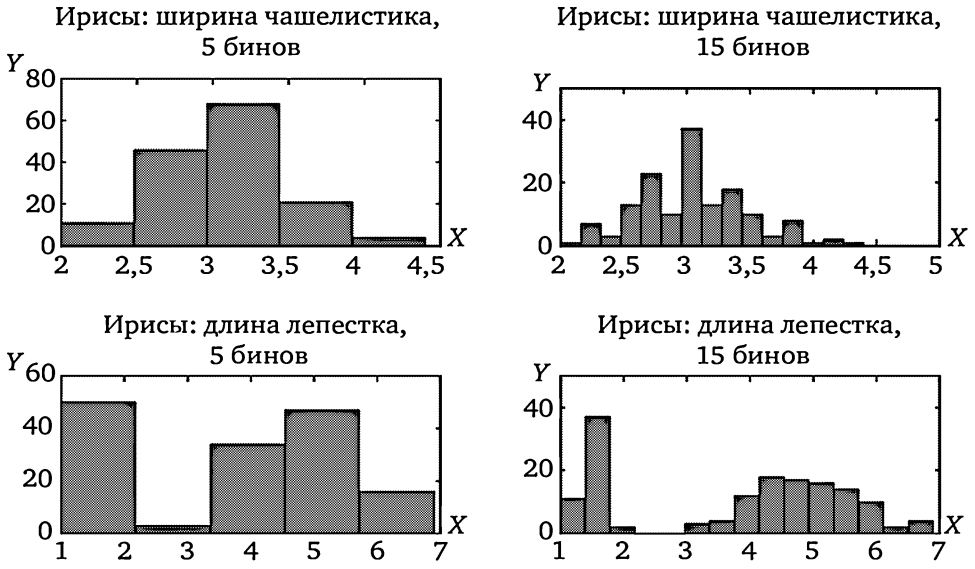


Рис. 2.2. **Гистограммы количественных признаков в таблице данных Ирисы:** рассматриваемый признак отмечен на оси *X*, количество объектов — на оси *Y*. Форма диаграммы зависит не только от распределения, но и от числа бинов

На рис. 2.3 и 2.4 представлены два часто встречающихся типа гистограмм. Первый демонстрирует так называемый степенной закон, или распределение Парето, оно же — распределение Ципфа (см. рис. 2.3). Этот тип гистограммы часто встречается в социальных системах. Согласно эмпирическим исследованиям, такие показатели, как уровень доходов, размер сообществ, уровень производительности и им подобные, распределены по степенному закону. Получается, что ничтожно малая часть индивидов или объектов обладает большим богатством/популярностью/производительностью, в то время как большая часть индивидов остается почти ни с чем. Тем не менее все индивиды — существенная часть системы, в которой «нищие» создают такую среду, что только несколько счастливчиков могут соперничать между собой.

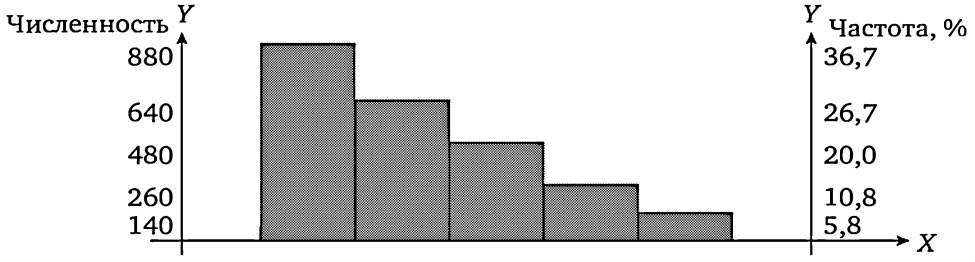


Рис. 2.3. **Распределение степенного типа**

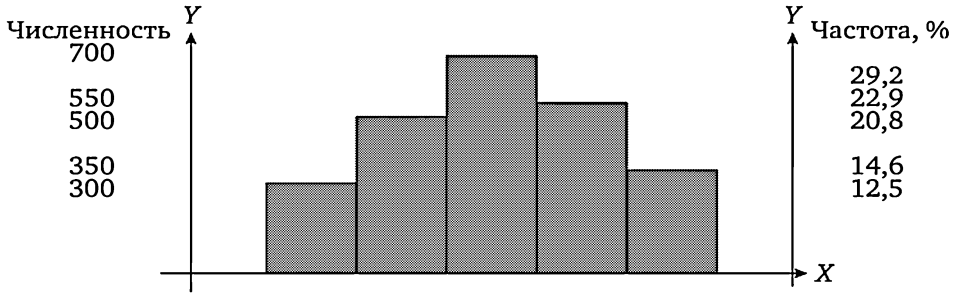


Рис. 2.4. **Распределение Гауссова типа (в форме колокола)**

Другой вид распределения, часто встречающийся в природе, показан на рис. 2.4. Этот тип гистограмм соответствует так называемому нормальному, или Гауссову, закону распределения. Распределение ошибок измерений, и, в целом, величин, получаемых под действием небольших независимых случайных эффектов, считается Гауссовым.

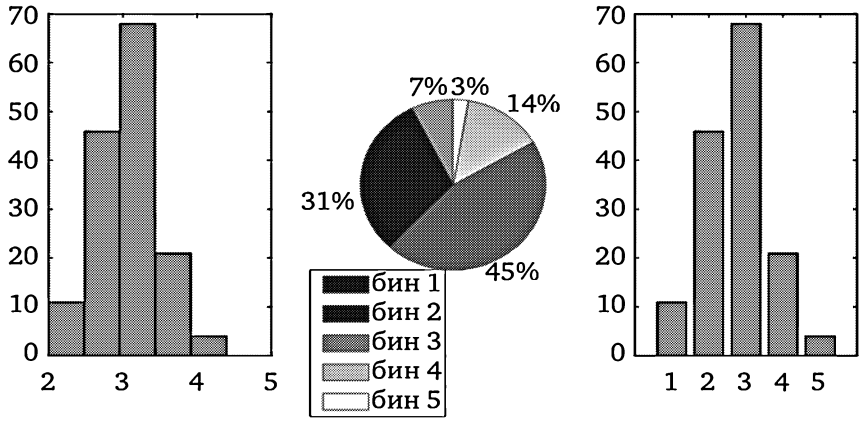


Рис. 2.5. **Распределение ирисов в 5 бинах:**
в виде гистограммы (слева),
круговой диаграммы (каравай, в центре),
столбцовой диаграммы (bars, справа)

Гистограммы и *круговые диаграммы* (*караваи*) служат для отображения разных свойств распределения. Гистограммы используются для того, чтобы показать, как распределены объекты вдоль оси X, а караваи представляют относительные размеры частей распределения, попадающих в разные бины. Существует множество других форматов визуализации распределений, таких как пузырьковые, кольцевые, паутинные диаграммы. Для их создания можно использовать программу Excel компании Microsoft, входящую в пакет Microsoft Office.

Ф2.3. Гистограмма и плотность распределения

Рассмотрим N объектов, пронумерованных от 1 до N: $i = 1, 2, \ldots, N$. Значения признака x на этих объектах представляют собой индексированный набор чисел $x_1, \ldots, x_N$. Этот набор чаще всего обозначается как $X = \{x_1, \ldots, x_N\}$.

Чтобы построить n бинов в интервале (a, b), где a — левый, а b — правый конец интервала, требуется $n - 1$ разделитель в точках $x_k = a + k(b - a)/n$ $(k = 1, 2, \ldots, n - 1)$. На самом деле эта же формула может быть использована и при $k = 0$, когда $x_0 = a$, и при $k = n$, когда $x_n = b$. Этот прием может оказаться полезным при нахождении числа объектов N_k, попадающих в k-й бин $k = 1, 2, \ldots, n$. Левая граница k-того бина находится в точке $x_{k-1} = a + (k - 1)(b - a)/n$, а правая в точке $x_k = a + k(b - a)/n$. Одну из границ следует исключить из бина, чтобы бины не пересекались даже в граничных точках. Числа N_k, $k = 1, 2, \ldots, n$, характеризуют *распределение* признака. *Гистограмма* — это визуальное представление распределения. Для k-того бина рисуют столбик высотой N_k $(k = 1, 2, \ldots, n)$ (см. рис. 2.2—2.5). Заметим, что выбор числа бинов определяется пользователем исходя из характера распределения и цели анализа; надежных теоретических рекомендаций не существует.

Гистограмму можно рассматривать как эмпирическое представление теоретической так называемой функции плотности распределения. Функция плотности $p(x)$ выражает понятие вероятности, но не напрямую с помощью своих значений $p(x)$, а с использованием интегралов от p(x) на интервалах значений признака [f, g]. Такой интеграл равен площади фигуры между осью абсцисс и кривой p(x) на интервале [f, g]. Он выражает вероятность того, что случайная величина, распределенная по p(x), попадет в интервал [f, g].

Следовательно, площадь под всей кривой должна быть равна 1. Чтобы получить такую функцию из произвольной неотрицательной функции, ее шкалируют с помощью деления на полную площадь фигуры, лежащей между графиком функции и осью абсцисс, т. е. на величину интеграла от функции, взятого от минус до плюс бесконечности.

Функция плотности степенного закона записывается как $p(x) = a/x^{\lambda}$, где $\lambda > 0$ характеризует степень уменьшения вероятности при увеличении x. Считается, что этот закон распределения выражает явление, которое называется эффектом Матфея. Это название связано с притчей о талантах из Евангелия от Матфея: «Ибо всякому имеющему дастся и приумножится; а у не имеющего отнимется и то, что имеет» (Мф. 25:29). Эффект Матфея проявляется, например, в популярном механизме «предпочтительного присоединения» в интернете. Согласно этому механизму, вероятность того, что новый пользователь зайдет на определенный сайт, пропорциональна популярности этого сайта, измеряемой, например, количеством «кликов» пользователей на сайт за единицу времени. При этом

у более популярного сайта больше посетителей, что делает его ещё более популярным.

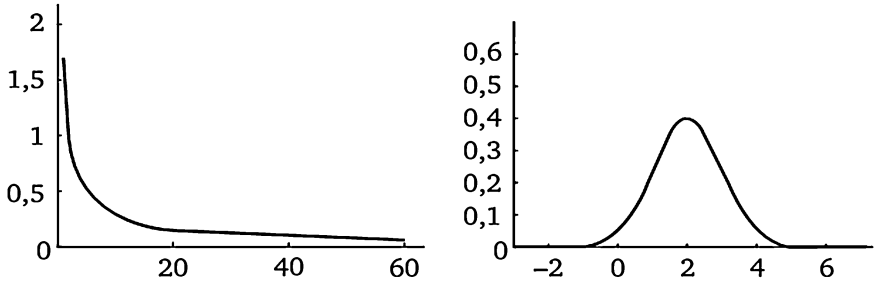


Рис. 2.6. **Функции плотности:**
слева степенной закон при $\lambda = -0.8$,
справа — нормальное распределение $N(2, 1)$

Функция плотности нормального, или Гауссова, распределения имеет форму $p(x) = C \exp[-(x-a)^2/2\sigma^2]$, где C — константа, выбранная так, чтобы площадь между кривой и осью абсцисс равнялась единице. Эту функцию обозначают как $N(a, \sigma)$. Распределения ошибок измерений, как и другие распределения, порожденные наложением многих малых случайных независимых эффектов, приближаются к Гауссовому, что может быть обосновано с привлечением математического аппарата теории вероятностей. Параметры Гауссова распределения, a и σ^2, имеют естественную интерпретацию: a выражает ожидаемое или среднее значение, а σ^2 — дисперсию.

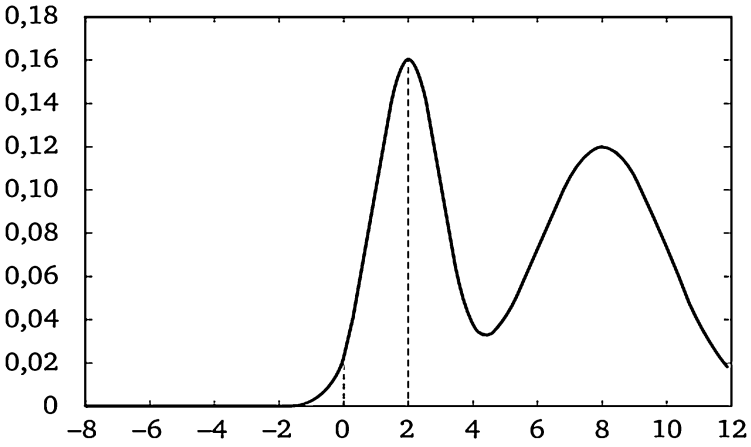


Рис. 2.7. **Функция плотности $p(x)$ в данном случае — смесь двух нормальных распределений, $N(2, 1)$ с весом 0.4, и $N(8, 2)$ с весом 0.6:**
площадь между двумя пунктирными линиями соответствует вероятности того, что величина x попадет в интервал между 0 и 2 — она не так уж и велика для данной $p(x)$!

Эти параметры и способы их оценки по данным описаны в параграфе 2.5. Следует отметить, что для Гауссова распределения вероятность того, что величина x попадет в интервал $a \pm \sigma$, составляет примерно 88 %, а вероятность попадания в интервал $a \pm 3\sigma$ — практически единицу, точнее, 99.7 %. Последнее означает, что при относительно скромных выборках, до тысячи объектов, случайные значения, выбранные в соответствии с нормальным распределением, как правило, не смогут оказаться вне интервала, определенного этим так называемым «правилом трех сигма». Гауссово распределение может быть приведено шкалированием к стандартному виду $N(0, 1)$ с нулевым математическим ожиданием и дисперсией равной 1. Для этого значение переменной x сдвигают к математическому ожиданию, a, а затем нормализуют квадратным корнем из дисперсии, σ, называемым стандартным отклонением. Это преобразование называется z-стандартизацией, по-английски *z-scoring*, и записывается как $y = (x - a)/\sigma$, где y — преобразованная переменная.

Еще одно популярное распределение — равномерное распределение на интервале $[a, b]$. Его функция плотности постоянна и равна $p(x) = 1/(b - a)$, так что вероятность интервала (l, r), лежащего внутри $[a, b]$, составляет $p = (r - l)/(b - a)$ и пропорциональна длине интервала.

В2.4. Вычисление гистограммы

Построим гистограммы, изображенные на рис. 2.2. Для этого потребуется сначала загрузить данные об Ирисах в качестве массива MATLAB, используя функцию

```
>> ir = load('Data\iris.dat');
% Предполагается, что данные об Ирисах хранятся
% в папке "Data" в файле с названием iris.dat
```

После этого признаки «Ширина чашелистика» и «Длина лепестка» могут быть выделены из массива в отдельные переменные

```
>> w2  = ir(:,2); w3  = ir(:,3);
% весь второй и весь третий столбцы –
% именно они соответствуют данным признакам
```

Вопрос 2.4. Построить рис. 2.2 с четырьмя окнами, используя систему MATLAB.

Ответ. Воспользуемся функцией subplot и зададим необходимые преобразования координат с помощью функции axis. Гистограмма строится с помощью функции hist:

```
>>subplot(2,2,1);hist(w2(:,2),5);subplot(2,2,2);
hist(w2,15);axis([2 5 0 50]);
```

```
>>subplot(2,2,3);hist(w3(:,3),5);subplot(2,2,4);
hist(w3,15);axis([1 7 0 50]);
```

Функция axis([a b c d]) задает координаты на рисунке, т. е. помещает на ось X интервал $[a, b]$ и интервал $[c, d]$ на ось Y. Круговые и столбчатые диаграммы строятся с помощью функций pie и bar, соответственно.

2.5. Дальнейшая суммаризация: центр и рассеяние

Для удобства признаки часто представляют всего двумя величинами. Первая величина выражает положение распределения, его центр или другую точку «нормы». Вторая величина характеризует уровень разброса распределения вокруг центра. Мы рассмотрим наиболее популярные характеристики центра (табл. 2.1) и разброса (табл. 2.2). Самым популярным является понятие *среднего значения*.

Рабочий пример 2.1

Среднее

Рассмотрим числовое множество $X = \{1, 1, 5, 3, 4, 1, 2\}$. Его среднее вычисляется суммированием всех элементов с последующим делением на их количество: c = (1 + 1 +5 + 3 +4 + 1 +2)/7 = 17/7 = 2.42857..., и составляет, с округлением до первого знака, c = 2.4. Почему до первого? Потому что исходные данные — целые числа, и нет смысла вести вычисления с большей точностью.

Самостоятельная работа

2.1. Рассчитайте средние значения количественных признаков в данных о Компаниях (см. табл. 1.2).

2.2. Рассчитайте средние значения признаков в таксоне 1 (*Iris setosa*) данных об Ирисах (см. табл. 1.3).

С одной стороны, среднее значение — это самое точное приближение к числам данного множества, которое можно получить. С другой стороны, среднее обладает не очень хорошим свойством: оно не устойчиво к выбросам. Если, например, добавить в множество X из Рабочего примера 2.1 значение 23, сильно отличающееся от остальных, т. е. *выброс*, то среднее сильно увеличится: c = (17 + 23)/8 = 5. Чтобы избежать сдвига среднего значения при наличии выбросов, пользуются понятием «усеченного среднего» (*trimmed mean*). Усеченное среднее рассчитывают после того, как удаляют максимальные и минимальные наблюдения из диапазона данных. *Медиана* — крайний вариант усеченного среднего, когда оставляют только средний элемент упорядоченного по возрастанию ряда чисел множества X.

Рабочий пример 2.2

Медиана

Вычислим медиану для множества $X = \{1, 1, 5, 3, 4, 1, 2\}$ из предыдущего примера. Сначала отсортируем X в порядке возрастания: 1, 1, 1, 2, 3, 4, 5. Медиана определяется как элемент, находящийся в середине отсортированного ряда. В этом случае медиана равна 2, что меньше среднего значения 2.4. Следовательно, распределение сдвинуто влево, в сторону маленьких значений. Если добавить ко множеству X элемент 23, выброс, то отсортированный ряд примет вид: 1, 1, 1, 2, 3, 4, 5, 23. В середине нового ряда находятся два элемента, 2 и 3. Медиана находится как среднее этих двух элементов, т. е. равна $(2 + 3)/2 = 2.5$. Заметим, что медиана изменяется гораздо меньше, чем среднее значение, которое для нового расширенного множества равно 5.

Самостоятельная работа

2.3. Рассчитайте медианы количественных признаков в данных о Компаниях (см. табл. 1.2).

2.4. Рассчитайте медианы признаков в таксоне 1 (*Iris setosa*) данных об Ирисах (см. табл 1.3).

Обзор характеристик центра распределения и их свойств дан в табл. 2.2.

Чем симметричнее распределение, тем ближе друг к другу среднее значение и медиана. Ширина чашелистика из таблицы Ирисов (см. табл. 1.3) имеет среднее значение, равное 3.05, и медиану, равную 3, т. е. эти показатели довольно близки. Среднее степенного закона всегда сдвинуто в сторону больших значений. Поэтому в качестве центрального значения часто используют медиану, которая менее чувствительна к выбросам, поскольку равномерно добавленные с обеих сторон отсортированного ряда выбросы не влияют на его середину.

Таблица 2.2

Обзор характеристик центра распределения.

#	Название	Объяснение	Комментарии
1	Среднее значение	Среднее арифметическое значение признака	Минимизирует сумму квадратов ошибок. Является оценкой математического ожидания распределения. Чувствительно к выбросам и форме распределения
2	Медиана	Середина упорядоченного ряда значений признака	Минимизирует сумму модулей ошибок. Является оценкой математического ожидания распределения. Не чувствительна к выбросам. Чувствительна к форме распределения

Окончание табл. 2.2

#	Название	Объяснение	Комментарии
3	Середина	Середина размаха	Минимизирует максимум модулей ошибок. Является оценкой математического ожидания распределения. Чрезвычайно чувствительна к выбросам. Не чувствительна к форме распределения
4	P-квантиль, где P — число между 0 и 1	Значение, разделяющее исходное множество объектов, предварительно отсортированное, на доли P и (1 – P) от общего количества объектов. При этом количество объектов с большими значениями признака пропорционально P (верхний P-квантиль) или же количество объектов с меньшими значениями признака пропорционально P (нижний P-квантиль)	Не чувствителен к выбросам. Чувствителен к форме распределения
5	Мода	Бин, на который приходится максимум гистограммы	Зависит от размера бина. Может существовать несколько мод

Середина соответствует среднему значению равномерного распределения, у которого частоты во всех бинах равны. В отличие от среднего значения и медианы, середина зависит только от размаха данных, а не от распределения. Очевидно, что она чрезвычайно чувствительна к выбросам, т. е. к изменениям максимального или минимального значений выборки.

Понятие P-квантиля расширяет понятие медианы, которая является 50 % квантилем.

Рабочий пример 2.3

P-квантиль (перцентиль) и фондовый коэффициент

Зададим уровень p = 10 % и определим верхний 10%-ный квантиль для признака «Ширина чашелистика» w2. Это должно быть 16-е значение в отсортированных по убыванию данных, т. е. 3.6. Почему надо выбирать именно 16-е значение? Потому что всего имеется 150 объектов, 10 % от их общего числа составляет 15. После удаления пятнадцати объектов с наибольшей «Шириной чашелистика», т. е. от 3.7 мм до 4.4 мм, первым становится именно 16-й объект, а для него значение признака равно 3.6.

Аналогично, нижний 10%-ный квантиль определяется удалением последних 15 объектов в ряду, отсортированном по убыванию w2, т. е. значением признака на объекте номер 135 в этом ряду, 2.5. Это позволяет сказать, что «Ширина чашелистика» у 80 % объектов заключена между 2.5 и 3.6, тогда как полный интервал значений меняется от 2.0 до 4.4.

Фондовый коэффициент характеризует уровень неравенства в распределении значений признака. Он определяется как отношение среднего самых больших 10 % значений и среднего самых малых 10 % значений. С использованием MATLAB его можно посчитать так:

```
>> w2s = sort(w2, 'descend');
>> fc = mean(w2s(1:15))/mean(w2s(136:150))
```

Первая команда сортирует w2 в порядке убывания, а вторая берет отношение средних для первых 15 и последних 15 значениях сортированной последовательности. Получаем fc = 1.67. Эта разница не очень велика. Фондовый коэффициент может быть значительно выше в социальных системах.

Самостоятельная работа

2.5. Рассчитайте 5 % верхний квантиль признака «Длина чашелистика» по данным об Ирисах (см. табл 1.3).

2.6. Рассчитайте 10 % нижний квантиль признака SH по данным о компьютерных атаках (см. табл 1.4).

Рабочий пример 2.4

Мода

Судя по гистограммам на рис. 2.2, некий бин в середине является модальным для распределения признака «Ширина чашелистика». В случае пяти бинов каждый бин занимает 1/5 от размаха признака, (4.4 – 2.0)/5 = 0.48. Средний бин — это интервал от 2.96 до 3.44, его частота может быть рассчитана в MATLAB с помощью команды

```
>> m5  = length(find(w2< = 3.44 & w2>2.96))
```

приводящей к m5 = 68. Относительная частота моды в этом случае равна 68/150 = 0.453 или 45.3 %. В случае 15 бинов каждый бин занимает 0.16 (почему?). Средний бин здесь — интервал от 2.96 до 3.12 и его частота — 37, т. е. 24.7 %.

Самостоятельная работа

2.7. Постройте гистограмму признака Длина чашелистика по данным об Ирисах (см. табл 1.3) с 10 бинами и определите модальный бин.

2.8. Постройте гистограмму признака SH по данным о компьютерных атаках (см. табл 1.4) с 5 бинами и определите моду.

Меры разброса используются для того, чтобы оценить степень ошибочности соответствующей характеристики центральности. *Стандартное отклонение* — это квадратный корень из средней квадратичной ошибки среднего значения. Популярность этой меры

связана с принципом наименьших квадратов, который в настоящее время превалирует в анализе данных. Использование принципа наименьших квадратов может быть объяснено хорошими свойствами решений, которые он дает, с точки зрения анализа данных, и свойствами нормального распределения, с точки зрения теории вероятности. Более строгое объяснение этого принципа приведено в параграфе Ф.2.6.

Обзор характеристик рассеяния и их свойств приводится в табл. 2.3.

Таблица 2.3

Обзор характеристик рассеяния

#	Название	Объяснение	Комментарии
1	Стандартное отклонение	Квадратный корень из среднего отклонения от среднего значения	Минимизируется средним значением Является оценкой квадратного корня из дисперсии распределения
2	Абсолютное отклонение	Среднее абсолютное отклонение от медианы	Минимизируется медианой
3	Полуразмах	Максимальное отклонение от середины размаха	Минимизируется серединой размаха

Абсолютное отклонение выражает среднее абсолютное отклонение от медианы. Как правило, его находят относительно среднего значения, поскольку именно среднее значение чаще всего берется в качестве центральной характеристики. Однако, среднее абсолютное отклонение лучше соответствует медиане, так как именно медиана минимизирует его.

Полуразмах выражает максимально возможное отклонение значений от середины интервала, поэтому имеет смысл использовать эти две характеристики вместе, как это делают исследователи, изучающие методы построения классификационных правил.

Самостоятельная работа

2.9. Рассмотрите характеристики рассеяния и характеристики центра распределения: под одинаковыми номерами в табл. 2.2 и 2.3 и установите параллели между ними.

Ф2.6. Центр и рассеяние: формулировки

Существует два принципиально разных взгляда на методы суммаризации и коррелирования данных. Согласно одному взгляду, наиболее четко выраженному в классической математической ста-

тистике, данные порождены неким математическим механизмом, обычно, вероятностным, который называют *генеративной моделью*. Поэтому их используют для восстановления механизма или хотя бы некоторых его свойств. С точки зрения подхода анализа данных, механизм порождения данных не существует или не очень интересен, а главная задача — это поиск закономерностей в самих данных как они есть.

Ф2.6.1. Подход анализа данных

Пусть дано множество наблюденных значений признака $X = \{x_1, \ldots, x_N\}$. Задача — представить это множество в «сжатом» виде некой центральной точкой a. Эта центральная точка a должна минимизировать среднее величин индекса расстояния от нее до всех наблюденных значений

$$D(X, a) = [d(x_1, a) + d(x_2, a) + \ldots + d(x_N, a)]/N. \quad (2.1)$$

В зависимости от того, как определен индекс расстояния $d(x_i, a)$, оптимальными могут быть разные значения a. Например, естественно определить $d(x_i, a) = |x_i - a|^p$ для некоторого вещественного положительного p (правило Минковского). К сожалению, нет единого простого метода минимизации (2.1) для произвольного p. Для трех значений $p = 1$, 2, и ∞ (бесконечность), впрочем, можно указать простые правила вычисления оптимального a.

Рассмотрим сначала принцип наименьших квадратов, соответствующий $p = 2$. Согласно этому принципу, индекс расстояния — это квадрат разности, $d(x, a) = |x - a|^2$. Тогда минимум среднего расстояния (2.1) достигается в точке a, равной среднему арифметическому значению c. Это доказывается приравниванием нулю производной от выражения (2.1) при квадратах разностей, подставленных вместо $d(x_i, a)$. Среднее арифметическое значение определяется выражением:

$$c = \sum_{i=1}^{N} x_i / N. \quad (2.2)$$

Следовательно, среднее расстояние $D(X, c)$ (2.1) в этом случае — ни что иное, как

$$s^2 = \sum_{i=1}^{N} (x_i - c)^2 / N. \quad (2.3)$$

Эта величина называется *дисперсией* среднего значения.

Если определить индекс расстояния более традиционным способом просто как величину отклонения $d(x, a) = |x - a|$, т. е. $p = 1$ в (2.1), то нетрудно доказать индукцией по N, что оптимальное зна-

чение a (центр) при минимизации (2.1) — это медиана, ms, а $D(X, a)$ в этом случае — среднее абсолютное отклонение от медианы

$$ms = \sum_{i=1}^{N} |x_i - m| / N. \tag{2.4}$$

На самом деле медиана — единственный оптимум только при нечетном N. Если же N четное, то оптимальной будет любая величина между двумя числами, $x_{N/2}$ и $x_{N/2+1}$, находящимися в середине упорядоченного ряда элементов X, включая медиану.

Если расстояние $D(X, a)$ в (2.1) определено не как среднее, а как максимум из расстояний, $D(X, a) = \max\{d(x_1, a), d(x_2, a), \ldots, d(x_N, a)\}$, то минимум (2.1) достигается на середине размаха mr. Вместе с тем само правило взятия максимума величин $d(x_i, a)$ может рассматриваться как предельный случай минимизации суммы (2.1) по правилу Минковского при $p \Rightarrow \infty$.

Рассмотренные выше утверждения объясняют связь между характеристиками центра и характеристиками разброса, приведенными в табл. 2.1 и 2.2. Каждая из характеристик центра минимизирует соответствующую ей меру разброса.

Задача минимизации среднего индекса расстояния, особенно в форме Минковского, может быть представлена в рамках подхода восстановления данных, который позволяет развить для аппроксимационной задачи минимизации (2.1) некоторое подобие теории. Согласно этому подходу, любой метод анализа данных перекодирует данные к более простому, в какой-то мере «идеальному», виду. В частности, в задачах вычисления центральной величины, все наблюденные значения рассматриваются как «зашумленные» реализации некого неизвестного значения a, так что имеют место равенства:

$$x_i = a + e_i \text{ при } i = 1, 2, \ldots, N, \tag{2.5}$$

где e_i — аддитивные, т. е. суммируемые, остатки, которые необходимо минимизировать, чтобы обеспечить наилучшее качество восстановления данных в случае их утери — замену каждого значением a. Чтобы не связываться с совершенно неясной проблематикой минимизации всех остатков одновременно, используется какой-либо интегральный критерий. Существует достаточно общее семейство таких критериев — критерий Минковского, математически называемый также нормой L_p. Норма Минковского для многомерного набора остатков определяется как

$$L_p = (|e_1|^p + |e_2|^p + \ldots + |e_N|^p)^{1/p},$$

где p — некоторое положительное число.

При разных значениях p задача минимизации L_p или, эквивалентно, ее p-й степени L_p^p, будет давать разные центры Минковского. Самые часто используемые значения $p = 1$, 2, и ∞ (бесконечность) как раз и дают вышеупомянутые критерии:

(1) Принцип наименьших квадратов: минимизировать $L_2^2 = e_1^2 + e_2^2 + \ldots + e_N^2$, при $p = 2$.

Минимизация L_2^2 по неизвестному a эквивалентна задаче минимизации среднего квадрата отклонений $e_i = x_i - a$. Оптимальное a в этой задаче — среднее значение.

(2) Принцип наименьших модулей: минимизировать $L_1 = |e_1| + |e_2| + \ldots + |e_N|$, при $p = 1$.

Минимизация L_1 по неизвестному a эквивалентна задаче минимизации среднего абсолютного отклонения. Оптимальное значение a в этой задаче — медиана, $a = ms$.

(3) Принцип наименьшего максимума (Чебышева) $L_\infty = \max(|e_1|, |e_2|, \ldots |e_N|)$, при $p = \infty$.

Минимизация L_∞ по неизвестному a эквивалентна задаче минимизации максимального отклонения. Оптимальное значение a в этой задаче — середина размаха, $a = mr$.

Может показаться, что критерий Минковского L_p^p для модели (2.5) является всего лишь тривиальной переформулировкой критерия минимизации расстояния (2.1). Как говорится, «старое вино в новые меха». Но это не так. Дело в том, что уравнение (2.5) позволяет не только оценить расстояние, но и разложить разброс данных на «объясненную» и «необъясненную» составляющие.

Особенно просто это можно сделать для случая $p = 2$, т. е. принципа наименьших квадратов. Величина критерия в точке a, равной среднему значению c, равна $L_2^2 = (x_1 - c)^2 + (x_2 - c)^2 + \ldots + (x_N - c)^2$. Раскроем скобки в этом выражении, приведем подобные и получим, что

$$L_2^2 = x_1^2 + x_2^2 + \ldots + x_N^2 - 2c(x_1 + x_2 + \ldots + x_N) + Nc^2 = \\ = x_1^2 + x_2^2 + \ldots + x_N^2 - Nc^2 = T(X) - Nc^2,$$

где $T(X)$ — *квадратичный разброс данных*, который определяется как сумма квадратов наблюденных значений $T(X) = x_1^2 + x_2^2 + \ldots + x_N^2$. При выводе данного разложения мы использовали равенство $(x_1 + x_2 + \ldots + x_N) = cN$, являющееся переформулировкой определения среднего арифметического.

Таким образом, квадратичный разброс данных согласно модели (2.5) равен

$$f(u) = Ce - \frac{(u-m)^2}{2\sigma^2}, \tag{2.6}$$

т. е. состоит из двух частей: первая, Nc^2, характеризует ту часть разброса, которая *объясняется* моделью (2.5), а вторая — ту, которая

остается *необъясненной*, L_2^2. Поскольку разброс данных — константа, минимизация L_2^2 эквивалентна максимизации Nc^2. Разложение разброса данных на две составляющие позволяет оценить адекватность модели (2.5) не только с помощью дисперсии, усредненного квадратичного критерия, но и с помощью относительной величины объясненной части $L_S^2 / T(X)$. Похожее разложение может быть найдено и для принципа наименьших модулей L_1 (см. [30]).

Вопрос 2.5. Какую часть разброса данных объясняет модель (2.5) для данных вопроса из Рабочего примера 2.1?

Ответ. Разброс данных $X = \{1, 1, 5, 3, 4, 1, 2\}$ по определению равен $T(X) = 1^2 + 1^2 + 5^2 + 3^2 + 4^2 + 1^2 + 2^2 = 1 + 1 + 25 + 9 + 16 + 1 + 4 = 57$. Согласно материалу раздела Ф. 2.2, объясненная часть разброса равна $N\bar{x}^2 = 7 \cdot (2.4286)^2 = 41.2857$, где $\bar{x}$ — среднее значение X. Таким образом, среднее для этих данных объясняет 41.2857/57 = 0.724, т. е. 72.4 % разброса данных. Для проверки можно рассчитать необъясненную часть разброса непосредственно:

$$L_2^2 = (\bar{x}-1)^2 + (\bar{x}-1)^2 + (\bar{x}-5)^2 + (\bar{x}-3)^2 + (\bar{x}-4)^2 + (\bar{x}-1)^2 + (\bar{x}-2)^2 = 2.04 + 2.04 + 6.61 + 0.33 + 2.47 + 2.04 + 0.18 = 15.71.$$

Её доля составляет 15.71/57 = 0.276, т. е. 27.6 %, что дополняет предыдущий результат до 100 % и этим подтверждает правильность вывода.

Вопрос 2.6. Рассмотрим не аддитивную, как в (2.5), а мультипликативную модель ошибки $x_i = a(1 + e_i)$, предполагая, что ошибки e_i пропорциональны величинам x_i. Каков будет центр a по принципу наименьших квадратов для этой модели?

Ответ. Согласно принципу наименьших квадратов, центр должен минимизировать сумму квадратов ошибок. По модели каждая ошибка может быть выражена как $e_i = x_i/a - 1 = (x_i - a)/a$. Следовательно, критерий записывается как

$$L_2^2 = e_1^2 + e_i^2 + \ldots + e_N^2 = (x_1 / a - 1)^2 + (x_2 / a - 1)^2 + \ldots (x_N / a - 1)^2.$$

По условию оптимальности первого порядка, найдем производную L_2^2 по a и приравняем ее нулю. Производная равна $(L_2^2)' = -(2 / a^3)\Sigma_i (x_i - a)x_i$. Допустим, что оптимальное значение a отлично от нуля, тогда условие первого порядка эквивалентно перепишется как $\Sigma_i (x_i - a)x_i = 0$, так что

$$a = \frac{\Sigma_i x_i^2}{\Sigma_i x_i} = \frac{\Sigma_i x_i^2 / N}{\Sigma_i x_i / N}.$$

Здесь знаменатель — это среднее значение c, а числитель может быть выражен через дисперсию s^2, так как имеет место соотношение $s^2 = \Sigma_i x_i^2 / N - \Sigma_i x_i / N$, которое несложно доказать. После преоб-

разований получим, что оптимальное $a = s^2/c + 1$. В статистике часто рассматривается близкая величина, коэффициент вариации s/c.

Заметим, что и стандартное отклонение, и абсолютное отклонение не превышают половины размаха признака. Этот факт может быть доказан математически [32].

Таблица 2.4

Центры Минковского для признака «Ширина чашелистика» из данных об Ирисах при разных p

p	p-центр	*Необъясненная доля разброса, %*
0.5	3.0	28.40
1	3.0 (медиана)	10.82
2	3.057 (среднее)	1.98
3	3.083	0.44
4	3.103	0.11
5	3.120	0.01

Вопрос 2.7. Величину c назовем p-центр Минковского, если она минимизирует среднее расстояние Минковского (2.1) с показателем степени p. Докажите, что центр Минковского не меняется при изменении масштаба; точнее, меняется в соответствии с масштабом.

Вопрос 2.8. Для признака Ширина чашелистика из данных об Ирисах вычислите центр Минковского при $p = 0.5, 1, 2, 3, 4, 5$.

Ответ. Решение приведено в табл. 2.4. Оно получено с помощью программы cm.m, разработанной в рамках Проекта 2.1.

Вопрос 2.9. Докажите, что p-центр Минковского возрастает с ростом p.

Ф2.6.2. Теоретико-вероятностный подход

В классической математической статистике множество наблюденных значений признака $X = \{x_1, x_2, \ldots, x_N\}$ считается случайной выборкой из генеральной совокупности, заданной вероятностным распределением с функцией плотности $f(x)$, где каждый элемент x_i выбран независимо от всех остальных элементов. При этом каждое наблюдение x_i как случайная величина моделируется тем же распределением $f(x_i)$. Среднее значение таких «случайных величин» само является случайной величиной, функция плотности которой — среднее всех функций плотности $f(x_i)$. Но поскольку эти плотности совпадают, то и средняя плотность — та же, т. е. $f(x)$. Аналоги среднего значения и дисперсии для генеральной совокупности определяются по функции $f(x)$. Теоретические величины среднего и дисперсии могут быть выражены через функцию плотности $f(u)$ как определенные интегралы $\mu = \int uf(u)du$ и $s^2 = \int (u-\mu)^2 f(u)du$. При этом и среднее значение, и медиана, и середина размаха — это несмещенные оценки среднего генеральной совокупности. Дисперсия

среднего значения оказывается в N раз меньше, чем дисперсия генеральной совокупности, поэтому его стандартное отклонение с ростом N уменьшается со скоростью $\sqrt{N}$.

В том случае, когда генеральная совокупность имеет Гауссово распределение $N(\mu, \sigma)$ с функцией плотности

$$f(u) = Ce^{-\frac{(u-m)^2}{2^2}}, \tag{2.6}$$

где $C = (2\pi\sigma^2)^{-\frac{1}{2}}$, то c из выражения (2.2) является оценкой μ, а s^2 из выражения (2.3) — оценкой σ^2 в (2.6), сделанными по выборке X.

Чтобы в этом убедиться, рассмотрим проблему оценки параметров нормального распределения (2.6) по выборке. Для этого будем считать, что μ и σ^2 в (2.6) неизвестны, а наблюдённые значения в X — это случайная независимая выборка из генеральной совокупности с распределением (2.6). Отсюда следует, что вероятность случайного наблюдения x_i равна $C\exp\{-(x_i-\mu)^2/2s^2\}$, а вероятность получения всей выборки X — произведение этих вероятностей в силу их независимости. Таким образом, вероятность выборки X равна

$$L(X) = \Pi_{i\in I} C\exp\{-(x_i-\mu)^2/2\sigma^2\} = C^N \exp\{-\Sigma_{i\in I}(x_i-\mu)^2/2\sigma^2\}.$$

Согласно широко применяемому в математической статистике принципу *максимального правдоподобия*, наиболее соответствуют данным выборки X те значения μ и $\sigma 2$, в которых вероятность наблюдённой выборки $L(X)$ или, эквивалентно, et логарифм $\ln(L(X))$, достигает максимума. При заданной дисперсии максимум логарифма $\ln(L) = N\ln(C) - \Sigma_{i\in I}(x_i-\mu)^2/2\sigma^2$ достигается на том значении μ, которое минимизирует выражение в показателе экспоненты, $E = \sum_{i\in I}(x_i-\mu)^2$, так как величина $N\lambda\nu(X)$ — постоянная.

Вопрос 2.10. Ответьте — почему минимизирует? Ведь мы хотим *максимизировать* логарифм.

Как мы помним из предыдущего параграфа, это оптимальное μ — не что иное, как среднее значение выборки. Таким образом, по принципу максимального правдоподобия искомое μ минимизирует сумму квадратичных расстояний, как и в (2.1). Это значит, что допущение независимости наблюдений в выборке и принцип максимального правдоподобия, применительно к Гауссовой генеральной совокупности, приводят к принципу наименьших квадратов. Разумеется, отсюда не следует, что принцип наименьших квадратов имеет смысл только в рамках гипотезы о независимой выборке из нормальной популяции. Этот принцип сам по себе достаточно универсален в анализе данных.

Аналогичным образом, оптимальное значение σ^2 максимизирует ту часть $\ln(L)$, которая зависит от нее, т. е. $g(\sigma^2) = -N\ln(\sigma^2)/2 -$

$-\sum_{i\in I}(x_i-\mu)^2/2\sigma^2$ (в предположении, что μ известно). Оптимальное значение σ^2 находится из условий первого порядка для $g(\sigma^2)$. Продифференцируем $\ln(L)$ по σ^2 и приравняем производную к 0: $dg/d(\sigma^2) = -N/(2\sigma^2) + \sum_{i\in I}(x_i-\mu)^2/2(\sigma^2)^2 = 0$. Отсюда $\sigma^2 = \sum_{i\in I}(x_i-\mu)^2/N$, что означает, что дисперсия s^2 является оценкой максимального правдоподобия параметра σ^2 Гауссова распределения.

Заметим, что в том случае, когда μ тоже неизвестно и рассчитывается для выборки по формуле среднего значения (2.2), выборочная дисперсия s^2 по формуле (2.3) оказывается смещенной оценкой дисперсии σ^2 и должна быть скорректирована. Для этого N в знаменателе заменяют на $N-1$. Такая замена может быть объяснена тем фактом, что равенство (2.2), введенное в формулу дисперсии, эффективно уменьшает число степеней свободы с N до $N-1$.

Если предположение о случайности, независимости и принадлежности к Гауссову распределению данных оправдано, то среднее значение и дисперсия — единственные теоретически обоснованные оценки центра и разброса данных. Доказано, что Гауссово распределение хорошо описывает ситуации, в которых небольшие эффекты добавляются друг к другу. Предположение нормальности или независимости может оказаться в высшей степени нереалистично в других случаях. Но даже и тогда не стоит отказываться от понятий среднего значения и дисперсии. Их использование может оказаться полезным в рамках аппроксимационного подхода к анализу данных, рассмотренного в предыдущем разделе.

В2.7. Центр и рассеяние: вычисления

В среде MATLAB есть функции mean(X) и median(X). Они могут быть применены не только к векторам, но и к матрицам. Они возвращают строку, содержащую средние значения или медианы, соответственно, столбцов матрицы. Чтобы найти середину размаха, можно воспользоваться комбинацией двух функций mr = (max(X) + min(X))/2.

Стандартное отклонение вычисляется с использованием функции std(x), если знаменатель формулы (2.3) равен $N-1$, или std(x,1), если равен N.

Стабильная версия размаха, которую можно использовать при больших значениях N или при наличии выбросов в данных, определяется с использованием понятия квантиля. В начале зададим величину пропорции p, например, 1 % или 5 %. Верхний (нижний) p-квантиль — это такое число xp из множества X, что доля объектов с большими (меньшими) чем xp, значениями признака составляет p. P-квантиль для вектора x в MATLAB вычисляется так: сначала необходимо отсортировать x в убывающем порядке командой sx = sort(x, 'descend'), а после этого квантиль вычисляется как sx(k), где k = ceil(p*length(x)).

Размах между 2p-квантилями определяется как интервал между значениями p-квантилей, «растянутый» в соответствии с долей наблюдений, не попавших в него, $(xp - px)/(1 - 2p)$, где xp и px — это верхний и нижний p-квантили, соответственно. Например, при $p = 0.05$ % и $N = 100000$ xp отрезает 50 наибольших, а px — 50 наименьших чисел из X.

2.8. Бинарные и категоризованные признаки

Категоризованные признаки отличаются от количественных не только тем, что их значения — строки символов, а не числа. Строки символов все равно кодируются числами при обработке. Более существенная разница в том, что вычисление среднего значения для количественного признака имеет смысл, а для качественного — нет. Например, для таких признаков как Тип протокола — со значениями tcp, icmp, udp — в данных Компьютерные атаки или Сектор экономики — Торговля, Энергия или Мануфактура — в данных Компании — среднее значение не имеет смысла, даже если категории кодированы числами. Иногда утверждают, что даже такие признаки, как количество поставщиков в данных Компании, нельзя рассматривать как количественные, поскольку их значения должны быть целыми, тогда как средние могут быть дробными. На наш взгляд, подобные утверждения излишне ограничительны; фраза «среднее число поставщиков составляет 3.15» имеет смысл, поскольку может быть переформулирована переходом от единиц к сотням единиц: в среднем на 100 компаний приходится 315 поставщиков.

Признак, принимающий значения «Да» или «Нет», иногда называют булевым, связывая его с Булевой алгеброй, где утверждения бывают либо истинными, либо ложными, но не количественными. Чтобы не вносить путаницу, мы будем называть такие признаки *бинарными* и пользоваться операциями арифметики, а не Булевой алгебры. Нам удобнее оперировать с такими признаками как с количественными. Значения признака будем кодировать числами, 1 вместо «Да», и 0 вместо «Нет». Усреднение таких данных действительно имеет смысл.

Среднее значение бинарного признака, кодированного нулями и единицами, показывает, какую долю составляют наблюдения, соответствующие категории «Да». Другие определенные выше характеристики центральности дают меньше информации в бинарном случае. Медиана равна 1 тогда и только тогда, когда доля единиц составляет 0.5 или больше, в противном случае она равна 0. В тех редких случаях, когда число наблюдений четно и доля единиц составляет ровно половину, медиана принимает значение 0.5. Мода равна 1 либо 0 в тех же случаях.

Для категоризованных признаков нет надобности вводить бины: сами категории выполняют роль бинов. В отличие от количественного случая, их порядок безразличен. Как правило, их гистограммы изображают столбцами или ростками (stem). На рис. 2.8 представлено распределение категорий tcp, icmp и udp признака «Тип протокола» из данных о компьютерных атаках.

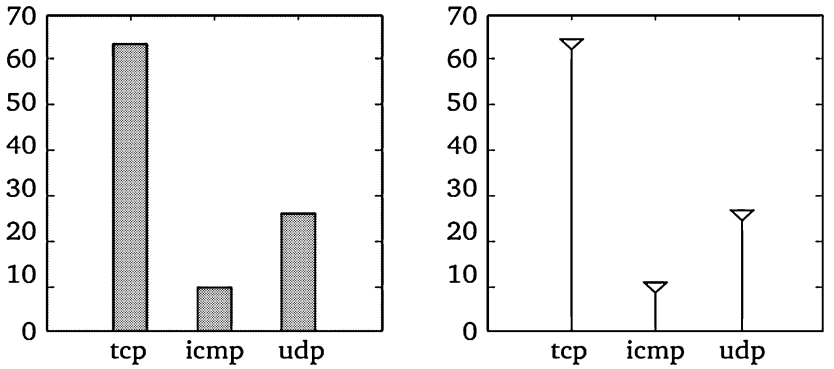


Рис. 2.8. **Распределение категорий tcp, icmp и udp признака Тип протокола из данных Компьютерные атаки, представленное столбцами (bar) слева и ростками (stem) справа**

Распределение признака может быть выражено абсолютным числом объектов, приходящихся на каждую категорию, т. е. $D = (64, 10, 26)$, в абсолютной шкале. При переходе к относительной шкале число объектов каждой категории делят на общее число объектов, $64 + 10 + 26 = 100$, и получают распределение частот $d = (0.64, 0.10, 0.26)$.

Рассмотрим пример еще более неоднородной выборки. В Великобритании полиция имеет право остановить и обыскать человека, показавшегося подозрительным, прямо на улице (процедура «Останови и Обыщи», ОО). Пресса критикует эту практику, подозревая в ней расистскую тенденцию. В частности, так были интерпретированы статистические данные за 2005—2006 гг. Распределение 878 153 случаев ОО по цвету кожи задержанного представлено в табл. 2.5.

Таблица 2.5

Распределение случаев ОО по цвету кожи в 2005—2006 гг.

Расовая категория	Количество ОО	Относительная частота ОО, %
Чернокожий (Ч)	131 723	15
Азиат (А)	70 250	8
Белый (Б)	676 180	77
Всего	878 153	100

Обращает на себя внимание, что доля ОО в категории Б трижды превышает долю двух других категорий, вместе взятых. Однако, доля Б во всем населении оказывается еще больше, что и приводит к утверждениям, что чернокожие подвергаются процедуре ОО непропорционально часто (см. подробнее в параграфе 3.3).

Вопрос 2.11. Какая из категорий является модальной для распределения в табл. 2.5?

Ответ. Мода, т. е. наиболее вероятная категория в табл. 2.5, — это Б.

Существует множество коэффициентов, позволяющих оценить, насколько распределение отличается от равномерного, т. е. такого, при котором вероятность попадания в какой-либо интервал значений признака зависит только от длины интервала, а не от его местоположения. Самые популярные из них — это энтропия и коэффициент Джини, определенные ниже в разделе Ф. 2.3. Последний также называют качественной дисперсией.

Понятие энтропии было введено для измерения количества информации в сигналах, передаваемых по каналам связи. Редкие сигналы несут больше информации, чем частые. Кроме того, количество информации в независимых сигналах можно суммировать, для того чтобы оценить всю переданную информацию. Эти два условия приводят к необходимости использования логарифма величины $1/p$, т. е. $-\log(p)$, в качестве меры количества информации в сигнале, вероятность которого равна p. Логарифм берется по основанию 2, поскольку все цифровые устройства используют двоичную систему счисления. Энтропия определяется как среднее количество информации, приходящееся на одну категорию качественного признака, рассматриваемую как сигнал. За единицу измерения количества информации принят один бит — энтропия равномерно распределенного бинарного признака, т. е. двоичного разряда с равновероятными значениями. Иными словами, бит — это количество информации в ответе на вопрос, допускающий только два ответа, при условии, что никакого знания о возможном ответе не было. Максимум энтропии для признака с m категориями, $H = \log(m)$, достигается при равномерном распределении.

Максимум индекса Джини, $(m - 1)/m$, также достигается на равномерном распределении. Индекс Джини позволяет оценить средний уровень ошибки метода *пропорционального предсказания*. Такое предсказание осуществляется в ситуации, когда объекты, у которых неизвестны значения некоторого качественного признака, появляются случайно и независимо один за другим. Пропорциональный классификатор будет случайным образом присваивать объектам категории признака в соответствии с вероятностями категорий. Средняя ошибка, т. е. вероятность того, что объекту категории, встречающейся с частотой p, будет приписана другая категория, равна

$p(1-p) = p - p^2$. Так, например, при $p = 20\ \%$, средняя ошибка составит $0.2(1 - 0.2) = 0.16\ \%$..

Рабочий пример 2.5

Энтропия и индекс Джини

В табл. 2.6 представлены все шаги, которые нужны, чтобы рассчитать энтропию и индекс Джини с использованием p — вероятности (относительной частоты) категории.

Таблица 2.6

Энтропия и индекс Джини для распределения обысков по расе из табл. 2.5

Распределение		Энтропия		Качественная дисперсия	
Категория	Относительная частота p	Информация $-\log(p)$	Взвешенная информация $-p\log(p)$	Ошибка $1-p$	Дисперсия $p(1-p)$
Ч	0.15	2.74	0.41	0.85	0.128
А	0.08	3.64	0.29	0.92	0.074
Б	0.77	0.38	0.29	0.23	0.177
Итого	1.00	—	0.99	—	0.378

Энтропия — это среднее количество информации в трех категориях, $H = -p_1\log(p_1) - p_2\log(p_2) - p_3\log(p_3)$. Отношение энтропии из табл. 2.6 к максимально возможной энтропии составляет $0.99/1.585 = 0.625$, так как для $m = 3$ максимум энтропии равен $H = \log(3) = 1.585$.

Индекс Джини G определяется как средняя ошибка пропорционального классификатора. Принцип действия пропорционального классификатора определяется для объектов, о которых ничего не известно, кроме распределения категорий $\{p_l\}$. Этот классификатор приписывает объектам категорию l с вероятностью p_l. В нашем случае, $G = p_1(1-p_1) + p_2(1-p_2) + p_3(1-p_3) = 0.378$. Максимум индекса Джини равен $(m-1)/m$ — значение, соответствующее равномерному распределению, $G = 2/3$. Тогда относительный индекс Джини составит $0.378/(2/3) = 0.567$, что не очень отличается от относительной энтропии.

Самостоятельная работа

2.10. Найдите распределение, энтропию и индекс Джини для признака Тип протокола данных о компьютерных атаках (табл 1.4).

2.11. Найдите распределение, энтропию и индекс Джини для признака Бол в данных о малых городах английского побережья (табл 1.5).

Сформулируем теперь основные понятия с использованием языка математики.

Качественный признак, например, «Тип протокола» в данных «Компьютерные атаки» разбивает множество объектов так, что каждый объект попадает в одну и только в одну категорию. Такие признаки называются номинальными.

Рассмотрим номинальный признак с L категориями $l = 1, 2, \ldots, L$. Его распределение характеризуется количеством объектов $N_1, N_2, \ldots, N_L$, которые попадают в каждую из категорий. Заметим, что сумма численностей категорий равна общему числу объектов: $N_1 + N_2 + \ldots + N_L = N$. Относительные частоты, определяемые как $p_l = N_l/N$, в сумме дают единицу ($l = 1, 2, \ldots, L$). Это свойство вытекает из альтернативного и повсеместного характера категорий номинального признака — каждый объект покрывается какой-либо категорией, причем ни один объект не может принадлежать двум или более категориям.

Рис. 2.9. **Итоги голосования по партийному признаку должны суммироваться к 100 % согласно принципу «один избиратель — один голос». Ведущая явно в замешательстве — в одной из областей это явно не так!**

Поскольку категории номинального признака не упорядочены, их лучше визуализировать с помощью круговых диаграмм, «караваев», а не гистограмм.

Характеристики центра, за исключением моды, не имеют смысла для распределений качественных признаков. Однако, рассеяние или разнообразие, распределения ($p_1, p_2, \ldots, p_L$) измерить можно. Две популярные меры рассеяния: индекс Джини, или качественная дисперсия, и энтропия.

Индекс Джини — это средняя ошибка правила пропорционального предсказания. Согласно этому правилу, каждая категория l, $l = 1, 2, \ldots, L$, предсказывается случайно с вероятностью p_l, так что частота предсказанной категории l равна Np_l. Средняя ошибка предсказания категории l в этом случае равна $1 - p_l$, так что суммарная средняя ошибка равна:

$$G = \sum_{l=1}^{L} p_l(1 - p_l) = 1 - \sum_{l=1}^{L} p_l^2.$$

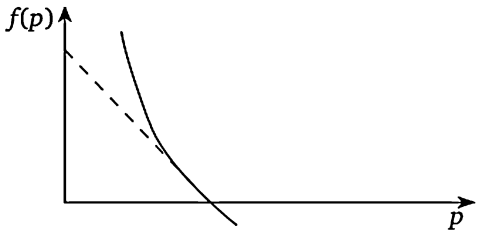


Рис. 2.10. **Графики функции ошибки $f(p) = 1 - p$ в индексе Джини (пунктирная линия) и количестве информации $f(p) = -\log(p)$**

Энтропия усредняет количество информации отдельных категорий. Величина информации о категории l определяется как $\log(1/p_l) = -\log(p_l)$ для любого l. Энтропия вычисляется по формуле

$$H = -\sum_{l=1}^{L} p_l \log p_l.$$

Энтропия не сильно отличается от индекса Джини, качественной дисперсии, так как при малых значениях p, величины $-\log(1 - p)$ и $1 - p$ почти совпадают. Этот факт хорошо известен (см. рис. 2.10).

Отдельный класс номинальных признаков — бинарные признаки. У бинарного признака только две категории. Такие признаки могут появляться сами по себе, как атрибуты, которые могут присутствовать у объекта, а могут и отсутствовать. Но часто они формируются в связи с категориями качественного признака. Например, категория «udp» Типа протокола в данных Компьютерные атаки может быть преобразована в бинарный признак в форме вопроса «Правда ли, что атака случилась при использовании протокола udp?». На этот вопрос может быть два ответа: «да» или «нет».

Бинарные признаки совмещают свойства качественных и количественных признаков. Принято считать, что основное различие между качественными и количественными типами шкал заключается в множествах допустимых преобразований. Допустимое числовое преобразование изменяет значения признака таким образом, что отношения между объектами по признаку сохраняются. Например, рост человека в сантиметрах может быть пересчитан в миллиметрах (для этого рост нужно умножить на 10), а температура, измеренная в градусах Фаренгейта, может быть преобразована в температуру в градусах Цельсия (для этого необходимо вычесть 32 и разделить результат на 1.8). Такое преобразование не изменяет отношения между различными областями, где температура была измерена в градусах Фаренгейта. Если в качестве новых температур выбрать произвольные значения, новое множество измерений будет представлять совершенно другую информацию. Этим определяется принципиальное различие между количественными и номи-

нальными признаками. Значения номинальных признаков можно сравнивать только на предмет совпадения-несовпадения категорий, так что допустимы всевозможные взаимно-однозначные преобразования их значений. У количественных признаков можно изменять масштаб и сдвигать точку отсчета (начало шкалы), т. е. допустимы только так называемые *аффинные* преобразования. Такие преобразования переводят значения x в значения y с помощью двух числовых параметров, $ax + b \Rightarrow y$, где a характеризует изменение масштаба, а b — сдвиг точки отсчета шкалы. Это различие между типами шкал, однако, не работает для бинарных признаков. Дело в том, что для бинарных признаков любое их числовое преобразование определяется всего двумя константами — теми, которые замещают 0 и 1 соответственно. Эти константы могут быть связаны со сдвигом точки отсчета и изменением масштаба. Точнее, чтобы преобразовать значения бинарного признака: 0 в α, а 1 в β, нужно всего два параметра: сдвиг b определяется величиной α, а масштаб a — разностью $\beta - \alpha$. Это означает, что бинарный признак одновременно и номинальный, и количественный.

Значения любого бинарного признака могут быть представлены двумя числами: 1 для «да», 0 для «нет». Иногда так закодированные категории называют *дамми* (*dummy*) или *фиктивными* переменными.

Вычислим дисперсию бинарного признака, у которого частота значения «да» равна p. Очевидно, среднее значение этого признака равно $c = p$. Дисперсия — это средний квадрат отклонения от среднего. Чтобы ее рассчитать, сложим Np величин $(1 - p)^2$ (квадрат отклонения единицы от c) и $N(1 - p)$ величин p^2 (квадрат отклонения нуля от c), получим $s^2 = p(1 - p) = 1 - p^2$. Стандартное отклонение — это квадратный корень из дисперсии, т. е. $s = \sqrt{p(1-p)}$. Очевидно, что стандартное отклонение достигает максимума при $p = 0.5$, т. е. в том случае, когда оба бинарных значения равновероятны. Размах всегда равен 1. При $p < 0.5$, медиана $m = 0$, а среднее абсолютное отклонение sm состоит из Np значений, равных 1, и $N(1 - p)$ значений равных 0, поэтому $sm = p$. При $p > 0.5$ $m = 1$ и число единиц равно $N(1 - p)$, отсюда $sm = 1 - p$. В целом, это означает, что, $sm = \min(p, 1 - p)$; эта величина не превышает стандартное отклонение. Действительно, если $p \leq 0.5$, то $p \leq 1 - p$ и, следовательно, $p^2 \leq p(1 - p)$, поэтому $sm \leq s$. Аналогично, если $p > 0.5$, то $p > 1 - p$ и $p(1 - p) > (1 - p)^2$, поэтому $sm < s$, что и доказывает утверждение.

Пусть качественный признак развернут во множество бинарных признаков, соответствующих его значениям $l = 1, 2, \ldots, L$. Тогда суммарная дисперсия всех L бинарных признаков равна индексу Джини, или качественной дисперсии, исходного признака

Использование бинарных признаков может быть включено в вероятностный контекст. Существуют две вероятностные модели для бинарных признаков: модель Бернулли и модель Пуассона. Согласно

модели Бернулли, при данном p, $0 \le p \le 1$, каждое наблюдение x_i равно 1 с вероятностью p, или равно 0 с вероятностью $1 - p$. По модели Пуассона, единицы рассыпаны случайно среди N бинарных разрядов, так что pN разрядов равны единице, а $(1 - p)N$ — нулю. Математическое ожидание в обеих моделях совпадает и равно p. Дисперсии же различаются: дисперсия распределения Бернулли равна $p(1 - p)$, как определено выше, а дисперсия распределения Пуассона равна p, что, очевидно, больше при любых положительных значениях p, поскольку сомножитель $1 - p$ в дисперсии Бернулли меньше 1. Похожие модели могут быть построены и для качественных признаков с более чем двумя категориями.

Существует вполне естественное, но почему-то упорно не признаваемое, отношение между дисперсиями количественных и бинарных признаков. Дисперсия количественного признака всегда меньше, чем дисперсия соответствующего бинарного признака. Следуя [33], пронормируем признак так, что диапазон (размах) изменения данных $X = \{x_1, \ldots, x_N\}$ — это интервал [0, 1]. Среднее c значений X разделяет интервал определенным образом. Обозначим долю значений, больших или равных c в X, через p; тогда доля наблюдений, меньших, чем c, будет равна $1 - p$. Какое распределение значений в X максимизирует дисперсию при заданном p? Согласно обозначениям выше, Np наблюдений попадают между 0 и c. Если любую из этих точек подвинуть в сторону границы отрезка, 0, дисперсия только возрастет. Аналогично, дисперсия возрастет, если сдвинуть любую из оставшихся $N(1 - p)$ точек, находящихся между c и 1, в сторону другой границы, 1. Отсюда следует, что дисперсия $p(1 - p)$ бинарной переменной с Np нулевыми и $N(1 - p)$ единичными значениям является максимальной при любом p. Это доказывает, что дисперсия и стандартное отклонение бинарного признака с распределением $(p, 1 - p)$ максимальны среди всех количественных переменных с таким же диапазоном значений.

Следовательно, не существует переменной с размахом [0, 1], дисперсия которой больше максимальной возможной дисперсии 1/4, которой обладает бинарный признак при $p = 0.5$. Стандартное отклонение этого бинарного признака равно 1/2, что составляет только половину размаха. Следовательно, стандартное отклонение любой количественной переменной не может быть больше, чем ее полуразмах.

Бинарные переменные обладают также максимальным абсолютным отклонением среди всех переменных такого же размаха. Это не трудно доказать по аналогии с рассуждениями выше.

Если распределение признака записано в массиве df, то команда MATLAB

```
>> bar(df,.4); h = axis; axis(1.1*h);
```

строит столбчатую диаграмму признака. Входные параметры здесь: 0.4 — ширина столбцов, 1.1 — коэффициент масштабирования, который позволяет выбрать удобное расстояние между столбцами и границами рисунка (см. рис. 2.8).

Энтропию и индекс Джини можно вычислить с помощью команд:

```
>> df = df/sum(df); h = -sum(df.*log2(df)); % h энтропия
>> df = df/sum(df); g = -sum(df.*(1-df)); % g индекс Джини
```

Вопрос 2.12. Сформируйте бинарные признаки для качественных категорий в данных о компьютерных атаках, после этого вычислите среднее значение и дисперсию каждого нового признака. Сравните сумму найденных дисперсий с индексом Джини исходного признака.

Ответ. Совпадают.

2.9. Более продвинутые понятия

Далее мы опишем несколько более продвинутых понятий в форме проектов. Эти проекты ориентированы на наиболее активных читателей. Конкретно, будут рассмотрены следующие понятия: критерий и центр Минковского; градиентный метод минимизации; метод оптимизации, инспирированный природой (Проект 2.1); метод бутстрэпа для валидации среднего значения, с опорой и без (Проект 2.2); метод скользящего среднего (перекрестной валидации) (Проект 2.3).

Проект 2.1. Вычисление центра по критерию Минковского

Рассмотрим множество значений признака x_i, $i = 1, 2, \ldots, N$, и положительную величину $m > 1$. Вычислим такое значение a, которое минимизирует критерий Минковского, т. е. сумму m-х степеней расстояний,

$$L_m = |x_1 - a|^m + |x_2 - a|^m + \ldots |x_N - a|^m \qquad (2.7)$$

При $m \neq 2$ не существует общего аналитического решения этой задачи. Есть несколько способов ее численного решения. Один из них — это использование итеративного метода приближения к точке (локального) минимума. Этот метод заключается в пошаговом перемещении в обратном градиенту направлении, как бы «спускаясь» по градиенту (отсюда «метод градиентного спуска»). Альтернативный способ решить поставленную задачу — использовать алгоритм, инспирированный природой. В таких алгоритмах популяция допустимых решений итеративно эволюционирует, причем лучшие из найденных решений запоминаются в соответствии с теми или иными правилами поддержания элиты.

Рассмотрим два метода минимизации L_m:

1) ГС — итерационный градиентный спуск;
2) ИП — итерационный алгоритм, инспирированный природой.

ГС. Градиентный спуск MC_AG

Исследуем свойства критерия L_m из (2.7). Для простоты примем, что $m \geq 1$. Упорядочим все N чисел из X в порядке возрастания, так что $x_1 \leq x_2 \leq \ldots \leq x_N$. Несложно доказать, что, во-первых, рассматриваемый критерий является выпуклой функцией (см. рис. 2.11) и, во-вторых, что оптимальное значение a находится между x_1 и x_N.

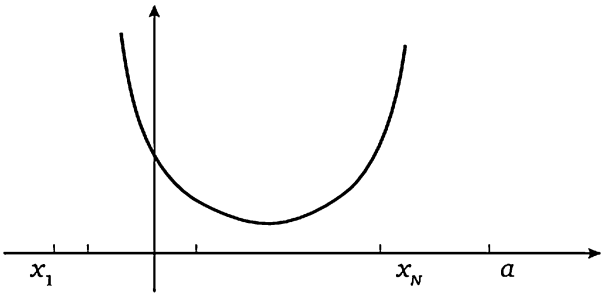


Рис. 2.11. **Выпуклая функция от *a***

Допустим обратное, т. е. что минимум достигается вне заданного интервала, например, при $a > x_N$. это противоречит тому, что $L_m(x_N) < L_m(a)$, потому что $|x_i - x_N| < |x_i - a|$ для любого $i = 1, 2, \ldots, N$, и то же самое верно для m-х степеней модулей. Что касается выпуклости, рассмотрим любое a в интервале между x_1 и x_N. Тогда критерий (2.7) имеет вид:

$$L_m(a) = \sum_{i \in I_+} (a - x_i)^m + \sum_{i \in I_-} (x_i - a)^m, \tag{2.8}$$

где I_+ — это множество индексов i, для которых $a > x_i$, и I_- — множество индексов i, для которых $a \leq x_i$. Производная $L_m(a)$ из (2.8) равна:

$$L'_m(a) = m\left(\sum_{i \in I_+} (a - x_i)^{m-1} - \sum_{i \in I_-} (x_i - a)^{m-1}\right), \tag{2.9}$$

а вторая производная:

$$L''_m(a) = m(m-1)\left(\sum_{i \in I_+} (a - x_i)^{m-2} - \sum_{i \in I_-} (x_i - a)^{m-2}\right).$$

Последнее выражение положительно при любых значениях a, если $m > 1$. Как известно из математического анализа, дифференцируемая функция, у которой вторая производная положительна, является выпуклой. Следовательно, $L_m(a)$ — выпуклая функция.

Тогда минимум достигается вблизи точки x_{i*}, на которой достигается минимум на имеющихся наблюдениях, $L_m(x_i)$ по всем $i = 1, 2, \ldots, N$. Точнее, минимум $L_m(a)$ принадлежит интервалу $(x_{i'}, x_{i''})$, где $x_{i'}$ — то из чисел x_i, которое ближе всего к x_{i*} с левой стороны, т. е. с той, где $L_m(x_i) < L_m(x_{i*})$. Аналогично, $x_{i''}$ — то из чисел x_i, которое ближе всего к x_{i*} с правой стороны, где $L_m(x_i) > L_m(x_{i*})$.

Перечисленные выше свойства позволяют сформулировать следующую версию алгоритма градиентного спуска при $m > 1$.

MC_FD

1. **Инициализация**: примем $a0 = x_{i*}$ и зададимся какой-либо скоростью обучения $\lambda > 0$.
2. **Итерационный шаг**: Вычислим разность $a0 - \lambda L'_m(a0)$ по формуле (2.9) и примем найденное значение за $a1$, если оно попадает в интервал $(x_{i'}, x_{i''})$. В противном случае несколько уменьшим λ, например, на 10 %, и повторим шаг уменьшения до тех пор, пока разность не окажется в интервале $(x_{i'}, x_{i''})$.
3. **Критерий остановки:** Проверим, совпадают ли $a1$ и $a0$ с точностью до заранее определенного порога. Если да, останавливаем процесс и возвращаем $a1$ в качестве оптимального значения a. Если нет, продолжаем вычисления.
4. **Критерий успешности итерации**: Проверим условие $L_m(a1) \leq L_m(a0)$. Если оно верно, присвоим $a0 = a1$ и $L_m(a0) = L_m(a1)$ и перейдем к шагу 2. Если нет, уменьшим λ и перейдем к (2), не меняя $a0$.

ИП. Метод MC_NI, инспирированный природой

Алгоритмы, инспирированные природой, отличаются от классических подходов тем, что на каждом шаге вычислений рассматривают не одно допустимое решение, а целую популяцию допустимых решений. Улучшение решений получается при помощи «случайной» эволюции популяции и поддержки ее «элиты» от одного поколения к последующему. Поскольку решение рассматриваемой задачи — число, а не многомерный вектор, то скорее всего любые случайные изменения будут достаточно быстро вести к оптимуму. Алгоритм MC_NI, реализующий эту идею, довольно хорошо показывает себя в экспериментах.

1. **Определение области допустимых решений.** Определим область A, которой принадлежат допустимые решения, т. е. точки, среди которых находится оптимальное решение или несколько оптимальных решений.

В данном случае это нетрудно сделать. Выше было доказано, что оптимум лежит между минимумом lb и максимумом rb выборки x_i, $i = 1, 2, \ldots, N$. Следовательно, область допустимых решений — это интервал (lb, rb).

2. **Инициализация популяции.** Зададим размер популяции pe. Пусть, например, $pe = 15$. Случайным образом выберем точки s_1, s_2, ..., s_{pe} из допустимой области (lb, rb).

3. **Инициализация элиты.** Оценим значение критерия (2.7) для каждого из элементов популяции и запомним в переменной s_e наилучший из них (элиту), т. е. ту из величин s_k, на которой достигается минимум критерия.

4. **Переход к новому поколению.** Во-первых, добавим случайный Гауссов шум r:

$$s'_k = s_k + \lambda r.$$

Во-вторых, если какое-то из значений выйдет за границы допустимой области A, вернем его обратно, на границу.

5. **Поддержка элиты.** Найдем значения критерия для всех элементов нового поколения, выберем наилучшее и наихудшее значения s_b и s_w, и сравним их с элитой s_e. Если s_b лучше, чем s_e, запомним s_b в s_e. Если s_b и, тем более, s_w хуже, чем s_e, заменим s_w в текущей популяции на s_e.

6. **Условие остановки.** Если число итераций не превышает заранее заданной величины, переходим к шагу 4. В противном случае выдаем элиту s_e как решение задачи.

Эксперименты показывают, что метод градиентного спуска в данной задаче работает быстрее, чем метод, инспирированный природой. Но первый метод работает только при показателе степени $m > 1$, а второй — при любых значениях m.

Проект 2.2. Оценка доверительного интервала среднего значения методом бутстрэп

Файл с данными short.dat приведен в Приложении А.5. Он представляет собой массив 50×3, столбцы которого — выборки из трех разных распределений, описанных в табл. 2.7.

Таблица 2.7

Сводные характеристики столбцов массива short.dat

Тип данных		Нормальное распределение	Двумодальное распределение	Закон Парето
Среднее значение		10.27	16.92	289.74
Стандартное отклонение	Действительное значение	1.76	4.97	914.50
	Деленное на $\sqrt{N}$	0.25	0.70	129.33

Первый столбец — это выборка из Гауссова распределения N(10,2) с математическим ожиданием, равным 10, и стандартным отклонением, равным 2. Второй столбец — выборка из двумодаль-

ного распределения, а третий — из степенно́го распределения. Их гистограммы изображены на левых сторонах рис. 2.15, 2.16. и 2.17. Судя даже по сводным данным из табл. 2.7, среднее значение степенно́го распределения не имеет особого смысла, поскольку оно значительно меньше стандартного отклонения.

Многие статистики могли бы оспорить обоснованность/правильность характеристик из табл, 2.7 не из-за формы распределений, которая действительно смущает по крайней мере в двух случаях из трех, а из-за небольших размеров выборок. Достаточно ли 50 наблюдений для того, чтобы представить всю генеральную совокупность двумя числами? Для решения этой проблемы в математической статистике выработаны методы, основанные на предположении, что все наблюдения выбраны случайным и независимым образом из одного, возможно не известного, но стационарного, распределения. Тогда в достаточно четко определенных ситуациях для таких показателей, как среднее значение, может быть построено свое теоретическое распределение и, следовательно, некоторые доверительные границы для значений показателя. Как правило, доверительные границы определяют по интервалу, в которой попадает 95 % наблюдений из генеральной совокупности, ведь ее распределение известно. Например, если распределение нормально, 95%-ный доверительный интервал вычисляется как среднее значение плюс-минус стандартное отклонение, умноженное на 1.96 и деленное на корень квадратный из числа наблюдений (корень из $N = 50$ равен 7.07). Для первой колонки данных теоретически обоснованный доверительный интервал имеет границы $10 \pm 1.96 \cdot 2/7.07 = 10 \pm 0.55$, т. е. (9.45, 10.55), при условии, что настоящие параметры распределения известны, или $10.27 \pm 1.96 \cdot 1.76/7.07 = 10.27 \pm 0.49$, т. е. (9.78, 10.76) для наблюдаемых параметров из табл. 2.7. Разница между построенными интервалами не велика, особенно, если учитывать, что понятие доверительного интервала само не очень-то понятно. В математической статистике используется так называемое распределение Стьюдента, чтобы компенсировать использование выборочного значения стандартного отклонения вместо точного. Если число наблюдений больше, чем несколько сотен, распределение Стьюдента мало отличается от нормального.

Во многих практических приложениях форма распределения неизвестна, к тому же, оно не обязательно стационарно. В таких случаях ценность теоретических доверительных интервалов не велика. Поэтому возникает закономерный вопрос: можно ли найти какие-нибудь доверительные границы вычислительным образом, используя только имеющуюся выборку, не используя сомнительные допущения. Разработано несколько подходов к вычислительной валидации показателей, построенных по выборке. Один из самых популярных — это бутстрэп. Ниже будут описаны две версии этого

метода: с опорой и без (*pivotal* и *non-pivotal*), как они определены в [38].

Бутстрэп основан на некотором количестве, например, 1000, случайных испытаний. Каждое **испытание** состоит из N случайных выборов объектов из выборки с возвращением, где N — это количество объектов в исходном множестве. Поскольку проводится выбор с возвращением, некоторые объекты могут быть выбраны несколько раз, а другие останутся не выбранными ни разу. Нетрудно убедиться, что в среднем $(e-1)/e = 63.2\ \%$ всех объектов попадут в выборку одного испытания (здесь $e = 2.7182818$ — знаменитое «математическое» число, основание натурального логарифма). Действительно, при каждом случайном выборе из множества размером N, вероятность не быть выбранным для любого объекта составляет $1 - 1/N$. Поэтому вероятность быть невыбранным при N независимых выборах равна $(1 - 1/N)^N \approx 1/e$ по так называемому «замечательному пределу» из математического анализа. А $1/e = 1/2.71828 \approx 36.8\ \%$ от общего числа объектов. Пример случайной независимой выборки из 15 объектов с возвращением: 8, 11, 7, 5, 3, 3, 11, 5, 9, 3, 11, 6, 13, 13, 9. Как видно, некоторые объекты — 1, 2, 4, 10, 12, 14, 15 (всего 6/15 = 40 %) — в нее не попали, а некоторые попали в выборку несколько раз.

Выборка, полученная в одном испытании, определяет выборочную таблицу данных, в которой N строк соответствуют элементам выборки, причем каждая строка взята из исходной таблицы данных — та, что соответствует данному элементу выборки. Совпадающие объектам соответствуют одинаковые строки. Затем рассматриваемый метод, в данном случае — вычисление среднего, применяется к данным рассматриваемого испытания; в результате получаем величину среднего для этого испытания. После 1000 или 5000 испытаний получается 1000 или 5000 выборочных оценок среднего. Эти-то величины и используются для получения доверительных границ.

В MATLAB имеется команда bootstrap, с помощью которой можно сгенерировать оценки среднего в любом заданном числе испытаний. Поскольку метод применим не только к валидации среднего, но и к валидации любых других результатов анализа данных, мы приведем команды MATLAB, порождающие 2000 испытаний для любого метода — они достаточно просты. Примем, что рассматриваемый признак обозначен через x, а n обозначает число объектов в рассматриваемой выборке. Например, команда >> n = 45; делает n равным 45.

```
>> r = ceil(n*rand(n,2000)); % создает матрицу n × 2000 случай-
ных индексов; столбец — испытание
>>ξρ = x(r); % формирует матрицу величин x, соответствующую
индексам матрицы r
>>mr = mean(xr); % вектор средних на 2000 испытаниях
```

Справа на рис. 2.12—2.14 показаны распределения среднего значения, полученные методом бутстрэпа, для всех трех типов данных после 1000 испытаний.

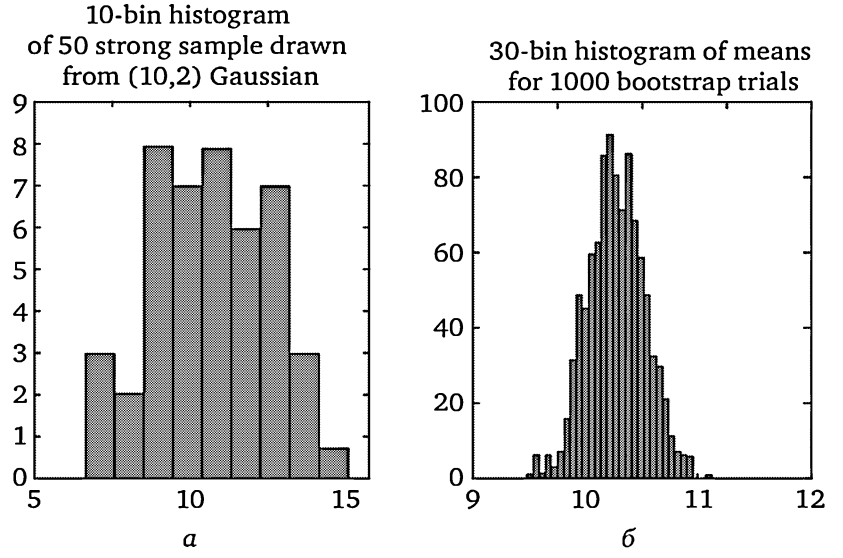


Рис. 2.12. **Гистограммы выборки 50 наблюдений из Гауссова распределения (*а*) и ее среднего значения, рассчитанного по методу бутстрэпа (*б*):** все значения среднего попадают между 9.7 и 11.1

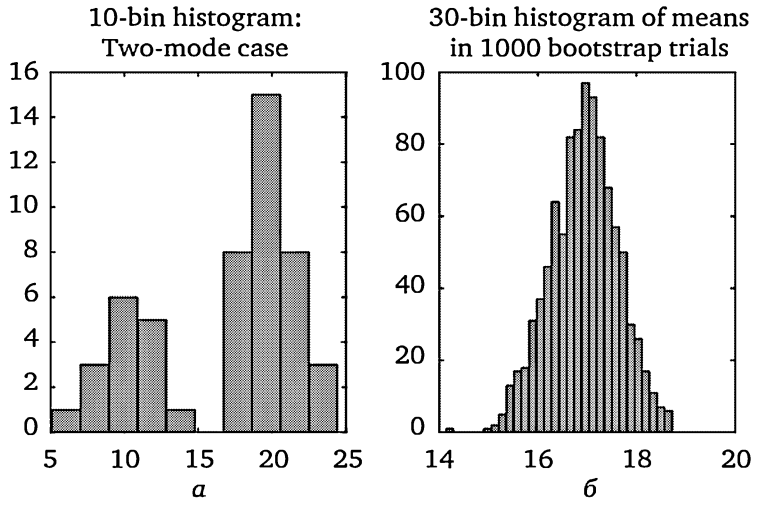


Рис. 2.13. **Гистограммы выборки из 50 наблюдений из двухмодального распределения (*а*) и 1000 испытаний для среднего значения по методу бутстрэпа (*б*)**

Метод валидации с опорой основан на предположении, что распределение значений средних значений бутстрэпа является Гаус-

совым. Это значит, что, имея оценку среднего, m_b, и стандартного отклонения s_b этого распределения, можно воспользоваться обычной «теоретической» формулой для нахождения 95%-ного доверительного интервала. Согласно теории нормального распределения, центральный интервал, покрывающий 95 % распределения, имеет центр в точке m_b, а границы — на расстоянии $1.96s_b$ от него, влево и вправо. В нашей задаче доверительный интервал это $m_b \pm$ $\pm 1.96 \cdot s_b = 10.24 \pm 1.96 \cdot 0.24 = 10.24 \pm 0.47$, т. е. интервал между 9.77 и 10.71. Этот результат близок к результату, полученному в предположении Гауссова распределения. Это не удивительно, поскольку в этом случае распределение действительно Гауссово.

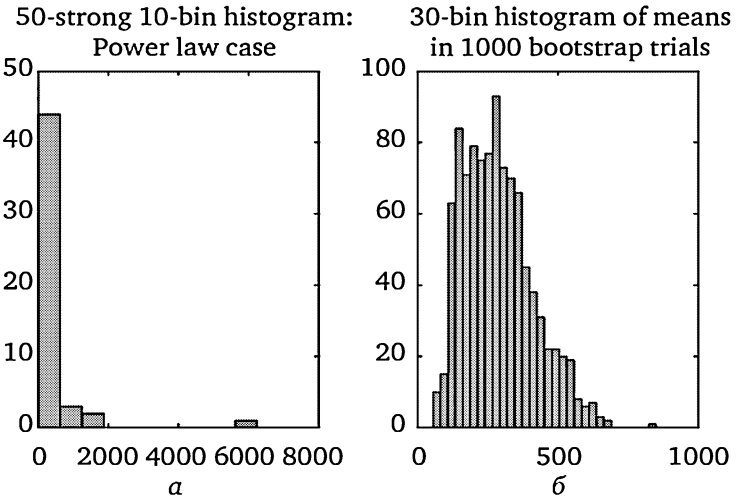


Рис. 2.14. **Гистограммы 50-элементной выборки степенного закона (*а*) и средних значений её 1000 испытаний по методу бутстрэпа (*б*)**

Безопорный метод бутстрэпа не использует никаких допущений о характере распределения средних значений бутстрэпа. Доверительные границы определяются непосредственно по полученному распределению средних значений путем отрезания 2.5%-ных нижнего и верхнего квантиля. Это делается так: весь список средних значений бутстрэпа сортируют в порядке возрастания, после чего отделяют нижние и верхние 2.5 % значений. Для случая 1000 испытаний для этого надо взять соответственно 26-ю и 975-ю компоненты отсортированного списка. В нашем случае доверительный интервал, построенный таким способом, лежит между 9.78 и 10.70. Этот результат очень близок к найденным раньше границам 95%-ного доверительного интервала для математического ожидания первой выборки.

Существует теоретическое доказательство того, что в случае, когда исходное распределение нормально, бутстрэп позволяет получить более узкие значения стандартного отклонения, Efron (1993).

Действительно, в табл. 2.8 средние значения почти одинаковы, а стандартное отклонение слегка уменьшилось.

Таблица 2.8

Общие характеристики результатов бутстрэпа (1000 испытаний для данных из массива short.dat)

Тип данных		Нормальное	Двумодальное	Степенной закон
Среднее значение		10.27	16.94	287.54
Стандартное отклонение	Исходная выборка	0.25	0.70	129.33
	Бутстрэп значение	0.25	0.69	124.38
	Среднее значение, %	2.46	4.05	43.26

К сожалению, бутстрэп не столь полезен для анализа двух других распределений. Он позволяет найти довольно точные доверительные интервалы для среднего значения как по двумодальному распределению, так и по закону Парето. Однако отыскание математического ожидания обоих этих распределений на практике считается бессмысленным. В первом случае смешиваются две разнородные группы, во втором — скрывается еще более глубинная неоднородность распределения. По-видимому, требуются другие методы, такие как кластерный анализ, для того чтобы сформировать более однородные множества наблюдений.

Читателю предлагается построить оценки 95 % доверительного интервала, с опорой и без, для двух оставшихся распределений из файла short.dat (двумодальное распределение и степенной закон).

Проект 2.3. Перекрестная валидация (скользящий контроль)

Ещё один метод валидации использует идею случайного разбиения данных на две части, обучающую и тестовую. Результаты, полученные на обучающем множестве, применяют к тестовому множеству и сравнивают с тем, что известно для тестового множества. Для того, чтобы каждый объект попадал в обучающее и тестовое множества с одинаковой частотой, пользуются специально разработанным методом *перекрестной валидации*, по-русски называемым иногда методом *скользящего контроля*.

Так называемая K-частная перекрестная валидация организована следующим образом. Случайным образом разбиваем множество объектов на K частей $Q(k)$ одинакового размера[1], $k = 1, \ldots, K$. Как правило, K выбирают равным 2, 5 или 10. В цикле по k каждая часть

[1] Это можно сделать, начав с пустых множеств $Q(k)$, повторно в цикле по $k = 1{:}K$ случайно выбирая из множества объект (без возвращения) и помещая его в $Q(k)$; процесс заканчивается, когда не остается нераспределенных объектов.

$Q(k)$ используется как тестовое множество, а объединение остальных образует обучающее множество. Рассматриваемый метод анализа данных применяют к обучающему множеству (стадия обучения), а результат применяют к тестовой выборке. Средняя оценка по всем тестовым множествам составляет оценку качества метода по K-частной перекрестной валидации.

Случай, когда K равно числу объектов N, особенно популярен. Ранее он назывался джек-найф (*jack-knife* — складной нож), но сейчас чаще пользуются названием «выставлять по одному» (*leave-one-out*). Это название отражает суть метода: проводится N обучений анализируемого метода на множествах, полученных исключением из X ровно одного объекта, каждый раз — другого.

Применим метод 10-частной перекрестной проверки к задаче оценки математического ожидания вышерассмотренных трех наборов данных short.dat. Для начала разобьем множество из 50 объектов на 10 непересекающихся классов по пять объектов каждый. Важно, чтобы объекты попадали в классы разбиения случайным образом. Это может быть достигнуто несколькими способами. Например, сначала инициализируются 10 пустых классов. Затем объекты в случайном порядке один за другим помещаются в следующий класс. Другой способ: случайно перемешиваем все индексы объектов и разделяем перемешанное на 10 частей, в каждой по 5 объектов. Для каждого класса $Q(k)$ (k = 1, 2, ..., 10), находим среднее значение и стандартное отклонение на всех остальных 45 объектах. Найденные 10 средних и стандартных отклонений усредняются.

Результаты представлены в табл. 2.9. Показатели, рассчитанные для исходного распределения и с использованием 10-частной перекрестной валидации, схожи. Означает ли это, что в использовании этого метода нет надобности? В данном случае — возможно, но когда речь идет о доверии к результатам более сложных методов анализа данных, результаты могут различаться существенным образом. Кроме того, отметим, что квадратичные отклонения на десяти тестовых множествах для Гауссова и двумодального распределения близки друг к другу, а для степенного закона — сильно различаются и варьируют от 391.60 до 2471.03.

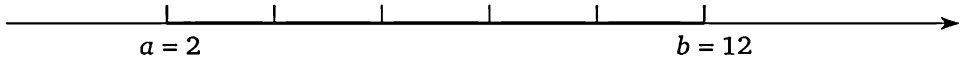


Рис. 2.15. **Интервал изменений признака, [2, 12], разделенный на 5 бинов**

Таблица 2.9

Квадратичные отклонения от математического ожидания, построенные для исходного множества и с использованием 10-частной перекрестной валидации

Тип данных		Нормальное	Двумодальное	Степенной закон
Стандартное отклонение	На множестве	1.94	5.27	1744.31
	10-частная валидация	1.94	5.27	1649.98

Вопрос 2.13. Каков размер бина на рис. 2.15?

Ответ. Равен 2.

Вопрос 2.14. Рассмотрим признак x, изменяющийся в пределах между 1 и 10. Разделим размах x на 9 бинов (в этом случае длина бина — 1). Частоты x в бинах в порядке нумерации равны: 10, 20, 10, 20, 30, 20, 40, 20, 30. Ответьте на следующиевопросы.

1. Сколько всего наблюдений x?
2. Что можно сказать о медиане x?
3. Посчитайте минимальную и максимальную оценки среднего значения x.
4. Что можно сказать о 20 % квантилях x?
5. Как выглядит распределение x, если число бинов равно 3? Какова качественная дисперсия (индекс Джини) для этого распределения?

Ответ.

1. Всего 200 наблюдений.
2. Медиана лежит между 100-м и 101-м значением в упорядоченном списке, т. е. в 6-м бине, следовательно, между 6 и 7.
3. Минимальная оценка математического ожидания рассчитывается по минимальным значениям признака в бинах: $(1 \cdot 10 + 2 \cdot 20 + 3 \cdot 10 + 4 \cdot 20 + 5 \cdot 30 + 6 \cdot 20 + 7 \cdot 40 + 8 \cdot 20 + 9 \cdot 30)/200 = 5.7$.

Максимальная оценка рассчитывается по той же формуле, только значения признака во всех бинах увеличиваются на 1. Получаем: $5.7 + 1 = 6.7$.

4. 20 % от 200 равно 40. Это означает, что 20%-ный квантиль с левого края — это 4, 20%-ный квантиль с правого конца попадает в 8-й бин, следовательно, лежит между 8 и 9.
5. Данное распределение для случая трех бинов будет 40, 70, 90 или, в относительных частотах, 0.2, 0.35, 0.45. Следовательно, индекс Джини равен $1 - 0.2^2 - 0.35^2 - 0.45^2 = 0.635$.

Вопрос 2.15. Центральные значения. Из 100 покупателей новогодних подарков 50 потратили по \$60, 20 потратили по \$100 и 30 — по \$150 каждый. Найдите (i) среднюю, (ii) медианную и (iii) модальную траты.

Подсказка. Как, учитывая то, что покупатели объединены в три группы, сделать вычисления центров более эффективными?

Ответ. **Среднее:** Для начала найдем доли покупателей, потративших по \$60, \$100 и \$150 каждый. Получим 0.5, 0.2 и 0.3, соответ-

ственно. Среднее может быть вычислено так: сложим все затраты, взвесив их соответствующими пропорциями. Получим: $c = 60 \cdot 0.5 + + 100 \cdot 0.2 + 150 \cdot 0.3 = 30.0 + 20.0 + 45.0 = 95$.

Медиана: согласно определению, медиана 100 чисел — это значение посредине между 50-м и 51-м объектом в отсортированном списке. В нашем случае, на этих местах стоят 60 и 100, поэтому медиана затрат — $80.

Мода: Модальное значение — наиболее вероятное, т. е. 60.

Вопрос 2.16. Рассмотрим два геологических разреза, для одного из которых имеется 7 образцов, а для другого — 5. Содержание минералов в образцах разреза A описано вектором a = (7.6, 11.1, 6.8, 9.8, 4.9, 6.1, 15.1), а в разрезе B — b = (4.7, 6.4, 4.1, 3.7, 3.9). Среднее содержание минералов в A составляет 8.77 и в B — 4.56. Протестируйте гипотезу «содержание минералов в разрезе A больше, чем в разрезе B» на 95%-ном уровне доверия с использованием бутстрэпа.

Ответ. Поскольку множества маленькие, число испытаний должно быть выбрано не слишком большим. При 200 испытаниях 95%-ный доверительный интервал образован 6-м и 195-м значением в списке отсортированных средних значений бутстрэпа. В нашем случае это интервал (6.66, 11.09) для A и (3.82, 5.44) для B. Поскольку все элементы первого интервала больше всех элементов второго, гипотеза может считаться подтвержденной. (Данное решение не совсем корректно, так как в утверждении «разрез A богаче разреза B» есть неточность. Например, можно считать, что A богаче B на 95%-ом уровне доверия, если случайная выборка из A богаче случайной выборки из B в 95 % случаев. Тогда 95%-ного интервала не достаточно, поскольку он покрывает только $0.95 \cdot 0.95 = 90.25$ % всех возможных пар средних значений бутстрэпа.) Посмотрим на минимальные и максимальные средние значения бутстрэпа. Размах средних значений равен (6.33, 11.94) для A и (3.82, 5.82) для B. Интервалы не пересекаются, один лежит строго правее другого, что означает, что гипотеза доказана, даже и при ограничениях метода.

Вопрос 2.17. Распределения признаков в данных об ирисах. Рассмотрите гистограммы признаков из данных об ирисах и покажите, что два признака бимодальны.

Ответ. Используем команду MATLAB

```
>> for k = 1:4;subplot(2,2,k);hist(iris(:,k),15);end;
% 15 здесь – число бинов
```

и получим рисунок типа рис. 2.16. Очевидно, что третий и четвертый признаки бимодальны.

Вопрос 2.18. Студент решил провести вычислительный эксперимент. Он случайным образом сгенерировал относительные частоты категорий для качественного признака с тремя категориями.

Сначала он решил сгенерировать три случайных числа в интервале (0, 1) и затем нормировать их общей суммой трех чисел, чтобы в результате получить сумму 1. Если, например, сгенерированы 0.7116, 0.1295, 0.6598, то после деления на их сумму 1.5009 получатся величины 0.4741, 0.0863, 0.4396, дающие в сумме 1. Правильны ли действия студента?

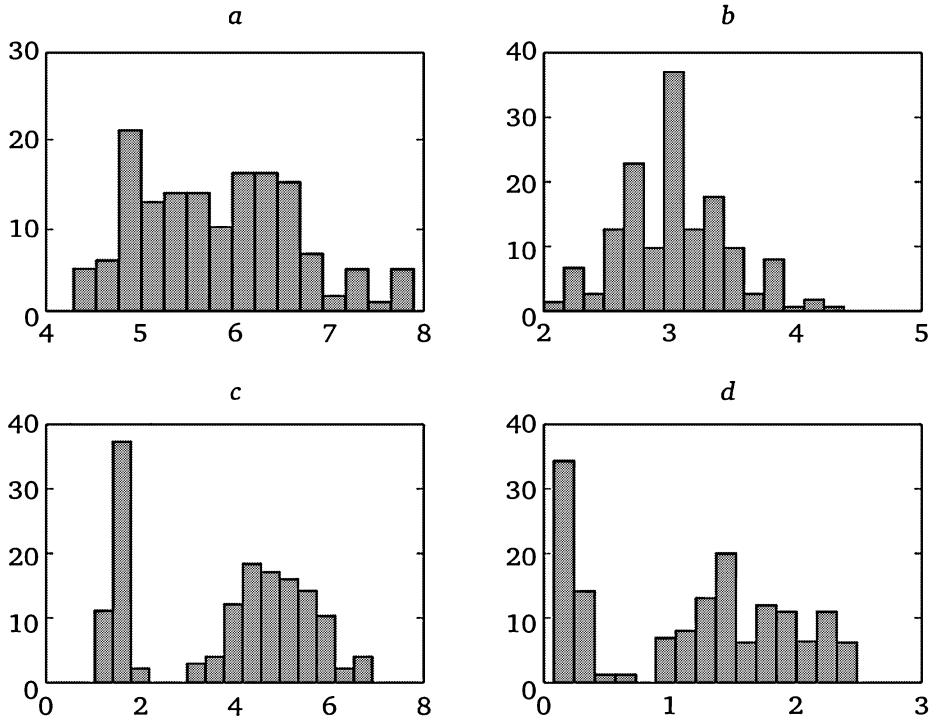


Рис. 2.16. **Гистограммы всех четырех признаков в данных Ирисы: (*c*) и (*d*) явно бимодальны**

Ответ. Нет, не совсем, поскольку создается сдвиг в сторону равных частот (см. рис. 2.17). На рис. 2.17, *а*, представлено распределение первой из пары частот, найденных описанным методом: два случайно сгенерированных числа делятся на их сумму. Распределение далеко от равномерного распределения, представленного на рис. 2.17, *б* (можете объяснить разницу?).

Более удачным был бы, например, следующий метод генерации случайной тройки чисел.

Сначала генерируются два случайных числа; их сортируют по возрастанию. Добавим 0 и 1 к выборке: $r_0 = 0 < r_1 < r_2 < r_3 = 1$. Затем определим искомые частоты как разности между соседними элементами, $p_k = r_k - r_{k-1}$ ($k = 1,2,3$). Например, если сначала были сгенерированы величины 0.8775, 0.5658, искомые частоты будут

определены как $p_1 = 0.5658$, $p_2 = 0.8775 - 0.5658 = 0.3117$, и $p_3 = 1 - 0.8775 = 0.1225$. При этом все получаемые частоты так же равномерно распределены, как и исходные две. Этот метод легко распространить на любое числа категорий.

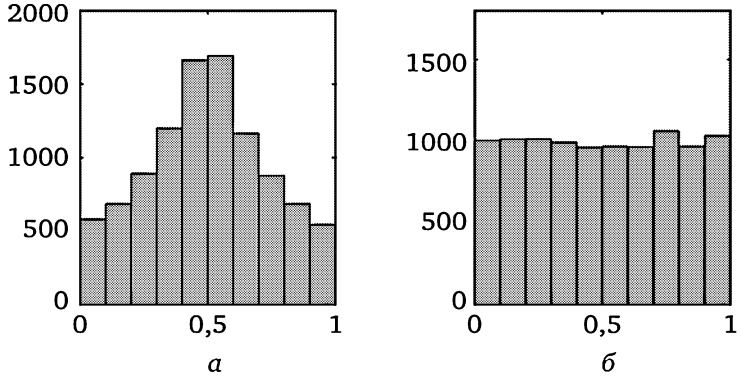


Рис. 2.17. **Гистограммы выборок из 100 000 элементов:**
а — первый элемент из случайной пары после деления на сумму двух величин; *б* — равномерная случайная величина

Кстати говоря

4. Категоризация (приписывание категорий отдельным наблюдениям)

4.1. Из телефонного разговора на работе:
— Как вас зовут?
— Славик.
— А отчество?
— С такой зарплатой — просто Славик...

4.2. У мудреца спросили:
— Как называется жена, которая всегда знает где ее муж?
— Вдова, — ответил мудрец.

4.3. — Как тебя зовут?
— Как Васю.
— Как какого Васю?
— Как любого Васю.

4.4. Студент: Я не думаю, что заслуживаю двойку.
Профессор: Я тоже так не думаю, но это самая низкая оценка, которую мне разрешено ставить.

4.5. Одного профессора права спросили:
— Что такое бигамия, и как она карается?
— Бигамия, — ответил ученый, — это наличие двух жен, которое карается наличием двух тещ...

4.6. Два медведя разговаривают:
— Я вчера кого-то съел, кого, не знаю — спинка черная, пузо белое, лапки грязные...
— Это же дачник!!!

4.7. — Доктор, я плохо выговариваю букву «Ч».

— Ну, что вы. Вы очень хорошо выговариваете букву «Ч».

— Ну, не чавчем.

4.8. — Простите, вы не видели моего брата-близнеца?

— Так вы уже спрашивали.

4.9. СМИ: «В Москве проходит аукцион винтажных вещей от люксовых брендов».

— А если попроще?

— Попроще? Старьё сбывают.

4.10. Экзамен по уголовному праву:

— Вы можете сказать, что такое обман?

— Это произойдет, профессор, если вы меня провалите.

— Каким образом, поясните.

— Согласно уголовному кодексу, обман совершает тот, кто, пользуясь незнанием другого лица, причиняет этому лицу ущерб.

4.11. После посещения выставки известного художника, журналист оставил в газете замечание, «Выставка могла быть и лучше». Оскорбленный художник потребовал письменного опровержения. На следующий день в газете появилось: «Опровержение: выставка могла быть и хуже». Художник в бешенстве потребовал нового опровержения. На третий день в газете: «Опровержение: Выставка хуже быть не могла».

4.12. — Посоветуйте, что купить жене на День рождения?

— Может, спросим у нее самой?

— Нет! Таких денег у меня нет!

4.13. Человек на приёме у врача с переломом ноги, вывихом руки и выбитой челюстью.

— В аварию попали?

— Нет... Чихнул в шкафу...

4.14. — Профессор, можете дать прогноз наших перспектив в экономике?

— Могу. Могу дать оптимистический, могу пессимистический и могу реальный.

— И какой реальный?

— Ну, реальный — в два раза хуже пессимистического.

4.15. Пришла в больницу девушка проведать своего парня. Увидев возле палаты женщину в белом халате, она обратилась к ней:

— Я могу увидеть больного?

— А кем вы ему приходитесь?

— Я его сестра.

— Рада познакомиться: я его мать.

4.16. — Как Вы догадались, что задержанный — вор?

— По шапке.

— Что, на нем горела?

— Нет, просто это была моя шапка.

4.17. — Мужчина! Почему вы в брюках и рубашке, когда все в плавках и купальниках?

— Видите ли, все вокруг купаются, а я тону!

Тема 3
ДВУМЕРНЫЙ АНАЛИЗ: СУММАРИЗАЦИЯ И КОРРЕЛЯЦИЯ ДВУХ ПРИЗНАКОВ

В этой теме приводится несколько важных характеристик суммаризации и корреляции двух признаков, а также некоторые способы их использования, такие как:

а) поле рассеяния (*scatter-plot*), линейная регрессия и коэффициент корреляции двух количественных признаков;

б) бокс-плот, табличная регрессия, корреляционное отношение, декомпозиция разброса количественного признака и классификатор по ближайшему соседу для случая смешанных шкал;

в) таблица сопряженности признаков, коэффициент Кетле, статистическая независимость и хи-квадрат Пирсона для двух номинальных признаков, для которого показано, что он характеризует уровень связи между признаками, а не только соответствие гипотезе о статистической независимости, как это обычно считается.

3.1. Введение

Анализ двух признаков на одном и том же наборе объектов может представлять интерес тогда, когда признаки связаны, т. е. изменяются более или менее одновременно. Такая связь — если она в самом деле наблюдается — может быть использована в различных целях, среди которых обычно различают следующие две:

(i) прогнозирование значений одного признака по значениям другого;

(ii) добавление новой связи к знанию о предметной области через ее интерпретацию в терминах данной области.

Признак, значение которого предсказывается, принято называть целевым, выходным или прогнозируемым, а второй признак — входным или предиктором. Примеры задач типа (i): прогнозирование компьютерных атак определенного типа или числа школ в малом городе с известным числом жителей. Кто-то может спросить, зачем, собственно, волноваться: ведь все значения признаков уже находятся в файле! Дело в том, что в задаче прогноза имеющиеся данные — всего лишь выборка из большой популяции, используемая как полигон для формирования решающего правила для про-

гнозирования целевых признаков на других, не попавших в данное множество, объектах. Обычно входной признак на этих других объектах известен или легко измеряем, в то время как целевой — нет. Что касается задачи (ii), то данные представляют собой простые эмпирические факты, не обязательно достойные внимания, до тех пор, пока они не обобщены в виде правил для принятия решений.

Математическая структура и визуализация контекста анализа связи между признаками зависят от шкал измерения признаков. Естественно рассматривать следующие случаи:

1) оба признака количественные;
2) один признак категоризованный, другой количественный;
3) оба признака категоризованные.

Рассмотрим эти случаи последовательно.

3.2. Два количественных признака: линейная регрессия и вокруг

3.2.1. Поле рассеяния, линейная регрессия и коэффициент корреляции

В случае, когда оба признака количественные, используются следующие три понятия: поле рассеяния (*scatter-plot*), корреляция и регрессия. Рассмотрим их на примере двух признаков из данных о прибрежных городах: численность населения — Нас и число начальных школ — Нш. Данные взяты из табл. 1.5 (ниже — данные для 4 из 45 городов; население измерено в тысячах):

Город	Нас (x)	Нш (y)	(x, y)-точка
Tavistock	10.222	5	(10.222,5)
Bodmin	12.553	5	(12.553,5)
Saltash	14.139	4	(14.139,4)
Brixham	15.865	7	(15.865,7)

Поле рассеяния или, по-английски, *scatter-plot* — это представление объектов в виде двумерных точек на координатной плоскости каких-либо двух признаков. На рис. 3.1, *а* показано поле рассеяния двух признаков торговых городов: населения — Нас (ось x) и количества начальных школ — Нш (ось y); на рис. 3.1, *б* добавлена линия регрессии Нш по Нас.

Обратим внимание на то, что для удобства мы изменили шкалу измерения признака Нас — разделили его значения на 1000, приблизив тем самым значения x и y друг к другу.

Предположим, что эти признаки связаны линейным уравнением $y = ax + b$, где a и b — константы «наклона» и «сдвига», соответственно. Действительно, число школ должно в какой-то степени зависеть

от числа детей; а число детей, в свою очередь, — от числа жителей в городе.

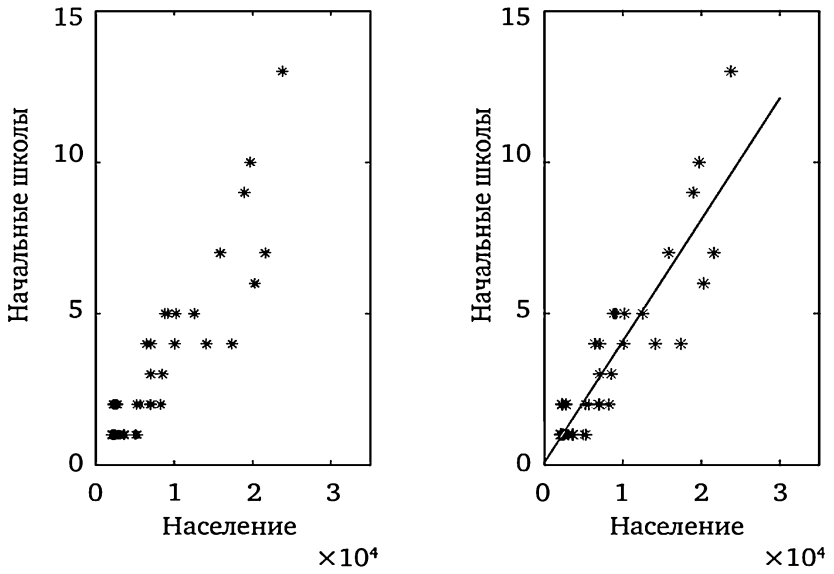


Рис. 3.1. **Поле рассеяния признаков Нас и Нш (*а*); линия регрессии Нш по Нас (*б*)**

Это уравнение называется *линейной регрессией* y по x. Очевидно, что многие другие связи невозможно описать подобной простой формулой, поскольку обычно на y влияют и другие факторы, такие как размеры школ, возраст населения и пр. Если бы одно уравнение подходило ко всем 45 городам, это было бы настоящим чудом — в реальности же ошибки в таком равенстве будут всегда. Возможные невязки в уравнении на тех или иных городах могут быть сведены в суммарную ошибку. Задача состоит в том, чтобы выбрать наклон a и сдвиг b таким образом, чтобы суммарная ошибка, измеряемая суммой квадратов невязок по всем 45 городам, была минимальной. Метод решения этой задачи описан далее в параграфе Ф.3.2.

Если параметры уравнения линейной регрессии оценены, уместно проверить его адекватность, т. е. соответствие имеющимся данным. Адекватная регрессия может быть использована как при прогнозировании, скажем, для целей планирования, задача (i), так и при описании, задача (ii).

В теории линейных регрессионных уравнений Гальтона — Пирсона (см. параграф Ф. 3.2) широко применяется понятие коэффициента корреляции, отражающего уровень «линейной связи» между двумя признаками. Его квадрат, называемый коэффициентом детерминации, может быть использован для быстрой оценки уровня адекватности уравнения линейной регрессии: он характеризует долю дисперсии y, объясненную его регрессией через x. Коэффици-

ент корреляции находится в интервале между —1 и 1, и если его величина близка к 1 или –1, то это означает, что признаки связаны линейным уравнением с точностью до малых ошибок. Коэффициент корреляции признаков Нас и Нш равен 0.909. В физике или химии такое высокое значение коэффициента корреляции — распространенное явление; в социальных науках — нет, так что рассматриваемый пример — скорее исключение, чем правило.

Многие другие признаки в данных о городах, такие как число почтовых отделений или докторов, также сильно связаны с признаком Нас, но, например, наличие фермерского рынка уже с Нас никак не связано. Низкое значение коэффициента корреляции, ниже 0.15, говорит о том, что размер города не является здесь существенным: вероятность наличия фермерского рынка в маленьком городе такая же, как и в большом.

Совсем низкое или нулевое значение коэффициента корреляции не всегда означает отсутствие взаимосвязи вообще. Речь идет только об отсутствии именно *линейной* связи. Нулевой коэффициент корреляции может соответствовать другому, более тонкому, типу функциональной зависимости. На рис. 3.2 представлены 3 различных поля рассеяния при нулевой корреляции в данных. Только один из них, тот, что слева, на самом деле свидетельствует о том, что между x и y нет связи, т. е. знание значения одного признака никак не помогает в прогнозе значения другого. Каждый из двух других случаев показывает довольно высокую степень связи x и y. В частности, в центре — график квадратичной зависимости, а справа — случай, когда совокупность объектов разнородна — она состоит из двух частей, таких что в каждой признаки связаны линейно, но связи взаимно противоположны.

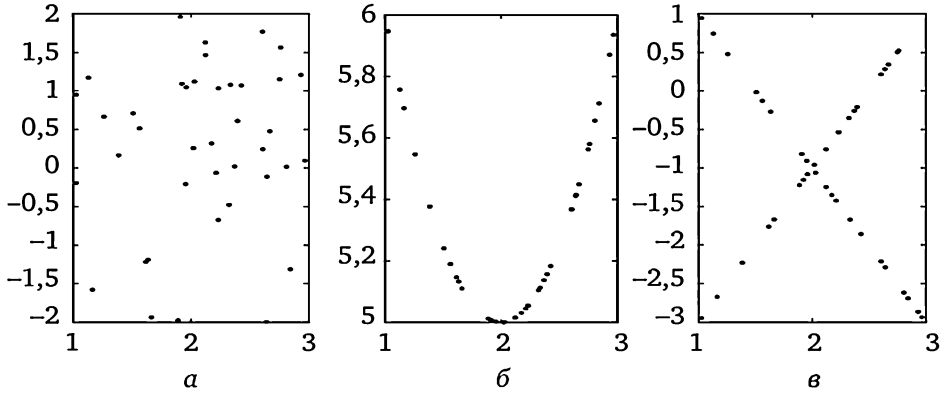


Рис. 3.2. **Типы полей рассеяния, соответствующих нулевому или почти нулевому значению коэффициента корреляции:**
а — отсутствие связи между x и y; *б* — неслучайная квадратичная зависимость $y = (x - 2)^2 + 5$; *в* — два симметричных линейных соотношения, $y = 2x - 5$ и $y = -2x + 3$, каждое из которых содержит ровно половину всех объектов

Регрессионное уравнение для признаков Нас и Нш, заданное формулами (3.4)—(3.6) ниже, будет иметь вид:

$$\text{Нш} = 0.401 \cdot \text{Нас} + 0.072, \qquad (3.1)$$

где население Нас выражено в тысячах человек, чтобы сделать наклон в тысячу раз больше того, как если бы он был выражен в абсолютных величинах. Наклон показывает, насколько изменится значение целевого признака при изменении входного признака на 1. Поскольку значения целевого признака выражены в целых числах, величину наклона можно трактовать следующим образом: рост населения в городе на 2.5 тысячи человек приведет, в среднем, к строительству одной дополнительной начальной школы.

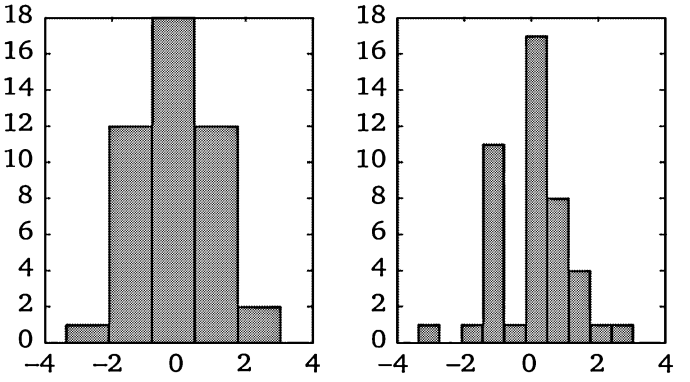


Рис. 3.3. **Гистограммы остатков, т. е. разностей между наблюдаемыми значениями НШк и значениями, вычисленными по Нас с использованием уравнения (2.1), с 5 столбцами (слева) и 10 столбцами (справа):** впадины на гистограмме справа могут быть связаны с тем, что выборка из 45 городов слишком мала, чтобы равномерно заполнить 10 столбцов

3.2.2. Анализ степени адекватности уравнения регрессии

Функция регрессии, построенная на некотором наборе данных, должна быть проверена на адекватность. Рассмотрим три способа проверки адекватности:

1) доля дисперсии целевого признака, учтённая в уравнении регрессии, коэффициент детерминации: чем его значение больше, тем точнее регрессионное уравнение описывает связь признаков;

2) доверительные интервалы для параметров регрессии — их диапазоны могут дать представление об уровне устойчивости регрессии;

3) непосредственное тестирование точности прогноза как на данных, использованных для построения регрессии, так и на тех, что не использовались для ее построения.

Рабочий пример 3.1

Коэффициент детерминации

Рассмотрим целевой признак Нш и входной признак Нас в данных по прибрежным городам (см. рис. 3.1). Коэффициент корреляции между ними равен 0.909. Коэффициент детерминации в случае линейной регрессии равен квадрату коэффициента корреляции, т. е. $0.909^2 = 0.826$. Это показывает, что использование линейной связи между Нш и Нас снижает разброс значений Нш на 82.6 % — довольно высокое значение.

Самостоятельная работа

3.1. Найдите коэффициенты корреляции и детерминации для признаков Длина и Ширина лепестка по данным об ирисах (см. табл 1.2). Дайте интерпретацию величины коэффициента детерминации.

3.2. Найдите распределение, энтропию и индекс Джини для признака Бол в данных о малых городах английского побережья (см. табл 1.5).

Если величина коэффициента детерминации мала, гипотезу о линейной зависимости признаков все же отвергнуть из-за этого одного факта нельзя. Это зависит от распределения остатков регрессии, т. е. от разности наблюдаемых значений Нш и значений, вычисленных по Нас с помощью уравнения (3.1). Это распределение должно быть Гауссовым или близким к таковому, так чтобы применение принципа максимального правдоподобия приводило к соответствующим выводам. Распределение, о котором идет речь, представлено на рис. 3.3. Оно действительно напоминает Гауссово распределение на гистограмме с 5 столбиками. Гистограмма с 10 столбиками имеет меньше сходства из-за наличия впадин — возможно, выборка слишком мала для такого уровня детализации: в среднем, только 4—5 объектов попадают в каждый из десяти столбиков.

Более простое тестирование корректности может быть проведено без обращения к какой-либо статистической теории вообще, а с использованием только лишь вычислительных средств. Бутстрэп — это способ диверсификации оценки интересующего нас параметра через вычисление этой оценки на случайных подвыборках из множества наблюденных данных.

Рабочий пример 3.2

Тестирование адекватности с помощью бутстрэпа

Рассмотрим линейную регрессию Нш на Нас в уравнении (3.1). Насколько устойчивы найденные значения наклона и сдвига при возникновении изменений в выборке? Именно это проверяется с помощью процедуры бутстрэп. Одно испытание, согласно этой процедуре, включает в себя три этапа.

1. Случайно выбрать, с повторением, столько объектов, сколько их в выборке — 45 в нашем случае. Например, случайно выбрана следующая последовательность индексов наших 45 городов:

r = {26, 17, 36, 11, 29, 39, 32, 25, 27, 26, 29, 4, 4, 33, 10, 1, 5, 45, 17, 16, 13, 5, 42, 43, 28, 26, 35, 2, 37, 44, 6, 39, 33, 21, 15, 11, 33, 1, 44, 30, 26, 25, 5, 37, 24}.

Некоторые индексы неоднократно попали в выборку, например, 26 — 4 раза, в то время как другие вовсе не попали в нее; таких значений всего 16, как например, 3, 7, 8. Доля отсутствующих индексов равна 16/45 = 0.356, что довольно близко к теоретической оценке $1/e$ = 0.3679, полученной в Проекте 2.2.

2. Найдем 45-мерные векторы значений признаков Нас и Нш на последовательности объектов r из пункта 1.

3. Найдем величины наклона и сдвига для этих новых версий признаков Нас и Нш.

Шаги вычислений в MATLAB те же, что и в Проекте 2.2. После 400 испытаний получим по 400 величин наклона и сдвига, 20-бинные гистограммы распределений которых представлены на рис. 3.4, *a* и *b*, соответственно. Гистограммы *c* и *d* получены в результате 4000 испытаний. Легко заметен сглаживающий эффект увеличения числа испытаний: при 4000 испытаний форма гистограмм очевидно Гауссова.

Процедура бутстрэп порождает разнообразие решений, необходимое для оценки доверия к средним значениям наклона и сдвига. Поэтому для каждого параметра можно построить доверительные границы.

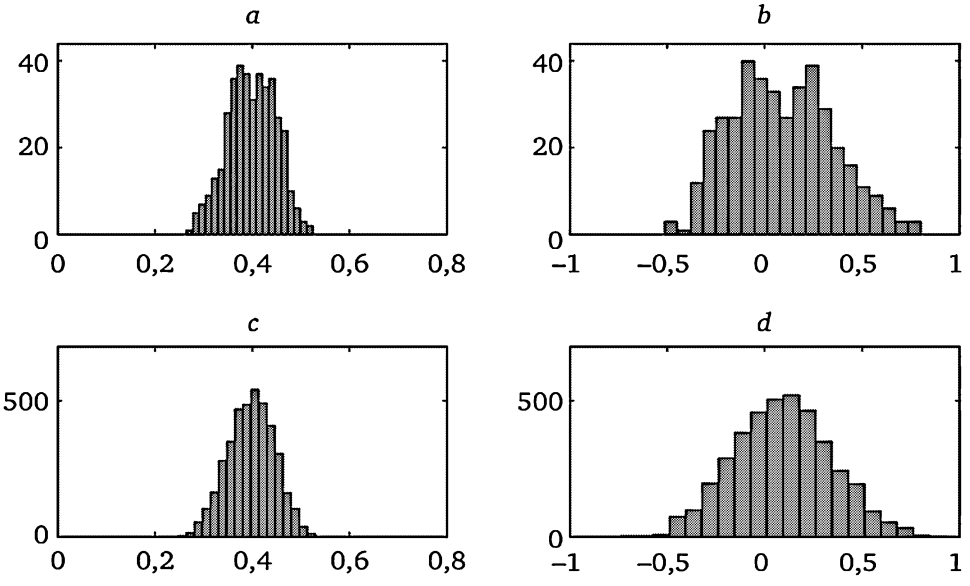


Рис. 3.4. **Гистограммы распределений наклона (слева) и сдвига (справа), найденные после 400 испытаний (верхний ряд) и 4000 испытаний (нижний ряд), бутстрэпа для регрессии количества школ Нш по населению Нас, выраженному в тысячах**

Таблица 3.1

Параметры линейной регрессии Нас по Нш, найденные по исходному множеству, а также в результате 400 и 4000 испытаний бутстрэпа

Регрессия / Параметр		400 испытаний			4000 испытаний		
		Среднее	2.5 %	97.5 %	Среднее	2.5 %	97.5 %
Наклон	0.401	0.399	0.296	0.486	0.398	0.303	0.488
Сдвиг	0.072	0.089	–0.343	0.623	0.092	–0.400	0.594

Примечание. Для испытаний указаны средние значения, нижние и верхние 2,5%-ные квантили.

Как можно получить доверительные границы, например, на 95 % уровне? По безопорному методу нижние и верхние 2.5 % квантили «вырезаются» из распределения симметричным образом, причем 95 % наблюдений попадают между ними. Для случая 400 испытаний 2.5 % — это 10, так что нижний квантиль относится к 11-му элементу отсортированного множества значений. Аналогично, верхний 2.5 % квантиль относится к 390-му элементу, отсекая «верхние» 10 значений. Для случая 4000 испытаний 2.5 % — это 100, так что верхний и нижний квантили относятся к 101-му и 3900-му элементам отсортированных множеств. Они приведены в табл. 3.1 для обоих случаев, 400 и 4000 испытаний. Можно заметить, что бутстрэп приводит к довольно узким границам для величины наклона: между 0.303 и 0.488 в 95 % всех испытаний (4000 испытаний); примерно те же границы получаются и при 400 испытаниях. Величина сдвига распределена с бóльшим разбросом, что ухудшает точность. Симметричные 95 % доверительные интервалы для величины сдвига: [–0.343,0.623] при 400 и [–0.400,0.594] при 4000 испытаниях.

Вопрос 3.1. Как здесь можно применить метод бутстрэпа с опорой, когда доверительный интервал строится так, как если бы распределения средних сдвига и наклона, полученные в результате бутстрэпа, были Гауссовы? Это бы привело к более устойчивым оценкам, нежели эмпирические распределения, используемые в безопорном бутстрэпе. Стандартные отклонения величин наклона и сдвига равны 0.0493 и 0.2606, соответственно, при 400 испытаниях бутстрэпа; они несколько меньше, 0.0477 и 0.2529, при 4000 испытаниях. Как извлечь из этого симметричный 95 % доверительный интервал для величин наклона и сдвига?

Подсказка. при Гауссовом распределении 95 % всех значений попадают в интервал «среднее ± 1.96 · стандартное отклонение».

Вопрос 3.2. Можете ли вы предложить оценку дисперсии разностей между наблюдаемыми значениями Нш и вычисленными по уравнению регрессии?

Последний тест адекватности уравнения регрессии, возможно, самый трудоемкий, основывается на понятии ошибки прогноза (см. Рабочий пример 3.3).

Рабочий пример 3.3

Ошибка прогноза по уравнению регрессии

Сравним наблюдаемые значения Нш с теми, которые вычислены по Нас согласно равенству (3.1). В табл. 3.2 представлены несколько примеров значений, взятых с обоих концов отсортированного признака Нас.

Таблица 3.2

Наблюдаемые значения числа начальных школ по сравнению со спрогнозированными значениями, исходя из численности постоянного населения (данные по прибрежным городам)

Нш наб.	Нш выч.	Нас	Нш наб.	Нш выч.	Нас
1	0.89	2040	2	2.35	5676
2	0.97	2230	2	2.9	7044
2	1.06	2452	4	4.12	10 092
2	1.19	2786	4	7.05	17 390
1	1.54	3660	13	9.62	23 801

В среднем прогнозы довольно близки, хотя и случаются более серьезные отклонения. Легко оценить относительную ошибку: $[(1 - 0.89)/1] \cdot 100 = 11$ % в первом случае, $[(2 - 0.97)/2] \cdot 100 = 51.5$ % во втором случае, и т. д. Средняя относительная ошибка в уравнении регрессии (3.1) составляет 30.7 %. Можно ли ее сделать меньше? На первый взгляд, нет, поскольку уравнение (3.1) по своей сути минимизирует ошибку. Но ошибка, которую мы минимизируем в (3.1), является средней квадратичной, а не относительной ошибкой. Эти две ошибки, безусловно, различаются, и уравнение (3.1) не обязательно оптимально для относительной ошибки.

3.2.3. Относительная ошибка прогноза в анализе данных и машинном обучении

Использованное в Рабочем примере 3.3 понятие относительной ошибки подразумевает сравнение абсолютной ошибки прогноза с истинным значением целевого признака. Расчет значений в табл. 3.2 опирается на неявное допущение, что истинными являются наблюденные значения целевого признака.

На самом деле здесь возможны две точки зрения на вопрос о том, что есть истина в данном случае, восходящие к давнему спору между двумя течениями философии — реализмом и номинализмом (см. параграф 2.1). Анализ данных, или, в современной ипостаси, наука данных, рассматриваемый как дисциплина, склонен

утверждать, что истина — это то, что мы наблюдаем, тогда как специалисты по машинному обучению, тем более сторонники «генеративного» подхода, склонны вслед за реалистами утверждать, что истина — в модели.

Так, казалось бы, чисто философские изыскания приводят нас к двум несовместимым правилам расчета относительной ошибки. Одно, так сказать, согласно «анализу данных», требует относить наблюденную ошибку к наблюденному значению выходного показателя; другое, так сказать, согласно «машинному обучению», требует относить наблюденную ошибку к значению выходного показателя, вычисленному согласно модели.

В табл. 3.3. представлены результаты использования уравнения (3.1) для прогнозирования значений признака Нш (2й столбец) по значениям признака Нас (1-й столбец) на выборках 10 самых меньших и 10 самых больших городов из табл. 1.11. Следующие столбцы содержат: вычисленные по уравнению (3.1) значения Нш (столбец 3), абсолютную величину разности значений вычисленного и реального значений Нш (столбец 4), а также относительные ошибки — результаты деления абсолютной ошибки на реальное значение Нш (столбец 5) и на вычисленное значение Нш (столбец 6), выраженные в процентах. Выделенные строки содержат средние значения показателей на соответствующих подмножествах: (а) самых малых городов, (б) самых больших городов, (с) всех городов из Табл. 1.11, соответственно.

С точки зрения науки данных истинны только данные, поэтому сравнивать абсолютную ошибку следует с наблюденными значениями Нш (столбец 5). С точки зрения «генеративного» подхода машинного обучения истинна модель и, соответственно, вычисленные по (3.1) значения Нш (столбец 3). Надо сказать, что в данном случае средние относительные ошибки, 30.68 % и 29.1 %, не очень отличаются. Если, однако, посмотреть на относительные ошибки в верхней части таблицы, на малых городах, то относительные ошибки отличаются, и довольно существенно, 19.21 % при подходе анализа данных и почти вдвое больше, 34.80 %, при подходе машинного обучения. В чем дело? Прямая линии регрессии (3.1) «ловит» средние тенденции, поэтому она должна завышать малые значения целевого показателя и занижать его большие значения. Значит, значения вычисленного знаменателя должны быть в среднем меньше реальных значений на больших городах, а результаты деления на них, соответственно выше. Очевидно, обратный феномен должен наблюдаться на подмножестве больших городов. Действительно, в этом случае относительная ошибка по методике анализа данных, 26.94 %, превышает относительную ошибку по методике машинного обучения, 22.88 %, хотя различие в этих значениях для больших городов не столь велико, как для малых.

Таблица 3.3

Относительные ошибки прогноза количества начальных школ Нш по сравнению с наблюденной численностью постоянного населения Нас (подход анализа данных) и прогнозной численностью постоянного населения (подход машинного обучения) на 10 самых маленьких и 10 самых больших городах табл. 1.11

№	1. Нас, тыс.	2. Нш набл	3. Нш мод	4. Ошибка абс	5. От. ош.-АД. %	6. От. ош.-МО, %
1	2.04	1.00	0.89	0.11	10.93	12.27
2	2.09	1.00	0.91	0.09	9.04	9.94
3	2.09	1.00	0.91	0.09	8.84	9.70
4	2.12	1.00	0.92	0.08	7.76	8.41
5	2.23	2.00	0.97	1.03	51.65	106.84
6	2.24	2.00	0.97	1.03	51.53	106.32
7	2.27	1.00	0.98	0.02	1.62	1.65
8	2.27	1.00	0.99	0.01	1.50	1.52
9	2.36	1.00	1.02	0.02	1.99	1.95
10	2.45	2.00	1.06	0.94	47.20	89.39
Среднее	2.22	1.30	0.96	0.34	**19.21**	**34.80**
36	10.22	5.00	4.17	0.83	16.51	19.78
37	12.55	5.00	5.11	0.11	2.19	2.15
38	14.14	4.00	5.75	1.75	43.66	30.39
39	15.87	7.00	6.44	0.56	8.02	8.71
40	17.39	4.00	7.05	3.05	76.27	43.27
41	18.97	9.00	7.68	1.32	14.63	17.14
42	19.71	10.00	7.98	2.02	20.18	25.29
43	20.30	6.00	8.22	2.22	36.96	26.99
44	21.62	7.00	8.75	1.75	24.99	19.99
45	23.80	13.00	9.62	3.38	**25.97**	**35.08**
Среднее	17.46	7.00	7.08	1.70	**26.94**	**22.88**
Общее среднее	7.35	3.02	3.02	0.81	**30.68**	**29.61**

Самостоятельная работа

3.3. Для той же табл. 1.11 сформируйте линейную регрессию признака По (Почтамт) по признаку Нас (Население) и постройте таблицу, аналогичную табл. 3.3.

3.4. Постройте уравнение регрессии признака «Ширина лепестка» по признаку «Длина лепестка» по данным об ирисах (табл 1.2), сравните спрогнозированные значения «Ширины лепестка» с наблюденными значениями и рассчитайте значения относительных ошибок. Вычислите и сравните средние значения относительной ошибки по методике анализа данных и методике машинного обучения.

3.5. Постройте уравнение регрессии признака Ба (число отделений банков) по признаку Нас (численность населения) по данным о городах английского побережья (табл 1.6), сравните спрогнозированные значения Ба с наблюденными значениями и рассчитайте значения относительных ошибок. Вычислите и сравните средние значения относительной ошибки по методике анализа данных и методике машинного обучения.

Классической теории оптимизации практически нечего предложить для минимизации относительной погрешности — этот критерий не относится к линейным, квадратичным, или выпуклым — а именно этим случаям уделяется основное внимание в математической теории. Попробуем применить подход эволюционной оптимизации, который все чаще используется для решения трудных оптимизационных задач в последнее время. В отличие от классического подхода, конструирующего единое решение, эволюционный подход оперирует с популяцией решений, которая эволюционирует случайным образом, итерация за итерацией, в поисках лучшего решения так, как это описано в Проекте 3.2. Применяя алгоритм из этого проекта для минимизации критерия относительной погрешности, можно найти другое решение, со средней относительной погрешностью в 26.4 %, т. е. снижению ошибки на 4.3 единицы (одна седьмая относительной погрешности уравнения регрессии (3.1)). Новое решение: Нш = 0.28 · Нас + 0.33, дает меньшее отношение темпа роста числа школ к темпу роста населения, 0.28, а не 0.4. Это еще раз показывает, что результаты анализа данных носят скорее индикативный характер, указывают направление, а не точные значения.

Задание 3.1. Правило Бершидского

Л. Бершидский обратил внимание на то, что поля рассеяния районов Москвы в системе двух признаков: стоимость жилья в районе и процент голосов, отданных на выборах мэра 8 сентября 2013 г., за одного из двух кандидатов (С. С. Собянина и А. Н. Навального), образуют противоположно направленные кластеры.[1] Рис. 3.5 из статьи Гурьянова[2] основан на официальных данных.

---

[1] URL: www.forbes.ru/mneniya-column/vertikal/244605-klassovyi-vybor-kak-tseny-na-zhile-predreshili-iskhod-golosovaniya-v

[2] *Гурьянов, В.* «Закон Бершидского: стоимость квадратного метра определила результаты выборов мэра // Квадратъ. 2013. № 44.

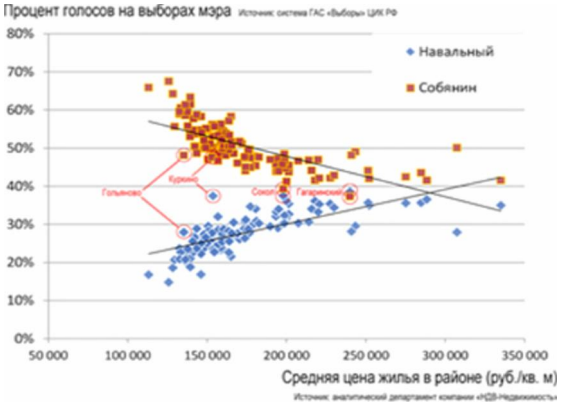


Рис. 3.5. **Поля рассеяния по результатам голосования за каждого из двух кандидатов на выборах в мэры в Москве (сентябрь 2013)**

Каждое поле проявляется в довольно четких линейных связях; одна с положительным коэффициентом корреляции, вторая — с отрицательным.

Задание 3.2 Распределение стран по признакам уровня и продолжительности жизни

На рис. 3.6 представлено поле рассеяния стран мира на плоскости признаков «уровень жизни» и «ожидаемая продолжительность жизни».

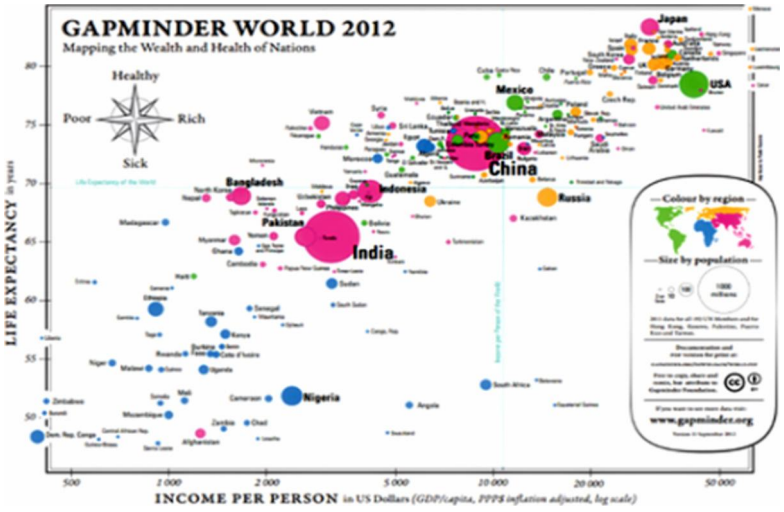


Рис. 3.6. **Поле рассеяния стран мира в пространстве признаков:** по оси абсцисс — среднедушевой доход (в логарифмической шкале); по оси ординат— средняя ожидаемая продолжительность жизни (в годах); страны представлены кружками радиуса, пропорционального населению. Цвета определяются регионом: Европа, включая Россию и Кавказ; Азия и Австралия; Африка; Америка, Северная и Южная[1]

[1] URL: http://www.gapminder.org

Последний измеряется с помощью специальных таблиц «дожития», оценивающих, для каждой возрастной категории, вероятности дожития до следующего года. Некоторая особенность этой таблицы — логарифмическое представление дохода на душу населения, позволяющее «спрямить» характер распределения. Кроме того, в визуализацию включена третья переменная — население страны, отражаемое в размере кружка, ей соответствующего. Четвертая, номинальная переменная, относящаяся к континенту, на котором расположена страна — отражена цветом, хорошо видном на веб-сайте, на котором вывешено данное изображение.

Хорошо видно, что африканские страны имеют низкие значения по обоим признакам, тогда как западноевропейские — высокие. Если посмотреть данное поле рассеяния в динамике, то можно заметить, что при сохранении общего характера графика, его значения растут как в Европе, так и в Африке, да и во всех других странах.

Ф3.3. Линейная регрессия: Формулировки

Ф3.3.1. Аппроксимационная перспектива: Линейная регрессия и коэффициент корреляции

Определим параметры линейной регрессии. Имеется N объектов, для каждого из которых заданы целевой признак y и входной признак x: (x_1,y_1), (x_2, y_2), …, (x_N, y_N). Вопрос в том, чтобы идентифицировать линейное уравнение, которое бы их связывало:

$$y = ax + b. \tag{3.2}$$

Точную подгонку коэффициентов можно произвести, только если все пары (x_i, y_i) лежат на одной и той же прямой в плоскости (x, y), что для реальных данных маловероятно. Это значит, что в уравнении (3.2) для каждой пары (x_i, y_i) будет невязка, не обязательно нулевая. В этом случае наше уравнение может быть переписано в виде

$$y_i = ax_i + b + e_i \ (i = 1, 2, \ldots, N), \tag{3.2′}$$

где e_i — невязки, или ошибки, регрессии. Явный учет невязок позволяет поставить задачу отыскания коэффициентов a и b так, чтобы остатки могли быть минимизированы по критерию наименьших квадратов. Минимизируется средняя квадратичная ошибка

$$L(a, b) = \Sigma_i e_i^2 / N = \Sigma_i (y_i - ax_i - b)^2 / N, \tag{3.3}$$

по всем возможным a и b при заданных x_i и y_i (i = 1, 2, …, N). Эта задача минимизации легко может быть решена с использованием понятий математического анализа. В самом деле, $L(a, b)$ — параболическая функция от a и b, ветвями вверх, поэтому ее минимум

находится в точке, где частные производные $L(a, b)$ по a и b равны нулю (условие оптимальности первого порядка):

$$\partial L/\partial a = 0 \text{ и } \partial L/\partial b = 0.$$

Предоставим математически ориентированному читателю найти выражения для этих частных производных, а также единственное их решение, в качестве упражнения. Решение можно выразить формулами (3.4)—(3.5) для a, и (3.6) — для b:

$$a = \rho\sigma(y) / \sigma(x) \tag{3.4}$$

где

$$\rho = [\Sigma_i(x_i - m_x)(y_i - m_y)] / [N\sigma(x)\sigma(y)] \tag{3.5}$$

— так называемый коэффициент корреляции, а m_x, m_y — средние значения x_i и y_i, соответственно;

$$b = m_y - am_x. \tag{3.6}$$

Подставив оптимальные значения a и b в (3.3), можно получить выражение для минимального значения критерия (3.3):

$$L_m(a, b) = \sigma^2(y)(1 - \rho^2) \tag{3.7}$$

Уравнение (3.2) называется линейной регрессией y по x, величина ρ в (3.4) и (3.5) — коэффициентом корреляции, а его квадрат ρ^2 в (3.7) — коэффициентом детерминации. Минимальное значение критерия L_m в (3.7) — это необъясненная часть дисперсии целевого признака y.

Смысл коэффициентов корреляции и детерминации при восстановлении данных в задачах анализа данных отражен в формулах (3.3)—(3.7). Ниже приведены некоторые их свойства.

Свойство 1. Коэффициент детерминации ρ^2 характеризует долю дисперсии признака y, учтенную в построенной линейной регрессии y по x (следует из уравнения (3.7)).

Свойство 2. Коэффициент корреляции ρ изменяется в интервале от -1 до 1. (Это следует из того, что ρ^2 лежит в интервале от 0 и 1, так как значение L_m в уравнении (3.7) не может быть отрицательным, потому что оно выражается через квадраты в уравнении (3.3).) Чем ближе ρ^2 к 1 или к -1, тем меньше остатки в линейном регрессионном уравнении. Например, величина $\rho = 0.9$ означает, что необъясненная часть дисперсии y L_m равна $1 - \rho^2 = 1 - 0.81 = 19$ % от исходной величины.

Свойство 3. Наклон a пропорционален ρ согласно уравнению (3.4); a положительно или отрицательно в зависимости от знака ρ.

Если $\rho = 0$, наклон нулевой: в этом случае y и x называются некоррелированными. («Не коррелированы» — вовсе не значит «не связаны»!)

Свойство 4. Коэффициент корреляции ρ не изменяется при сдвиге и изменении масштаба x и (или) y, о чем свидетельствует уравнение (3.5). Выражение (3.5) становится гораздо проще, если признаки x и y стандартизованы с помощью преобразования, называемого в статистике z-скоринг. Преобразование z-скоринг данного признака заключается в следующем: его среднее значение m вычитается из всех его значений, а результат делится на стандартное отклонение σ:

$$x_i' = (x_i - m_x) / \sigma(x),\ y_i' = (y_i - m_y) / \sigma(y),\ i = 1, 2, \ldots, N.$$

С использованием стандартизации z-скоринг формула (3.5) может быть переписана в виде:

$$\rho = \Sigma_i x_i' y_i' / N = < x', y' > / N, \qquad (3.5')$$

где $<x', y'>$ означает скалярное произведение стандартизованных векторов $x' = (x_i')$ и $y' = (y_i')$, $<x', y'> = \Sigma_i x_i' y'$.

Ф3.2.2. Вероятностная перспектива: двумерное Гауссово распределение и линейная регрессия

Следующее свойство связано с одним из фундаментальных открытий К. Пирсона — интерпретацией коэффициента корреляции в терминах двумерного Гауссова распределения. Общая формула для функции плотности этого распределения с учетом предварительной стандартизации по методу z-скоринг имеет вид:

$$f(u, \Sigma) = Ce^{-\frac{1}{2}u^T \Sigma^{-1} u}, \qquad (3.8)$$

где $u = (x, y)$ двумерный вектор рассматриваемых переменных x и y, а Σ — так называемая матрица корреляции

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

В формуле (3.8) ρ — параметр с четким геометрическим смыслом. Рассмотрим множество точек $u = (x, y)$ на плоскости (x, y), на которых функция плотности постоянна, т. е. $f(u, \Sigma) = c$, где c — некоторая константа. Формула (3.8) гарантирует, что в этом случае величина $u^T \Sigma^{-1} u$ тоже будет постоянной, независимо от u. Это означает, что множество точек $u = (x, y)$ постоянной плотности c удовлетворяет уравнению $x^2 - 2\rho xy + y^2 = \text{const}$. Это уравнение задает хорошо известную квадратичную фигуру — эллипс. При $\rho = 0$ уравнение превращается в уравнение окружности $x^2 + y^2 = \text{const}$. Чем больше

отличие ρ от 0, тем «уже» становится эллипс. При ρ = 1 эллипс превращается в прямую линию $y = x + b$, потому что левая часть уравнения становится полным квадратом, в этом случае $x^2 - 2xy + y^2$, так что $(y - x)^2 = \text{const}$. Размер эллипса пропорционален константе c: чем она больше, тем больше размер.

Гауссово двумерное распределение в общем случае определяется точкой математического ожидания $\mu = (x_\mu, y_\mu)$ и матрицей ковариации

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \varrho\sigma_x\sigma_y \\ \varrho\sigma_y\sigma_x & \sigma_y^2 \end{pmatrix}, \tag{3.9}$$

где ρ — коэффициент корреляции, а σ_x^2 и σ_y^2 — дисперсия x и y, соответственно. Функция плотности Гауссова распределения в общем случае имеет вид:

$$f(u, \Sigma) = Ce^{-\frac{1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu)}, \tag{3.10}$$

где $u = (x, y)$, Σ — ковариационная матрица (3.9), а C — константа, определяемая условием, что определенный интеграл от функции плотности (3.10) по всей плоскости (x, y) равняется единице, $C = 2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}$.

Для заданной двумерной функции плотности $f(x, y)$, маргинальная плотность вероятности для x определяется как интеграл от $f(x, y)$ по y на интервале $[a, b]$ определения y:

$$f_x(x) = \int_a^b f(x, y)dy.$$

Аналогично определяется маргинальная функция плотности по y.

Условная плотность вероятности y для заданного значения $x = x_0$ определяется как

$$f(y \mid x = x_0) = \frac{f(x_0, y)}{f_y(x_0)}. \tag{3.11}$$

Оказывается, и маргинальная, и условная плотности для двумерного Гауссова распределения являются Гауссовыми же функциями плотности. Более того, можно доказать, что математическое ожидание условной плотности $f(y \mid x = x)$ имеет математическое ожидание, равное

$$E(y \mid x) = ax + (\mu_y - a\mu_x), \tag{3.12}$$

где $a = \rho\frac{\sigma_y}{\sigma_x}$, и дисперсию, равную

$$\sigma^2(y \mid x) = \sigma_y^2(1-\rho^2). \tag{3.13}$$

Совершенно очевидно, что уравнения (3.12) и (3.13) — теоретическое выражение того самого решения, которое было получено нами для задачи отыскания параметров уравнения линейной регрессии в соответствии с критерием наименьших квадратов (см. равенства (3.4), (3.6) и (3.7)).

Это означает, что, так сказать, «навязанное» квадратичным критерием уравнение линейной регрессии для случая Гауссова распределения на самом деле выражает внутренние свойства этого распределения. А именно, справедливо следующее утверждение.

Свойство 5. Для независимой случайной выборки из Гауссова распределения с плотностью $f(x.y)$ уравнение линейной регрессии по методу наименьших квадратов совпадает с математическим ожиданием условной плотности $f(y|x = x)$, а остаточная дисперсия — с дисперсией условной плотности.

При этом имеет место свойство 6.

Свойство 6. Коэффициент корреляции (3.5) — это не что иное, как оценка параметра ρ в (3.8), построенная по выборке из Гауссова распределения, при стандартном предположении о случайности и независимости точек (x_i, y_i), попавших в выборку.

Эти поразительные факты — свойства 5 и 6 — лежат в основе неправильного мнения, популярного среди психологов, социологов и экономистов. Согласно этому мнению, использование коэффициента корреляции корректно только при том условии, что выборка взята из двумерного Гауссова распределения. Подобная логика носит ограниченный характер. Она применима, если речь идёт об оценке функции плотности двумерного распределения. Нас же интересует совсем другой вопрос — качество линейного представления одной переменной через другую. Коэффициент корреляции имеет абсолютно другой смысл в контексте аппроксимации, не имеющий ничего общего с Гауссовым распределением, как это отражено выше в свойствах 1—4 и уравнениях (3.4)—(3.7). В этом плане никаких ограничений на использование коэффициента корреляции нет.

Вычисление линейной регрессии, согласно формулам (3.4)—(3.6), можно провести с помощью следующего псевдокода:

```
>> rho = corr(x,y); % коэффициент корреляции
>> slope = rho*std(y)/std(x); % наклон
>> intercept = mean(y) - slope*mean(x); % сдвиг
```

Ф3.3.3. Ложная корреляция: влияние выбросов и неоднородности

Когда Ф. Гальтон и К. Пирсон обнаружили коэффициент корреляции как важную характеристику связи двух признаков, они некоторое время считали, что коэффициент корреляции отражает наличие причинных связей. Однако, оказалось, что это не так.

Были обнаружены высокие корреляции между абсолютно несвязанными характеристиками. Читатель может обратиться к сайту https://tylervigen.com/spurious-correlations, на котором приводятся десятки примеров ложной корреляции между очевидно несвязанными показателями, такими, например, как потребление шоколада и количество Нобелевских лауреатов в стране, отнесенные к ее населению (см. F. H. Messerli (2012), The New England Journal of Medicine, 367(16), 1562).

В общем случае наука мало что может сказать о причинах возникновения ложной корреляции. Но одна причина известна достаточно подробно — это неоднородность данных. Объединяя в одном множестве разнородные объекты, мы можем как увеличивать, так и уменьшать корреляцию между признаками. Далее рассмотрим на примерах два случая разнородности данных: наличие выбросов и парадокс Симпсона.

П3.3.3.1. Влияние выбросов на коэффициент корреляции

Наличие выбросов — сильно выделяющихся наблюдений — может сильно исказить картину корреляции между признаками, обычно в сторону увеличения корреляции.

Сгенерируем для примера два независимых случайных признака, один распределенный по однородному, второй — по нормальному закону:

```
>> x = 10*rand(500,1)-4;
>> y = 5*randn(500,1)+4;
```

Признак x задан на 500 объектах и меняется между –4 и 6, поскольку генератор rand производит псевдослучайные числа между 0 и 1, а последующее умножение на 10 переводит их в интервал от 0 до 10, который сдвигается к (–4, 6) после вычитания 4. Аналогично, признак y распределен нормально с центром в точке 4 и стандартным отклонением 5. Его считаем заданным на тех же 500 объектах. Вот начало таблицы данных (полный размер 500 × 2):

Объект	x	y
1	–2.3782	8.3207
2	3.9428	4.5671
3	–0.8878	5.9918
4	1.2853	8.4198
5	–2.3435	4.9013

Поскольку вышеприведенные определения признаков никак не связаны, корреляция между ними должна быть нулевая. Команда

```
>>cc = corr(x,y)
```

приводит к значению коэффициента корреляции cc = 0.0689, которое, действительно близко к нулю, но вст-таки отличается от него. Почему? Из-за случайности выборки.

Теперь добавим к сгенерированным пятистам объектам два выброса, т. е. сильно отличающиеся наблюдения. Для удобства оформим модификацию как другую пару объектов, номер 501 и номер 502:

```
>> xn = x; yn = y; % копирование старых признаков в новые
>> xn(501) = –100; yn(501) = –100;
>> xn(502) = 100; yn(502) = 200;
```

Нами взяты сильно отклоняющиеся значения, но для наглядности так, чтобы не слишком отклонялись от биссектрисы прямого угла между осями координат. Теперь, при добавленных к имеющимся пятистам двух новых наблюдениях, сильно ли изменится корреляция между признаками? Вычисление

```
>>cn = corr(xn,yn)
```

приводит к значению cn = 0.7910, близкому к единице!

В чем дело?

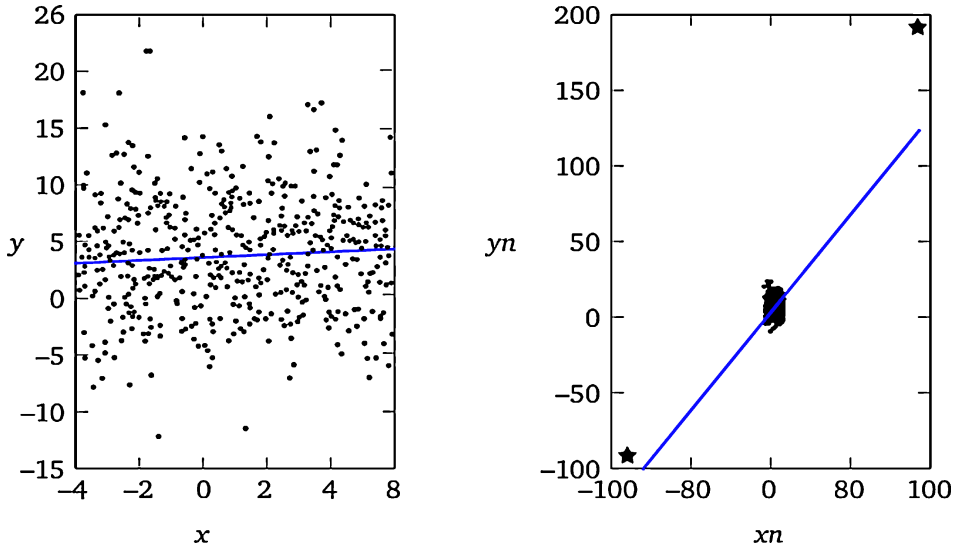


Рис. 3.7. Поля **рассеяния признаков (*x, y*) (слева) и (*xn, yn*) (справа):** на правом рисунке введенные «выбросы», объекты 501 и 502, изображены звездами. Обратите внимание на разницу в масштабах

Ситуация проясняется, если посмотреть на поля рассеяния признаков (рис. 3.7). Слева — поле рассеяния исходных данных, случайное нагромождение точек, практически нулевая корреляция. Справа — совсем другое дело! Добавление двух выбросов, изображенных для наглядности увеличенными звездочками, полностью поменяло масштаб.

В новом масштабе исходные 500 объектов — не более чем случайное нагромождение реализаций как бы единой «средней» точки, а «реальная» картина связи между *xn* и *yn* определяется взаиморасположением трех «главных» сил — двух выбросов и кляксы посередине.

Понятно, что эта глобальная картина крайне неустойчива. При случайных выборках двух третей объектов выбросы, как правило, в выборку не попадут (их ведь всего два!), так что основная доля испытаний бутстрэпа (см. далее) будет давать нулевую корреляцию.

Задание 3.3. Проверьте, насколько изменится коэффициент корреляции, если выбросы будут не x_{501} = (–100, –100) и x_{502} = = (100,200), как сейчас, а еще в 5 раз выше, т. е. (–500, –500) и (500, 1000).

ПЗ.3.3.2. Эффект скрытого признака: Парадокс Симпсона

Вернемся к данным об ирисах. Обозначим через *x* длину, а через *y* — ширину чашелистика для множества ирисов из табл. 1.3.

Поле рассеяния *x* и *y* представлено на рис. 3.8 слева. Это облако точек без видимого направления, напоминающее множество точек с нулевой корреляцией на рис. 3.2 слева. Однако в данном случае величина коэффициента не только мала, порядка –0.12, но еще и отрицательна, что довольно странно, поскольку интуитивно корреляция между длиной и шириной чашелистика должна быть положительной — ведь оба напрямую характеризуют размер!

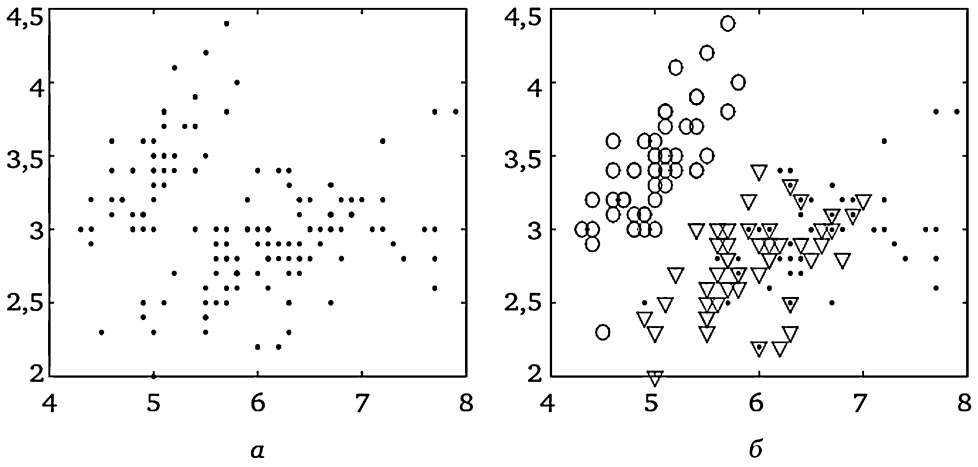


Рис. 3.8. **Диаграмма разброса длины и ширины чашелистика, отражающая данные из табл. 1.3:**
а — в общем виде; *б* — с разбиением на таксоны. Таксон 1 представлен кружками, таксон 2 — треугольниками, таксон 3 — точками

Чтобы понять, в чем причина такой низкой и даже отрицательной корреляции, необходимо учесть, что выборка не является однородной: множество Iris состоит из 50 образцов каждого из трех таксонов.

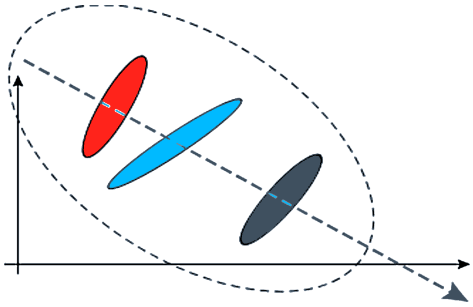


Рис. 3.9. **Структура парадокса Симпсона, проявленного в примере, представленном на рис. 3.8:**
объединение трех эллипсоидов, каждый с выраженной положительной корреляцией между признаками, порождает объемлющий эллипсоид — с выраженной отрицательной корреляцией.

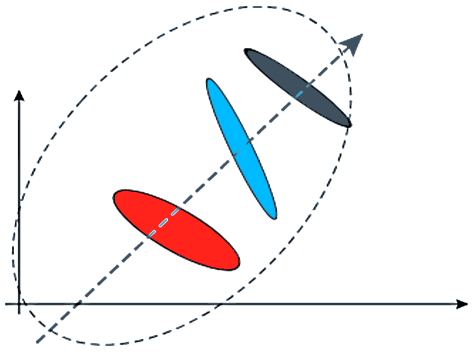


Рис. 3.10. **Структура парадокса Симпсона, в котором объединены три множества с отрицательной корреляцией между признаками. В объединении — корреляция положительна**

Когда таксоны разделены (см. рис. 3.8 справа), положительная корреляция обнаруживается. Коэффициенты корреляции равны 0.74, 0.53 и 0.46 для объектов внутри таксонов 1, 2 и 3 соответственно. Это пример эффекта неоднородности выборки на величину коэффициента корреляции. В литературе этот эффект называется парадоксом Симпсона.

Схема парадокса Симпсона, проявленного в данном примере, изображена на рис. 3.9. Подобная же структура, но противоположной направленности, представлена на рис. 3.10.

ФЗ.3.4. Метод линеаризации для оценки нелинейной регрессии

Нелинейные зависимости также могут быть аппроксимированы, причем с помощью того же критерия минимизации среднего квадрата ошибки. Рассмотрим довольно распространенный случай экс-

поненциальной регрессии. В этом случае речь идет о возможности описания связи целевого признака y и входного признака x в виде $y = ae^{bx}$, где a и b — неизвестные константы, e — основание натурального логарифма. При заданных a и b средний квадрат невязки вычисляется как

$$E = ([y_1 - a\exp(bx_1)]^2 + \ldots + [y_N - a\exp(bx_N)]^2)/N =$$

$$= \Sigma_i [y_i - a\exp(bx_i)]^2/N. \qquad (3.14)$$

Не существует метода, который бы непосредственно давал глобально оптимальное решение задачи минимизации E в (3.14), поскольку функция E довольно сложна. Вот почему частенько коэффициенты уравнения экспоненциальной регрессии отыскиваются с помощью предварительной линеаризации: преобразования исходной задачи к задаче линейной регрессии. В самом деле, возьмем логарифм от обеих частей уравнения $y = ae^{bx}$. В результате получим уравнение $\ln(y) = \ln(a) + bx$. Его можно переписать в виде уравнения линейной регрессии $z = \alpha x + \beta$, где $z = \ln(y)$, $\alpha = b$ и $\beta = \ln(a)$, что наталкивает на следующую идею. Положим целевую величину равной $z = \ln(y)$ со значениями $z_i = \ln(y_i)$. Отыскав коэффициенты уравнения линейной регрессии z по x с использованием данных x_i и z_i, мы получим оптимальные α и β. Теперь вычислим коэффициенты экспоненциального уравнения: $a = \exp(\beta)$ и $b = \alpha$. Эти значения не обязательно минимизируют (3.14), но, предположительно, близки к оптимуму. К сожалению, иногда это может быть совсем не так, как описано далее в Проекте 3.2.

Вопрос 3.3. Найдите производные L по a и b и решите уравнения, выведенные из условий оптимальности первого порядка.

Вопрос 3.4. Найдите оптимальное значение L в (3.7), подставив оптимальные a и b.

Вопрос 3.5. Докажите или найдите доказательство в литературе, что линейное уравнение $y = ax+b$ соответствует прямой линии на декартовой плоскости xy, причем a — наклон этой прямой, а b — сдвиг вдоль оси y.

Вопрос 3.6. Найдите обратную матрицу Σ^{-1} для $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Ответ. $\Sigma^{-1} = \dfrac{\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}}{1-\rho^2}$.

Проект 3.1. Линейная регрессия и бутстрэп

Зададим в MATLAB данные из таблицы данных об ирисах в виде массива 150 × 4 в переменной iris. Выберем любые два признака и зададим множество объектов в виде множества точек декартовой

плоскости. Например, пусть длина лепестка — это x, а его ширина — это y:

```
>> x = iris(:,3);
% длина лепестка является 3-м столбцом массива iris
>> y = iris(:,4); % ширина лепестка в 4-м столбце iris
```

Цветок ириса 1 (первая строка) задается точкой с координатами $x = 1.4$ — длина лепестка, и $y = 0.3$ — его ширина. Чтобы построить поле рассеяния данных признаков, используем команду:

```
>> plot(x,y,'k.')
% k соответствует черному цвету,
% «.» – точечному представлению объектов;
% 'rp' соответствовало бы пентаграмме (пятиконечная звезда)
% красного цвета;
```

К сожалению, такое представление не совсем удачно, так как некоторые точки попали на границы поля (см. рис. 3.11, *а*). Чтобы отдалить границы, можно воспользоваться командой axis, например, так:

```
>>plot(x,y,'k.') ; d = axis; axis(1.2*d-.1);
% здесь рамка увеличена на 20 % и смещена вниз
```

или так

```
>>plot(x,y,'k.') ; axis([-.5 8 -.5 3]);
% здесь рамка задана значениями x (первая пара) и y
```

Эти преобразования изображены на рис. 3.11, *б* и *в*. Чтобы отобразить все три графика в одном и том же окне, используем функцию subplot MATLAB:

```
>> subplot(1,3,1); plot(x,y,'k.');
>> subplot(1,3,2); plot(x,y,'rp'); d = axis; axis(1.2*d-10);
>> subplot(1,'1'; plot(x,y,'k.') ; axis([-.5 8 -.5 3]);
% здесь первые два аргумента функции subplot характеризуют
% количество рядов и столбцов для размещения рисунков,
% а третий – номер конкретного окна в этой структуре
% для размещения последующего plot
```

Поле рассеяния признаков длины и ширины лепестков ириса выглядит довольно обещающим с точки зрения наличия между ними линейной связи.

Уравнение линейной регрессии имеет вид $y = slope*x + intercept$. Оценим его параметры в MATLAB, используя формулы (3.4)—(3.6):

```
>> cc = corrcoef(x,y); rho = c(1,2); % rho = 0.9629
>> slope = rho*std(y)/std(x); % slope = 0.4158;
>> intercept = mean(y) – slope*mean(x);
% intercept = –0.3631;
```

Здесь использованы команды более старых версий MATLAB, в которых функция corr отсутствует, зато есть функция corrcoef, порож-

дающая матрицу коэффициентов корреляции для данного множества признаков.

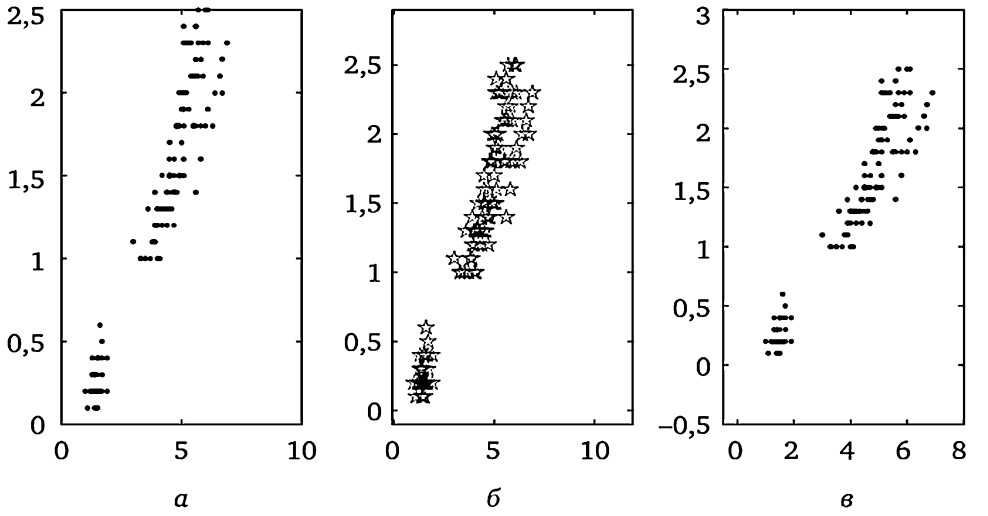


Рис. 3.11. **Поле рассеяния признаков длина и ширина лепестка с разными масштабами рамки**

С функцией corr первая строка должна выглядеть следующим образом:

```
>> rho = corr(x,y); % rho = 0.9629
```

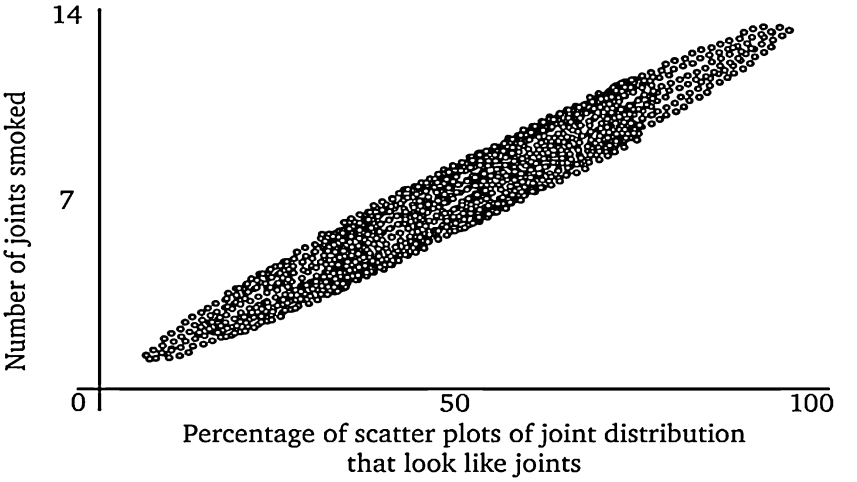


Рис. 3.12. **Поле рассеяния признаков, связанных непереводимой игрой слов английского языка**

Таким образом, получаем уравнение линейной регрессии $y = 0.416x - 0.363$. Величина наклона говорит о том, что каждый до-

полнительный сантиметр длины увеличивает ширину в среднем на 0.416 см. Как обычно в анализе данных, требуется определённая осторожность при формировании подобных заключений: строго говоря, они верны только в диапазонах реально наблюденных значений.

В целом, уравнение регрессии объясняет $\rho^2 = 0.927 = 92.7$ % от общей дисперсии y — это довольно высокий процент, как это нередко бывает в естественнонаучных исследованиях и практически никогда в социальных и гуманитарных науках.

Оценим надежность данного уравнения регрессии с помощью бутстрэпа. Это довольно популярный вычислительный инструмент для оценки доверительных интервалов для результатов анализа данных, который был описан в Проекте 2.3 на примере задачи валидации среднего значения.

Бутстрэп основан на заранее заданном числе испытаний, например, 5000. Каждое испытание состоит из следующих шагов:

(i) Из данной выборки случайно, с возвращением, N раз выбирается по объекту. При этом некоторые объекты могут быть выбраны несколько раз, в то время как другие могут не попасть в испытание вообще (как показано выше в Проекте 2.3, в среднем только 62 % объектов попадают в выборку). N — это число объектов рассматриваемого множества, в нашем случае $N = 150$. Используем следующие команды MATLAB:

```
>> N = 150;ra = ceil(N*rand(N,5000));
% rand(N,5000) задает 5000 столбцов из N случайных
% действительных чисел от 0 до 1.
% Умножение на N позволяет перейти к числам из интервала (0,N);
% операция ceil округляет выбранные числа до ближайших бóльших
% целых, так что элементы матрицы ra - это целые числа
% в интервале от 1 до 150.
```

Каждый столбец сгенерированной матрицы ra имитирует отдельное испытание — случайный выбор с возвращением N объектов из заданной совокупности N объектов.

(ii) Значения признаков на выбранных индексах — элементах матрицы ra — определяются как:

```
>>xt = x(ra); yt = y(ra);
% здесь x и y - входной и целевой признаки соответственно;
```

При этом одни и те же объекты получают одни и те же значения признаков.

(iii) Обращаемся к тому методу анализа данных, для которого производится валидация, в нашем случае это метод построения линейной регрессии. Для каждого испытания $k = 1, 2, \ldots, 5000$ этим методом вычисляются коэффициент корреляции rho, наклон (slope) и сдвиг (intercept). К сожалению, при этом используются операции,

не применимые к матрицам. Поэтому вычисления проводим в цикле по отдельным испытаниям — столбцам матрицы ra:

```
>> for k = 1:5000; r = ra(:,k);
     xt = x(r); yt = y(r);
     rh(k) = corr(xt,yt);
     sl(k) = rh(k)*std(yt)/std(xt);
     inte(k) = mean(yt)-sl(k)*mean(xt);
end
% результаты: rh (5000 значений коэффициента корреляции),
% sl (5000 величин наклона) и inte (5000 величин сдвига)
```

Теперь можно посчитать среднее и стандартное отклонение полученных 5000 значений параметров:

```
>>msl = mean(sl);ssl = std(sl);
```

что дает msl = 0.4159 и ssl = 0.0098. Это означает, что исходная величина наклона 0.416 подтверждается процедурой бутстрэпа. Но бутстрэп порождает разнообразие оценок среднего, позволяющее вычислить величину стандартного отклонения, 0.0098. Аналогичным образом вычисляются величины среднего и стандартного отклонения для сдвига и коэффициента корреляции: –0.363/0.0277 и 0.9629/0.0, соответственно.

Далее построим 30-бинные гистограммы для найденных 5000 значений наклона и сдвига (см. рис. 3.13):

```
>> subplot(1,2,1); hist(sl,30)
>> subplot(1,2,2); hist(inte,30)
```

Функция subplot(1,2,1) создает ряд из двух окон для изображений, причем помещает гистограмму наклонов в первом из них (рис. 3.13, *а*). Функция subplot(1,2,2) вносит гистограмму сдвигов во второе окно.

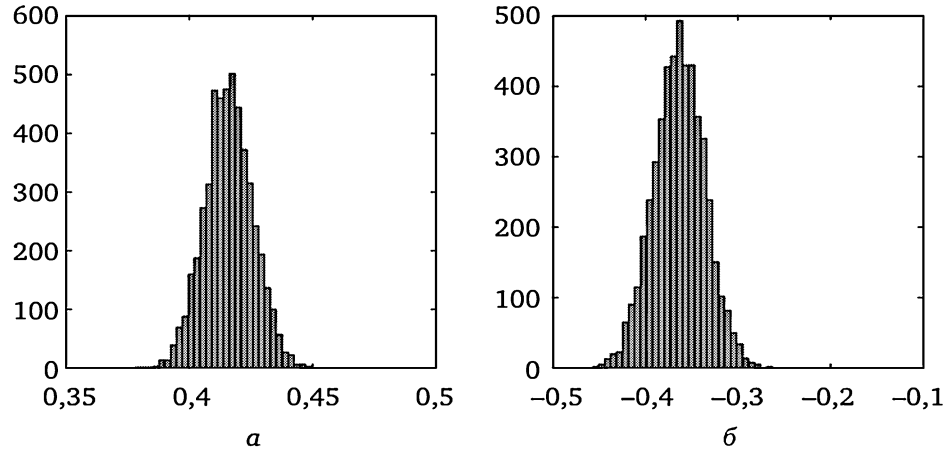


Рис. 3.13. **30-бинные гистограммы для наклона (*а*) и сдвига (*б*) линейной регрессии ширины лепестка по его длине, полученные в результате 5000 испытаний бутстрэпа**

Таблица 3.4

Параметры распределений бутстрэпа, а также границы 95%-ного доверительного интервала, полученные с использованием и без использования опоры

Параметр	Среднее	Стд. откл.	Границы с опорой		Границы без опоры	
			Левая	Правая	Левая	Правая
Наклон	0.4159	0.0098	0.3967	0.4351	0.3966	0.4351
Сдвиг	–0.3636	0.0277	–0.4178	–0.3093	–0.4185	–0.3092
Коэфициент корреляции	0.9630	0.0051	0.9530	0.9730	0.9519	0.9721

Для получения 95%-ного доверительного интервала для наклона, сдвига и коэффициента корреляции можно использовать как метод с опорой, так и метод без опоры. Метод с опорой использует предположение о том, что выборка 5000 значений в процедуре bootstrap — случайная независимая выборка из Гауссова распределения. Параметры этого распределения определяются как среднее и стандартное отклонение выборочных значений:

```
>> msl = mean(sl); ssl = std(sl);
```

Поскольку 95 % площади Гауссового распределения попадает в интервал «среднее ± 1.96 · сто», границы 95%-ного доверительного интервала для величины наклона получаются следующим образом:

```
>> lbsl = msl - 1.96*ssl; rbsl = msl + 1.96*ssl
```

Безопорные вычисления определяются только по выборке. Нужно отсортировать все значения выборки, после чего 2.5%-ные квантили на краях отсортированного ряда, в данном случае, отбросив по 125 крайних значений (125 = 2.5 % от 5000):

```
>> ssl = sort(sl); lbn = ssl(126); rbn = ssl(4875);
```

Действительно, чтобы построить 95%-ный доверительный интервал, необходимо удалить из выборки 5 % объектов по краям отсортированного ряда. Так как 5 % от 5000 это 250, и придерживаясь центрального интервала для отбора 95 % значений, надо удалить из отсортированной выборки бутстрэпа первые 125 и последние 125 наблюдений. т. е. величины ssl(126) и ssl(4875) и есть левая и правая границы доверительного интервала величины наклона по безопорному методу.

Аналогично вычисляются границы 95%-ного доверительного интервала, lbin и rbin, для полученного распределения величин сдвига.

Все эти оценки представлены в табл. 3.4. Опорные и безопорные границы отличаются не слишком сильно.

Полученные результаты можно представить с помощью трех регрессионных прямых, обычной, а также двух вспомогательных, соответствующих нижним и верхним границам доверительных оценок, соответственно:

```
>> yreg = slope*x+intercept; % регрессия по исходной выборке
>> yregleft = lbsl*x+lbin; % линия с с самыми левыми границами
>> yregright = rbsl*x+rbin; % линия с с самыми правыми границами
```

и затем отобразить все три на поле рассеяния данных (рис. 3.14):

```
>> plot(x'y'*k',x,yreg,'k',x,yregleft,''x,yregright,'r')
% x,y,'*k' данные по студентам отображаются черными звездочками;
% x,yreg,'k' график обычной регрессионной прямой
% задается черной линией x,yregleft,'r' and x,yregright,'r'
% красный для граничных регрессий
```

Как видим (см. рис. 3.14), ограничивающие прямые не слишком далеки от построенной линии регрессии, что, конечно же, объясняется очень высокой корреляцией рассматриваемых признаков.

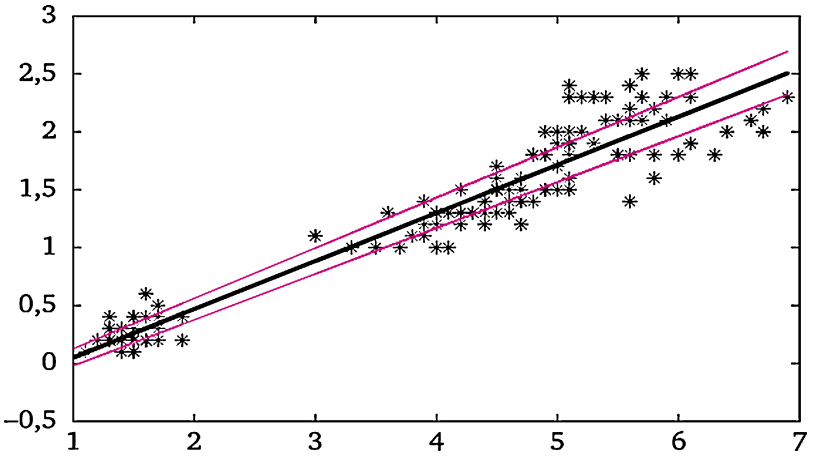


Рис. 3.14. **Регрессия ширины лепестка ирисов (утолщенная прямая) относительно их длины, а также граничные прямые 95%-ного доверия**

Проект 3.2. Нелинейная и линеаризованная регрессии: инспирированный природой алгоритм

Связь между признаками не обязательно линейна. Например, в экономике процессы, связанные с инфляцией, моделируются с помощью экспоненциальной функции. Подобный подход применяется и к процессам роста в биологии. Переменные, описывающие климатические условия, очевидно, имеют циклический характер «зима — лето» . Для многих социальных процессов характерен степенно́й закон.

Рассмотрим, например, степенну́ю функцию $y = ax^b$, где x — входной, прогнозирующий признак, а y — выходной, прогнозируемый

признак, тогда как коэффициенты a и b неизвестны. Для каждого наблюдаемого объекта $i = 1, \ldots, N$ известны значения x_i и y_i. Задача степенно́й регрессии может быть сформулирована как задача минимизации суммы квадратов или абсолютных значений ошибки по всем парам коэффициентов a и b. Заметим, что не существует метода, который сразу непосредственно приведет к глобально оптимальному решению задачи, поскольку минимизация суммы экспонент — сложная проблема. На практике уравнение степенной регрессии часто переписывают в виде уравнения линейной регрессии путем применения операции логарифмирования. При этом $\log(x)$ становится входным, а $\log(y)$ — выходным признаком: $\log(y) = b\log(x) + \log(a)$. При такой линеаризации x_i и y_i заменяются на $v_i = \log(x_i)$ и $z_i = \log(y_i)$, затем осуществляется линейная регрессия z_i по v_i, после чего найденные коэффициенты преобразуются в коэффициенты для исходной степенной функции. Этот способ часто дает высокое значение коэффициенту детерминации, поскольку логарифмирование сильно сглаживает данные.

Однако надо понимать: то, что найденные коэффициенты оптимальны для уравнения линейной регрессии, не обязательно означает, что при обратном переходе к степенно́й записи они также будут оптимальными. Данный проект посвящен иллюстрации этого утверждения.

Рассмотрим задачу минимизации суммы квадратов невязок согласно исходной степенной регрессии без ее линеаризации. Эта задача, скажем так: «не по зубам» классическому математическому подходу. Для ее решения мы применим подход, становящийся все более популярным — оптимизацию сложной функции методом, навеянным природными процессами. Вместо получения единственного решения путём его последовательного улучшения, как это делается в классических методах оптимизации, этот подход использует целое множество, так называемую «популяцию», решений, которая итеративно эволюционирует от поколения к поколению согласно правилам, имитирующим природу. Обычно такие правила включают: (а) небольшие случайные, но целенаправленные, изменения от поколения к поколению; (б) принципы отбора наилучших из найденных решений в так называемую «элиту». После того как осуществлено заранее заданное число итераций, одно из элитных решений выдается как результат данной процедуры.

Прежде чем начать эволюционный процесс оптимизации, необходимо определить границы области допустимых решений так, чтобы ни один член популяции не выходил за ее пределы. Это гарантирует, что в процессе эволюции популяция не «взорвется» путем устремления решения в бесконечность. В данном случае предлагается следующее рассуждение. В условиях гипотезы степенной зависимости переменных $y = ax^b$, для любых двух объектов i и j должны

выполняться следующие равенства: $z_i = b \cdot v_i + c$ и $z_j = b \cdot v_j + c$, где $c = \log(a)$, $z_i = \log(y_i)$ и $v_i = \log(x_i)$. Тогда неизвестные b и c могут быть выражены через известные z_i, v_j, z_i, v_j: $b = (z_i - z_j)/(v_i - v_j)$, $c = (v_i \cdot z_j - v_j \cdot z_i)/(v_i - v_j)$, что может привести к различным значениям b и c при различных i и j. Для тех i и j, для которых $v_i - v_j \neq 0$, обозначим минимальное и максимальное значения отношения $(z_i - z_j)/(v_i - v_j)$ через bm и bM, соответственно, а минимальное и максимальное значения $(v_i \cdot z_j - v_j \cdot z_i)/(v_i - v_j)$ через cm и cM. Допустимые b и c должны находиться в этих пределах, что позволяет задать область допустимых решений неравенствами $(bm, cm) \leq (b, c) \leq (bM, cM)$. При этом оптимальные значения (b, c) не должны сильно отклоняться от средних значений определенных выше отношений, т. е. находиться ближе к центру данного прямоугольника, чем к его границам. Кроме того, вычисления пойдут гораздо быстрее, если мы ограничимся только теми парами (i, j), для которых v_i, v_j и z_i, z_j не слишком близки к 0 из-за большой чувствительности логарифма в зоне 0. Все эти соображения учтены в программном коде ddr.m для MATLAB, который можно найти в Приложении.

Чтобы определить правила перехода от «текущего» поколения к следующему, обозначим массив популяции размера $2p$ на текущей итерации через f, а через f' — массив популяции на следующей итерации. Переход от f к f' осуществляется в три этапа. Во-первых, возьмем ряд средних значений по столбцам f и повторим его p раз в массиве mf размера $p \times 2$. Затем сделаем случайный Гауссов сдвиг:

```
fn = f + randn(p,2).*mf/20
```

Здесь randn(p,2) — массив размера $p \times 2$ (псевдо)-случайных чисел из Гауссова распределения $N(0, 1)$ с нулевым математическим ожиданием и единичной дисперсией. Символ «.*» обозначает операцию «умножения» матриц путем умножения друг на друга соответствующих элементов в матрицах, так что $(a_{ij}).*(b_{ij})$ — это матрица, (i, j)-ый элемент которой равен $a_{ij}*b_{ij}$. Эта случайная матрица масштабируется долей двумерного вектора средних, повторенного p раз в $p \times 2$ матрице $mf/20$, так что сдвиг составляет около 5 % от средних значений f.

Поскольку вышеопределенное изменение f допускает выход за допустимые границы, каждый a-элемент (первый столбец fn), больший, чем aM, должен быть заменен на aM, и каждый a-элемент, меньший, чем am, заменяется на am. Подобная замена производится и для b-элементов. Обозначим результат через fr.

Теперь возьмем массив el, тоже размера $p \times 2$, в строках которого — значения (a, b) для отобранной ранее элиты популяции, и определим популяцию следующего поколения как результат смешивания fr и el:

$$f' = 0.7fr + 0.3el.$$

Такое смешивание сдвигает популяцию *fr* в направлении элиты *el*, т. е. наилучшего найденного к данной итерации решения, на 30 %. Эксперименты показали, что такой сдвиг хорошо работает в данной задаче.

Под элитой понимается та пара популяции (a, b), на которой минимизируемый критерий достигает минимального значения. Элита пересчитывается на каждой итерации так же, как и в Проекте 2.2. А именно, найдем значения критерия для всех пар (a, b) нового поколения, выберем наилучшую и наихудшую пары (a', b') и (a'', b'') и сравним их с элитой (a, b). Если (a', b') лучше, чем (a, b), запомним (a', b') в качестве элиты (a, b). Если (a', b') и, тем более, (a'', b'') хуже, чем элита (a, b), заменим (a'', b'') в текущей популяции на (a, b).

Описанная процедура реализована на MATLAB в программном коде nlr.m, который использует вышеописанную программу ddr.m (см. раздел А4 Приложения).

Задание 3.4. Применение линеаризации и прямой минимизации к одним и тем же данным

Сформируем двумерную таблицу данных. Входной признак x создадим как 50-мерный вектор случайных положительных величин от 0 до 10: x = 10 · rand(1,50), а выходной признак определим как $y = 2 \cdot x^{1.07}$. Добавим к y нормальную ошибку из распределения $2 \cdot N(0,1)$, математическое ожидание которого равно 0, а стандартное отклонение 2, причем так, чтобы результат не мог стать меньше, чем 1 (это необходимо для того, чтобы логарифмирование в программе ddr.m не приводило к отрицательным числам), при помощи следующей строчки в коде:

```
>>for ii = 1:50; yy = 2*x(ii)^1.07 + 2*randn;
  y(ii) = max(yy,1.01); end;
```

Признак *yy* моделирует тренд 7%-ного роста, сильно зашумленного Гауссовой ошибкой.

Задача — оценить неизвестные параметры модели $y = ax^b$ по этим данным. Для этого можно применить оба подхода: (А) регрессионную модель линеаризации и (Б) эволюционный процесс, инспирированный природой.

(А) Применим логарифмическое преобразование уравнения $y = ax^b$ и построим линейную регрессию $\log(y)$ через $\log(x)$. Это приводит (переходом к экспоненте для найденного c) с помощью программы llr.m к $a = 3.0843$ и $b = 0.8011$, т. е. модели $y = 3.08x^{0.8}$.

При этом средняя ошибка квадрата невязки $y - ax^b$ равна 4.41, так что стандартная ошибка равна 2.10, т. е. приблизительно 20 % от среднего значения y, 10.1168. Это говорит не только о том, что ошибка высока, но и о том, что сам степенной закон оказался найден неверно. У заданной функции темп $b = 1.07 > 1$, а у полученной функции ниже единицы ($b = 0.80 < 1$).

(Б) Используем код nlr.m, реализующий вышеописанный эволюционный алгоритм минимизация средней ошибки квадрата $y - ax^b$ для исходной модели. Это дает $a = 2.0293$ и $b = 1.0760$, причем средний квадрат ошибки составляет 0.0003, а стандартное отклонение равно 0.0180. В противоположность значениям, найденным для линеаризованной модели, полученные эволюционно значения параметров a и b довольно близки к тем, которые использовались при генерации данных.

Пример показывает, что процедура линеаризации может приводить к совершенно неправильным оценкам. Поэтому лучше ею не пользоваться. Более точные результаты получаются путем минимизации критерия исходной нелинеаризованной модели с помощью эволюционного алгоритма.

Задание 3.5. Моделирование роста инвестиций

Применим рассмотренный в Проекте 3.2 подход к переменным x и y, значения которых для 20 моментов времени представлены в табл. 3.5.

Таблица 3.5

Величина инвестиционного фонда y в моменты времени x в промежутке 0.10—2.00.

x	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
y	1.30	1.82	2.03	4.29	3.30	3.90	3.84	4.24	4.23	6.50
x	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90	2.00
y	6.93	7.23	7.91	9.27	9.45	11.18	12.48	12.51	15.40	15.91

Переменная x отражает движение времени, а y — размеры фонда. Эти данные получены следующим образом. Компоненты x — это натуральные числа от 1 до 20, деленные на 10. Значения y получены в MATLAB по формуле $y = 2*\exp(1.04*x)+0.6*randn$, где randn — нормальная (Гауссова) случайная величина с математическим ожиданием 0 и дисперсией 1. Задача — определить тренд динамики фонда.

Для начала применим традиционный подход к определению темпа роста инвестиционного фонда в течение всего периода времени. Согласно этому подходу, средний рост инвестиций за 19 шагов выражается корнем 19 степени (степенью 1/19) из отношения $y_{20}/y_{01} = 15.91/1.30 = = 12.238$. Корень 19-й степени из этой величины равен 1.1409, что соответствует среднему годовому приросту 14 %, что гораздо выше, чем 4 %, использованные при генерации данных. Традиционный подход в данном случае оказывается не очень-то применимым.

Попытаемся теперь оценить связь между y и x, применяя процедуру линеаризации из раздела Ф.3.3.4, где рассматривалась экспоненциальная зависимость $y = ae^{bx}$ (в Проекте 3.2 зависимость — степенная). Задача линейной регрессии для линеаризованной

зависимости приводит к значениям коэффициентов $b = 1.1969$ и $c = 0.4986$. Переходя обратно к экспонентам, получаем значения для коэффициентов $a = e^c = 1.6465$ и $b = 1.1969$.

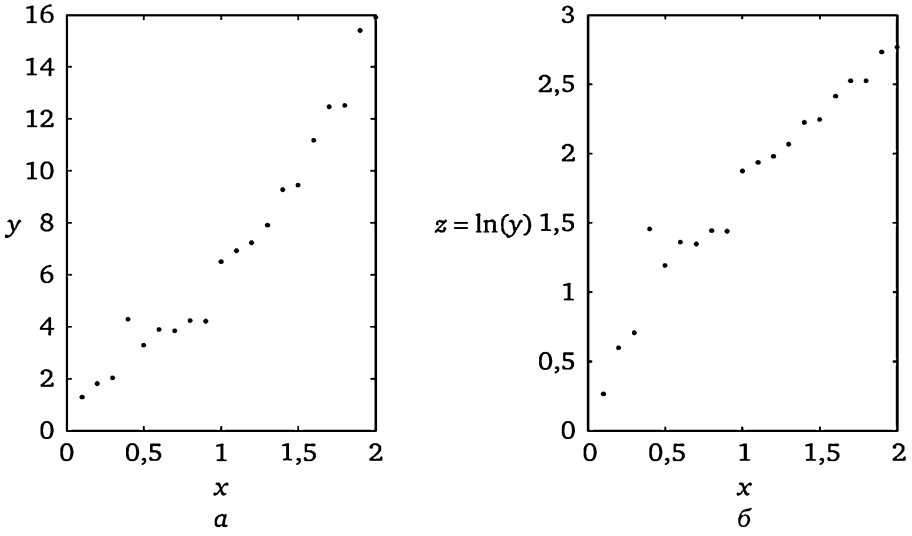


Рис. 3.15. **График пар (*x*, *y*):**
а — *y*, зашумленная экспоненциальная функция от *x*; *б* — график пар (*x*, *z*), где $z = \ln(y)$. Правый график выглядит несколько более линейным, хотя коэффициенты корреляции для данных полей рассеяния близки по своим значениям: 0.970 для левого графика и 0.973 для правого.

Легко заметить, что эти значения отличаются от истинных значений $a = 2$ и $b = 1.04$ приблизительно на 15—20 %. Средняя величина квадрата ошибки $E = 0.700$.

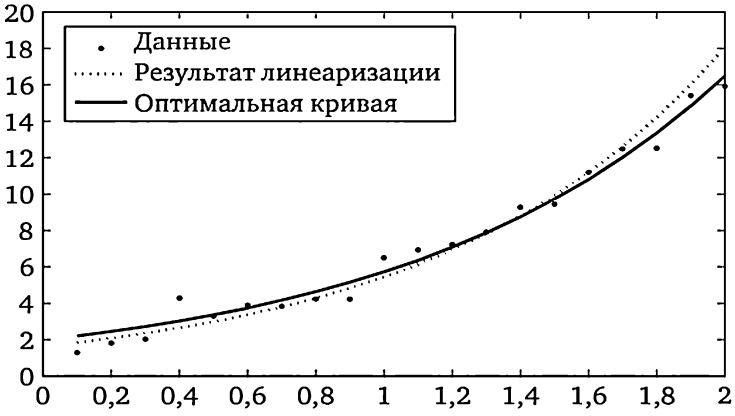


Рис. 3.16. **Два экспоненциальных приближения для задания 3.4**

Теперь исходную нелинейную задачу проанализируем с применением процесса, инспирированного природой.

Программа nlr.m из Приложения, реализующая эволюционный подход, описанный в Проекте 3.2, модифицированная таким образом, что вычисляющая функцию строчка в подпрограмме delta заменена на yp(ii) = a*exp(b*x(ii));, приводит к $a = 1.9908$ и $b = 1.0573$. Отличие от истинных значений $a = 2$ и $b = 1.04$ составляет не более 2 %. Средний квадрат ошибки E = 0.373 гораздо ниже среднего квадрата ошибки уравнения, найденного с помощью линеаризации.

Оба полученных решения отражены на рис. 3.16. Можно видеть, что результат линеаризации даёт более крутую экспоненту, что особенно заметно в более поздние периоды времени.

Вопрос 3.7. Рассмотрим бинарный признак, значения которого известны для 7 объектов: первые три из них принадлежат категории *A*, остальные четыре — категории *B*. Зададим две фиктивные 1/0 (бинарные) переменные *xA* и *xB*, так что *xA* = 1 для первых трех объектов и *xA* = 0 для остальных четырех, при этом *xB* = 0 для первых трех и *xB* = 1 для остальных объектов. Что можно сказать о коэффициенте корреляции между *xA* и *xB*?

Ответ. Коэффициент корреляции *xA* и *xB* равен −1, так как *xA* + + *xB* = 1 для всех объектов, так что имеет место линейная связь *xA* = = −*xB* + 1.

Вопрос 3.8. Подумайте, как распространить алгоритм, инспирированный природой, на задачу определения коэффициентов линейной регрессии с нетрадиционным критерием, таким как средняя относительная ошибка, заданная формулой

$$\frac{1}{N}\sum_{i=1}^{N}|e_i / y_i|.$$

Ответ. Для этого достаточно поменять три-четыре строчки в кодах вспомогательных функций esq = delta(tt,x,y) и [ab,bb] = ddr(x,y) программы nlr.m. В первую надо вставить вычисление относительной ошибки вместо квадратичной, а также вычисление целевого значения по линейной регрессии вместо степенной. Во второй надо изменить проверяемые равенства (см. стр.119): они должны относиться к исходным признакам, а не логарифмам от них.

3.4. Случай смешанных шкал: номинальный и количественный признаки

3.4.1. Целевой количественный признак: табличная регрессия

Рассмотрим неколичественный признак *x* на том же множестве объектов, что и количественный признак *y*. В качестве такой пары можно взять, например, признаки Тип протокола и SH, Количество

соединений источника за последние две секунды, в таблице данных о компьютерных атаках.

Распределения *y* внутри категорий *x* могут быть использованы для изучения связи между *x* и *y*. Распределения могут быть визуализированы с использованием лишь диапазона признаков следующим образом: на оси *x* изобразим категории отдельными бинами, и проведем две линии параллельно оси *x*, чтобы отобразить минимальное и максимальное значения признака *y* (на всем множестве данных). Затем представим диапазоны значений *y* внутри каждой категории так, как показано на рис. 3.17.

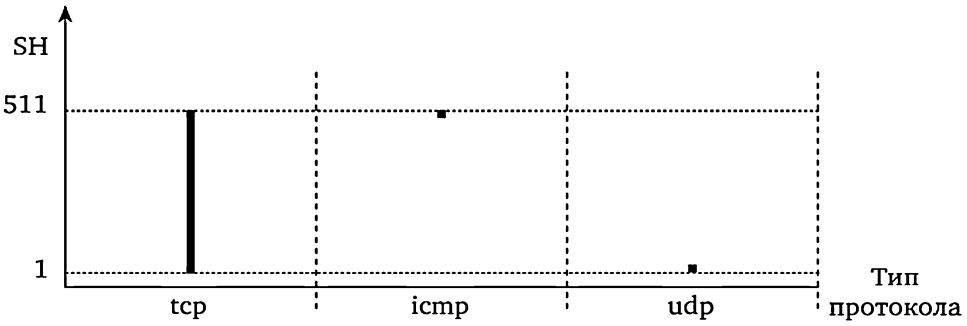


Рис. 3.17. **Графическое представление диапазонов количества соединений SH для различных типов протокола**

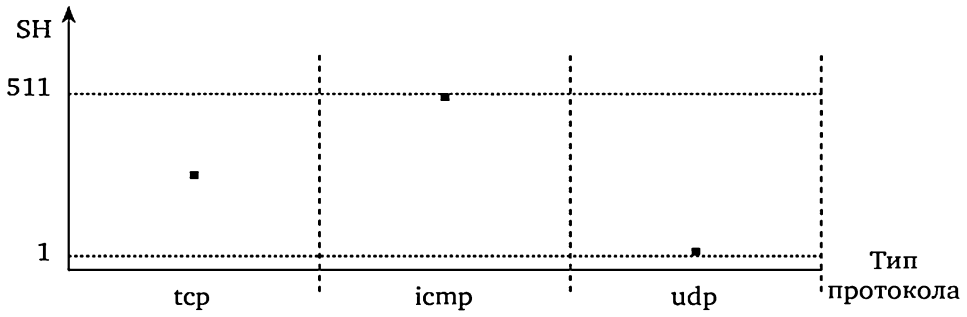


Рис. 3.18. **В случае абсолютной корреляции с нулевой дисперсией внутри категории знание типа протокола обеспечит точный прогноз количества соединений SH для каждого типа протокола**

Уровень связи между *x* и *y* тем выше, чем меньше разброс *y* внутри категорий *x*. На рис. 3.18 приведен идеальный случай полной корреляции — все величины *y* внутри категорий одинаковы, что позволяет дать абсолютно точный прогноз количества соединений SH для каждого типа протокола.

Рис. 3.19 представляет еще один крайний случай, когда знание типа протокола не может дать никакого уточнения в прогнозировании количества соединений SH.

В статистике принято информацию о связи количественного и качественного признаков представлять в виде так называемой табличной регрессии.

Табличная регрессия *y* по *x* — это трехстолбцовая таблица, строки которой соответствуют категориям *x*. В каждой размещается следующая информация о категории:

1) численность категории;
2) среднее значение y внутри категории;
3) стандартное отклонение *y* внутри категории.

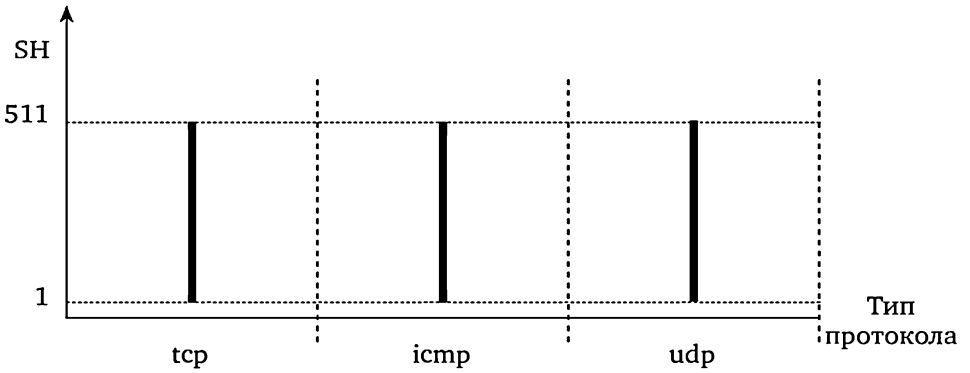


Рис. 3.19. **Распределения внутри категорий: случай, когда знание типа протокола не дает никакой дополнительной информации о количестве соединений**

Последняя, «маргинальная», строка содержит такую же информацию для всего множества объектов.

Рабочий пример 3.4

Табличная регрессия количества соединений SH (количественный целевой признак) по типу протокола (категоризованный входной признак) для данных о компьютерных атаках

Построим табличную регрессию количества соединений SH по типу протокола (см. табл. 3.6).

Таблица 3.6

Табличная регрессия количества соединений SH (количественный целевой признак) по типу протокола (категоризованный входной признак) для данных о компьютерных атаках

Тип протокола	Численность	SH, среднее	SH, стд. откл.
Tcp	64	98.98	177.70
Icmp	10	508.40	5.13
Udp	26	2.15	1.38
Всего	100	114.75	198.09

Эта таблица позволяет спрогнозировать количество соединений пакета, зная тип протокола. Например, для udp среднее количество соединений составляет 2.15 плюс-минус 1.38. Не имея никаких сведений о типе протокола, можно лишь утверждать, что количество соединений пакета в среднем 114.75 плюс-минус 198.09 — это значительно менее точная оценка.

Если количественный признак y задан без какой-либо дополнительной информации, то среднее значение этого признака, определяемое по формуле $\bar{y} = \sum_{i \in I} y_i / |I|$, представляет собой разумную суммаризацию имеющихся данных. Однако если становятся известными категории номинального признака x, то можно получить более детальную информацию: средние значения y в категориях. Обозначим через S_k множество объектов в категории k признака x. Среднее значение внутри этой категории равно $\bar{y}_k = \sum_{i \in S_k} y_i / |S_k|$.

Эти средние можно считать решением уравнения табличной регрессии по методу наименьших квадратов. Подход восстановления данных применительно к этой ситуации можно сформулировать следующим образом. Найдем внутригрупповые центральные значения c_k так, чтобы минимизировать суммарную квадратичную ошибку $L = \sum_{i \in I} e^{-2}$, где $e_i = y_i - c_k$ — невязка уравнения

$$y_i = c_k + e_i, \; i \in S_k, \tag{3.15}$$

«декодирующего» каждое наблюденное значение y характеристическим числом c_k, представляющим категорию k ($k = 1, 2, \ldots, K$).

Это уравнение лежит в основе табличной регрессии, называемой также «кусочно-постоянной» регрессией. Нетрудно показать, что оптимальная по критерию наименьших квадратов величина c_k в (3.15) равна среднему значению внутри категории $\bar{y}_k$. Отсюда следует, что минимальное значение критерия L равно $L_m = \sum_{k=1}^{K} \sum_{i \in S_k} (y_i - \bar{y}_k)^2$. Разделив и умножив внутреннюю сумму на $|S_k|$ — число элементов в множестве S_k, можно увидеть, что $L_m = N\sigma_w^2$, где σ_w^2 — внутригрупповая (*within-group*) дисперсия, определяемая формулой средневзвешенного среднего дисперсий y во всех группах:

$$\sigma_w^2 = \sum_k p_k \sigma_k^2, \tag{3.16}$$

где $p_k = |S_k| / N$ — доля категории k, а σ_k^2 — дисперсия y_k в S_k.

Для дальнейшего анализа рассмотрим тождество:

$$(y_i - \bar{y}_k)^2 = y_i^2 + \bar{y}_k^2 - 2 y_i \bar{y}_k$$

и просуммируем его по всем $i \in S_k$:

$$\sum_{i \in S_k} (y_i - \bar{y}_k)^2 = \sum_{i \in S_k} y_i^2 - |S_k| \bar{y}_k^2.$$

Суммируя эти уравнения по k и перенося последнее выражение из правой части уравнения в левую, получим:

$$\sum_{i\in I} y_i^2 = \sum_{k=1}^{K} |S_k| \bar{y}_k^2 + \sum_{k=1}^{K} \sum_{i\in S_k} (y_i - \bar{y}_k)^2. \quad (3.17)$$

Справа в уравнении (3.17) стоит сумма квадратов невязок модели (3.15) L_m. Это позволяет интерпретировать уравнение (3.17) как декомпозицию квадратичного разброса переменной y: эта величина (слева) разделяется на два слагаемых (справа), называемых объясненной и необъясненной частями квадратичного разброса.

Объясненная часть суммирует вклады $|S_k|\bar{y}_k^2$ отдельных категорий k. Величина вклада пропорциональна и частоте категории, и квадрату среднего: чем больше эти значения, тем выше вклад. Еще одна интерпретация декомпозиции (3.17) может быть сделана, если признак y центрирован, так что его среднее значение равно нулю. В этом случае, разделив уравнение (3.17) на N, получим:

$$\sigma^2 = \sum_{k=1}^{K} p_k \bar{y}_k^2 + \sum_{k=1}^{K} p_k \sigma_k^2, \quad (3.18)$$

где σ^2 — дисперсия y, самая правая сумма — внутригрупповая дисперсия σ_w^2 (3.16), а левая сумма — это взвешенная сумма квадратов расстояний между общим средним $\bar{y} = 0$ и средними внутри-групповыми значениями $\bar{y}_k$.

В статистике уравнение (3.18) хорошо известно как разложение дисперсии на внутригрупповую и межгрупповую составляющие. Оно лежит в основе широко используемого метода сравнения средних внутригрупповых значений, называемого *дисперсионный анализ*, по-английски ANOVA (ANalysis Of VAriance). В контексте модели табличной регрессии (3.15), имеющей смысл модели восстановления данных, разложение (3.17) представляется более подходящим.

Корреляционное отношение характеризует среднее уменьшение дисперсии признака y при прогнозировании его величины с помощью (3.15), или, иными словами, относительную долю объясненной части дисперсии. Корреляционное отношение обычно обозначается η^2 и определяется следующей формулой:

$$\eta^2 = 1 - \sigma_w^2 / \sigma^2. \quad (3.19)$$

Из определения непосредственно вытекают следующие свойства этой величины:

— η^2 принимает значения в интервале от 0 до 1.

— $\eta^2 = 1$, если и только если все внутригрупповые дисперсии равны нулю, $\sigma_k^2 = 0$ (т. е. когда y постоянно внутри каждой группы S_k).

— $\eta^2 = 0$, если и только если все σ_k^2 порядка σ^2

Вопрос 3.9. Рассмотрим два количественных признака x и y. Область значений x разделим на пять интервалов одинакового размера

для задания категориальной переменной *xc*. Существует ли какая-либо связь между коэффициентом корреляции *x* и *y* и величиной корреляционного отношения *xc* и *y*?

Ответ. Прямой связи нет, можно привести случаи, когда коэффициент корреляции больше величины корреляционного отношения, и случаи, когда меньше.

Самостоятельная работа

3.6. Постройте уравнение регрессии признака «Ширина лепестка» по признаку «Длина лепестка» по данным об ирисах (см. табл 1.2), сравните прогнозируемые значения «Ширины лепестка» с наблюденными значениями и рассчитайте значения относительных ошибок. Вычислите среднее значение относительной ошибки.

3.7. Постройте уравнение регрессии признака «Ба» (число отделений банков) по признаку «Нас» (численность населения) по данным о городах английского побережья (см. табл 1.5), сравните прогнозируемые значения «Ба» с наблюденными значениями и рассчитайте значения относительных ошибок. Вычислите среднее значение относительной ошибки.

Данные табл. 3.6 можно визуализировать с указанием средних значений и стандартных отклонений внутри каждой категории, однако в анализе данных более популярен иной способ визуализации, так называемый *бокс-плот*.

Бокс-плот в MATLAB формируется следующим образом. При заданном уровне квантиля, по умолчанию — 25 %, диапазон (a1, a2) значений количественного признака от нижнего 25 % квантиля a1 до верхнего 25 % квантиля a2 представляется прямоугольником — «боксом», расположенным вдоль оси ординат от уровня a1 до уровня a2. 50%-ный квантиль, т. е. медиана, изображается сплошной линией внутри прямоугольника. Полный диапазон изменения признака отмечается вертикальной линией, заканчивающейся «усиками», соответствующими минимальному и максимальному значениям. При этом из диапазона удаляются так называемые выбросы, т. е. значения признака, удаленные от медианы более, чем на полтора интервала a2-a1, покрываемого прямоугольником. Именно такие значения оказываются дальше от среднего, чем «три сигма», в случае, когда распределение признака Гауссово. Эти значения помечаются в MATLAB крестиками.

Рабочий пример 3.5

Бокс-плот количества соединений SH (количественный целевой признак) по типу протокола (категоризованный входной признак) для данных о компьютерных атаках

Построим с помощью MATLAB бокс-плот признака SH в категориях признака «Типа протокола» в данных о компьютерных атаках. Для этого по данным табл. 1.4 переформатируем ее столбцы 1—3, содержащие фик-

тивные переменные для различных типов, в номинальный признак gr со значениями tcp, icmp, udp:

```
>> x = load('smalln.dat'); % загрузка файла,
                           % содержащего данные табл. 1.4
>> g = {'tcp','icmp','udp'};  % формирование переменной
                              % с названиями протоколов
>> n = 100; for k = 1:3; for j = 1:n; ...
     if x(j,k) == 1, gr{j} = g{k}; end; end;
% 100x1 массив gr содержит названия протоколов,
% соответствующих каждому пакету
```

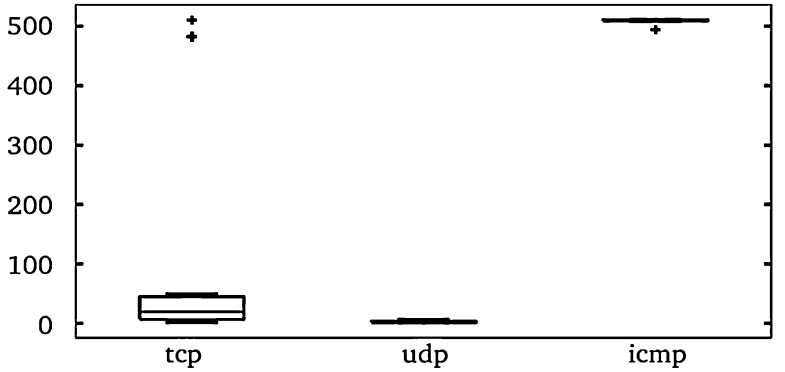


Рис. 3.20. **Бокс-плот, отображающий связь типа протокола с SH с точностью до 25 % квантилей:** высота прямоугольников — боксов — характеризует диапазон признака для 50 % его значений в соответствующей категории. Полные интервалы значений показаны «усиками». Горизонтальные отрезки отражают медианы признака SH внутри категорий

Теперь формируем бокс-плот признака SH в категориях признака gr:

```
>> boxplot(x(:,5), gr);
% SH – в 5-м столбце массива x,
% см. рис. 3.20.
```

Расположение бокса категории tcp не соответствует информации о признаке SH для этой категории в табл. 3.6, где указано среднее значение 98.98 с вдвое превышающим его стандартным отклонением. Проверим, как это получилось. Отсортируем по возрастанию от 1 до 510 все 64 значения SH, попавшие в категорию tcp, и определим нижнюю и верхнюю 25 % квантили. Это 17-е и 48-е значения отсортированного ряда, так как 25 % в данном случае — ровно 16 объектов. Они равны, соответственно, 7 и 44. Медиана вычисляется как среднее значений 19 и 20 на 32-м и 33-м объектах соответственно, т. е. 19.5. Сильное отличие медианы от среднего объясняется присутствием в категории нескольких очень высоких значений, которые отмечены на рис. 3.20 плюсами.

Рабочий пример 3.6

Анализ связи между типом атаки и количеством соединений SH для данных о компьютерных атаках

Построим с помощью MATLAB бокс-плот признака SH в категориях признака Тип атаки по данным о компьютерных атаках. Это делается так же, как и в предыдущем примере (рис. 3.21).

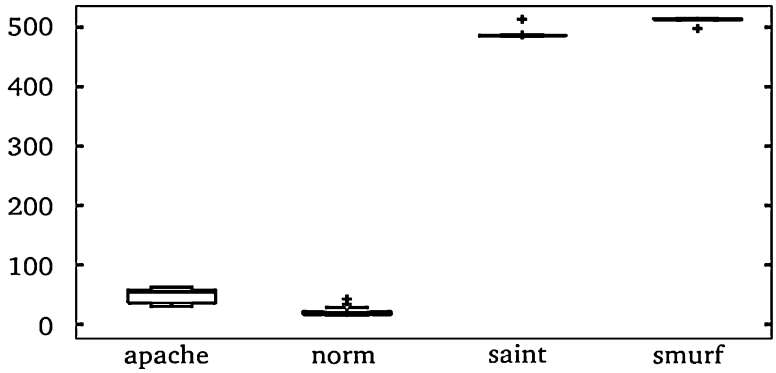


Рис. 3.21. **Бокс-плот, отображающий связь типа атаки с SH (расшифровка та же, что и на рис. 3.20)**

Интуитивно чувствуется, что связь Типа атаки с SH выше, чем связь Типа протокола: боксы на рис. 3.21 выглядят тоньше, чем на рис. 3.20. Это указывает на то, что прогнозировать значения признака SH, зная тип атаки, можно точнее, чем зная тип протокола.

Вычислим табличную регрессию SH по типу атаки (табл. 3.7).

Таблица 3.7

Табличная регрессия количества соединений SH (количественный целевой признак) по типу атаки (категоризованный входной признак) для данных о компьютерных атаках

Тип атаки	Численность	SH, среднее	SH, ст. откл.
Apache	23	33.61	12.13
Norm	56	5.12	5.59
Saint	11	484.64	8.42
Smurf	10	508.40	5.13
Всего	100	114.75	198.09

Эта таблица, в общем, подтверждает мнение о том, что тип атаки больше связан с SH, чем тип протокола, так как стандартные отклонения в группах одинаковых атак (табл. 3.7) в общем-то меньше, чем в группах одинаковых протоколов (табл. 3.6).

Но как установить это с непреложной точностью? К сожалению, в анализе данных точность и полнота могут противоречить друг другу. Табличная регрессия отражает так называемую кусочно-постоянную модель связи категоризованного и количественного признака. Применение метода наи-

меньших квадратов к оценке этой модели приводит к коэффициенту корреляционного отношения, показывающему, какую долю дисперсии количественного признака позволяют учитывать категории номинального признака. Этот-то коэффициент и принимается часто за характеристику силы связи номинального и количественного признака. Коэффициент корреляционного отношения оценивает, насколько дисперсия внутри групп меньше, чем дисперсия значений признака до его разбиения на категории; это аналог коэффициента детерминации для обычной регрессии.

Рабочий пример 3.7

Корреляционное отношение

Вопрос: выше ли связь между признаками в табл. 3.6 по сравнению с табл. 3.7? Чтобы ответить, рассчитаем корреляционное отношение для табл. 3.6, 3.7 по формулам (3.19) и (3.16):

— количество соединений SH/Тип протокола — 48.5 %;

— количество соединений SH/Тип атаки — 99.8 %.

Уменьшение дисперсии количественного признака в категориях качественного признака, выраженное корреляционным отношением, выше во второй таблице, т. е. действительно корреляция между типом атаки пакета и количеством в нем соединений SH выше, чем между этим последним и типом протокола. Более того, величина коэффициента корреляционного отношения в первом случае составляет почти единицу — внутригрупповые дисперсии SH в этом случае значительно ниже, чем общая дисперсия. Напротив, для протокола tcp внутригрупповая дисперсия сравнима по величине с общей дисперсией, что и предопределяет значительно меньшую связь между SH и типом протокола.

Вопрос 3.10. Влияют ли уровни величины внутригрупповых средних на величину корреляционного отношения?

Ответ. Нет; уровни величины средних значений не имеет отношения к уровню связи, измеряемому корреляционным отношением. Уровень связи в табл. 3.7 выше, чем в табл. 3.6, из-за меньшего разброса количественного признака в категориях табл. 3.7.

3.4.2. Номинальный целевой признак

Если прогнозирование ведется по количественной переменной, а прогнозируемый признак категориальный, можно использовать любой из многочисленных методов распознавания образов или машинного обучения. Задача может быть сформулирована по-разному. В задачах машинного обучения, как правило, множество объектов случайно разделено на два подмножества, «обучение» и «тест». На «обучении» целевой признак известен, на «тесте» — нет. Надо использовать обучение для выработки правила, позволяющего прогнозировать значения целевого признака на тесте, конечно, при условии,

что значения входных признаков известны. Пользователь оценивает качество правила, сличая прогнозы на тесте с известными ему значениями выходного признака — чем больше совпадений, тем выше качество прогноза. Применительно к рассматриваемому случаю речь идет о выработке правила, позволяющего по значению входного количественного признака прогнозировать категории выходного номинального признака. Такое правило часто называют *классификатором*.

Рассмотрим два подхода к решению задачи.

3.4.2.1. Классификатор по правилу ближайшего соседа

Один из наиболее популярных методов — классификатор по правилу ближайшего соседа [*nearest neighbor* (NN) *classifier*]. Он применим к любым данным, для которых можно задать расстояние или иную меру сходства/различия между объектами. NN-классификатор работает следующим образом: берётся объект из «обучения», наиближайший к рассматриваемому объекту, который относится к той категории, к которой принадлежит этот наиближайший объект. Можно посмотреть на результаты применения NN-классификатора к данным о компьютерных атаках и об ирисах на примерах табл. 3.8 (атаки) и табл. 3.10 (ирисы). Результаты сильно отличаются: для данных об атаках получены хорошие результаты, см табл. 3.8, в то время как для данных об ирисах — значительно хуже (табл. 3.10). Объяснение этому — различие в уровне связи между признаками — очень высокий в одном случае (табл. 3.7) и не очень высокий в другом (табл. 3.9).

NN-классификатор легко распространить на так называемый k-NN-классификатор, который определяет категорию признака на основе большинства из k ближайших соседей рассматриваемого объекта. Этот классификатор также имеет возможность «отказаться» от прогноза, когда правило большинства не дает четкого победителя.

Рабочий пример 3.8

Классификатор по наиближайшему соседу

Рассмотрим два признака из таблицы данных о компьютерных атаках: тип атаки Att — как целевой признак, и количество подключений к текущему хосту за последние 2 секунды — входной признак SH. Для ускорения работы метода отсортируем объекты по возрастанию SH.

Таблица 3.8

Применение NN-классификатора SH ⇒ «Тип атаки» к случайной выборке из множества данных о компьютерных атаках

Случайная выборка 10 объектов	9 29 37 51 63 70 72 80 86 89
Категория целевого признака	apa nor nor nor nor nor nor sai sai sai
Значение входного признака	24 10 1 14 2 3 1 482 482 483
Значение наиближайшего соседа	23 11 1 13 2 3 1 482 482 482
Категория наиближайшего соседа	apa nor nor nor nor nor nor sai sai sai

Далее случайным образом выберем подмножество из 10 элементов (верхняя строка в табл. 3.8) вместе со значениями Att для них (вторая строка) и величинами SH (третья строка) — эти объекты относятся к «экзамену», а остальные 90 — к обучению. Выберем обучающие объекты, для которых значения SH близки к величинам в третьем ряду, и запишем их в четвертый ряд. Также в пятую строку запишем значения признака Att для объектов в четвертом ряду (самый нижний ряд). Поразительный успех: все 10 предсказаны верно!

Таблица 3.9

Табличная регрессия длины чашелистика по таксону по данным об ирисах

Таксон	Количество	Среднее	Стд. откл
Таксон 1 Таксон 2 Таксон 3	50 50 50	5.00 5.94 6.59	0.35 0.52 0.64
Всего	150	5.84	0.83
Корреляционное отношение	0.6135		

Стандартное отклонение внутри категорий сравнительно невелико, порядка 10 % от среднего. Таблица объясняет порядка 61 % дисперсии длины чашелистика.

Вопрос 3.11. Постройте табличную регрессию длины чашелистика по таксону и найдите корреляционное отношение.

Ответ. См табл. 3.9.

Вопрос 3.12. Примените NN-классификатор для прогнозирования таксона по длине чашелистика из таблицы данных об ирисах.

Ответ. См табл. 3.10.

Таблица 3.10

Применение NN-классификатора «Длина чашелистика ⇒ Таксон» к случайной выборке из таблицы данных об ирисах**

Случайная выборка цветов ириса	123	99	32	40	22	34	92	91	146	119
Таксон	T3	T2	T1	T1	T1	T1	T2	T2	T3	T3
Длина чашелистика	6.7	6.1	5.0	5.4	4.8	5.4	5.5	5.5	7.3	6.0
Наиближайший сосед	6.7	6.1	5.0	5.4	4.8	5.4	5.5	5.5	7.4	6.0
Таксон наиближайшего соседа	**T2***	T2*	T1*	T1*	T1*	T1*	**T1***	**T1***	**T2**	**T2***

* Выбрана самая частая категория (из нескольких).

** Неверный прогноз (5 из 10) выделен жирным шрифтом.

Вопрос 3.13. Рассмотрим таблицу данных о 8 студентах с двумя признаками:

Студент	Оценка	Профессия
1	50	ИТ
2	80	ИТ
3	80	ИТ
4	60	БА
5	60	БА
6	40	БА
7	50	БА
8	40	БА

(i) Постройте регрессионную таблицу для прогнозирования оценки по профессии.

(ii) Спрогнозируйте оценку нового студента профессии БА.

(iii) Найдите корреляционное отношение для этой таблицы.

Ответ. (i) Табличная регрессия оценки по профессии. Строки соответствуют категориям профессии, их частостям, а также средней оценке и её стандартному отклонению внутри каждой категории:

Профессия	Оценка		
	Количество	Среднее	Стандартное отклонение
ИТ	3	70	14.1
БА	5	50	8.9

(ii) Для студента профессии БА наиболее вероятная оценка лежит в интервале 50±8.9.

(iii) Корреляционное отношение определяется взвешенной дисперсией в категориях: $(3 \cdot 14.1 + 5 \cdot 8.9)/8 = (42.3 + 44.5)/8 = 10.85$, и общей дисперсией для всех выборочных данных с средним = 57.5, которая равна 14.79. Корреляционное отношение $\eta^2 = 1 \cdot 10.85/14.79 = 0.266$. Это означает, что данная табличная регрессия объясняет лишь 26.6 % дисперсии оценки.

Вопрос 3.14. Постройте табличную регрессию длины лепестка по таксону, используя таблицу данных об ирисах, и найдите корреляционное отношение.

Ответ. Очень высокий уровень корреляционного отношения, $\eta^2 = 0.94$, в табл. 3.11 определяется, по-видимому, тем, что внутригрупповые стандартные отклонения количественного признака значительно меньше, чем его стандартное отклонение на всем множестве.

Таблица 3.11

Табличная регрессия длины лепестка w3 по таксону по данным об ирисах

Таксон	Количество	Среднее	Ст. откл.
T1	50	1.46	0.17
T2	50	4.26	0.47
T3	50	5.55	0.55
Всего	150	3.76	1.77
Корреляционное отношение	0.9406		

Самостоятельная работа

3.8. Постройте табличную регрессию признака «Нас» (Численность населения) по признаку «Бас» (Число бассейнов) по данным о городах английского побережья (см. табл. 1.5), а также величину корреляционного отношения. Дайте интерпретацию этой величины.

3.4.2.2. Классификатор с интервальными предикатами

Другой, более удобный для человеческого восприятия, классификатор может быть построен с использованием интервалов количественного признака x. Чтобы спрогнозировать категорию k целевого признака, этот классификатор ссылается на интервальный предикат $x(a(k), b(k))$, значение которого «истина» тогда и только тогда, когда величина x лежит между $a(k)$ и $b(k)$. Далее строится решающее правило, которое будет справедливо в этой модели, если $x(a(k), b(k)) \Rightarrow k$. Рассмотрим для примера атаку «Saint» в данных о компьютерных атаках: в табл. 1.4 представлено 11 случаев атак такого типа и все, кроме одного, имеют значения признака SHCo, равные 482 или 483. Таким образом, правило интервального предиката SHCo(482, 483) $\Rightarrow$ Saint даст 10 корректных ответов из 11, т. е. лишь 9 % ошибок.

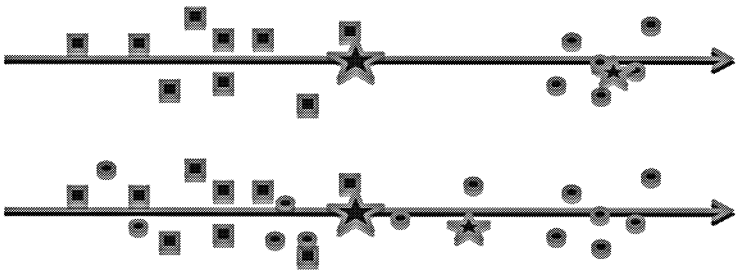


Рис. 3.22. **Группа кружков сильно удалена от остальных объектов на верхней картинке, но на нижней кружки уже перемешаны с остальными объектами; это отражается во взаиморасположении звездочек, представляющих общее и групповое средние**

Как можно понять, какая из категорий наилучшим образом покрывается правилом интервального предиката? Одно из предло-

жений: надо руководствоваться величинами вкладов категорий в разброс x в (3.13), $p_k(\overline{x}-\overline{x}_k)^2$, в обозначениях этого раздела, где p_k — доля объектов в категории k, $\overline{x}$ — общее среднее количественного признака (на всех объектах), а $\overline{x}_k$ — среднее в категории k. Геометрическую иллюстрацию этого предложения можно увидеть на рис. 3.22 (вверху): чем дальше удалено среднее в категории от общего среднего, тем больше шанс, что и все объекты категории удалены, так что ее нетрудно описать интервальным признаком. Однако в каких-то случаях некоторые объекты категории могут находиться далеко от среднего по категории, что приводит к ошибкам «интервального» прогноза (рис. 3.22, нижняя часть).

Рабочий пример 3.9

Вклад категории как основа формирования решающих правил с использованием интервальных предикатов

Рассмотрим уже использовавшиеся признаки Att и SHCo из данных о компьютерных атаках (см. пример 3.8) и определим вклады отдельных категорий Att по формуле (3.13), $p_k(\overline{x}-\overline{x}_k)^2$ (см табл. 3.12).

Таблица 3.12

Вклады категорий по формуле (3.13)

Тип атаки	Доля	Среднее	Вклад
Apache	0.23	33.61	1514.3
Saint	0.11	484.64	15 049.8
Smurf	0.10	508.40	15 496.0
Normal	0.56	5.13	6729.9
Всего	1.00	114.75	38 790.0

С учетом табл. 3.12 попытаемся построить интервальные предикаты для категорий с наибольшими вкладами, Saint и Smurf. Мы уже отмечали, что правило SH(482,483) ⇒ Saint порождает 9%-ную ошибку типа «ложное против». Причина — значение SH = 510, которое соответствует Saint (90-я строка таблицы данных об атаках), но не удовлетворяет правилу. Аналогичный предикат SH(490, 512) ⇒ Smurf также один раз не верен, причем на том же самом объекте — но на этот раз это «ложное да»: удовлетворяет посылке, но не Smurf. Следующая по величине вклада — категория Normal: в ней x меняется от 1 до 28, что частично покрывает интервал (16, 42), относящийся к категории Apache. Если ограничить область изменения до 15, взяв предикат SH(1,15) ⇒ Normal, он оказывается истинным в 53 из 56 случаев; три «ложных против» составляют всего лишь около 5 %. Правило SH(16,42) ⇒ Apache неверно на тех же трех объектах, но теперь это ошибки типа «ложные за».

Вопрос 3.15. Постройте таблицу вкладов категорий, подобную табл. 3.12, для признаков «длина чашелистика» и Таксон по данным об ирисах.

Ответ. См. табл. 3.13.

Таблица 3.13

Вклады таксонов в длину чашелистика.

Таксон	Доля	Среднее	Квадрат разн.	Вклад
T1	1/3	5.006	0.7011	0.2337
T2	1/3	5.936	0.0086	0.0029
T3	1/3	6.588	0.5545	0.1848
Все множество	1.00	5.843	—	—

Относительная успешность импликаций, основанных на интервальных предикатах в Рабочем примере 3.9, обусловлена высокой корреляцией между SH и Attack. В ситуации, когда особой корреляции нет, как, например, для пары «Длина чашелистика — Таксон» на данных об ирисах, интервальные правила могут приводить к многим ошибкам. Рассмотрим, например, категорию с наибольшим вкладом, T1, в табл. 3.13. Диапазон длины чашелистика в таксоне T1 составляет от 4.3 до 5.8. Если взять весь этот интервал и взять правило ДЧ(4.3, 5.8) $\Rightarrow$ T1, оно не дает ни одного случая «ложного против», но вместе с тем приводит к очень многим «ложным за»: 24 объекта таксона T2 и 6 объектов таксона T3 лежат в интервале (4.3, 5.8), что в сумме образует 30 «ложных за»! Можно попытаться сузить границы интервала, чтобы значительно уменьшить число «ложных за» ценой добавления немногих «ложных против». Рассмотрим, например, правило ВИ(4.3, 5.5) $\Rightarrow$ T1: теперь число «ложных за» — 12 (11 T2 и 1 T3), тогда как число «ложных против» — 3, что в итоге даёт 15 ошибок — вдвое меньше, чем 30, но всё-таки немало. Тем не менее, интервальные правила удобны для человека, так что иногда подобные правила могут быть приняты даже несмотря на высокие ошибки. Кроме того, эти правила можно уточнять, добавляя к ним интервальные предикаты, порожденные другими признаками.

3.5. Случай двух номинальных признаков: таблица сопряженности

3.5.1. Таблица сопряженности и концептуальная связь

Для анализа связи между двумя номинальными признаками составляют так называемые таблицы сопряженности. Строки таблицы сопряженности соответствуют категориям одного признака, а столбцы — категориям другого признака. Элемент на пересечении строки и столбца — количество объектов, обладающих соответствующими категориями и того, и другого признаков.

Рабочий пример 3.10

Таблица сопряженности на малых городах

Чтобы создать перекрестную классификацию двух признаков торговых городов, «Банки» (Ба) и «Фермерский рынок» (Фр), необходимо прежде всего категоризовать количественный признак «Банки». Рассмотрим, например, разбиение на 4 категории в табл. 3.14.

Эти категории перекрестно классифицируются с категориями «Есть» и «Нет» признака Фр в таблице сопряженности табл. 3.15. Кроме численностей объектов в категориях перекрестной классификации, в таблице содержатся и суммарные численности категорий — в последних, добавленных, строке и столбце таблицы — вот почему их называют маргинальными. Общее число объектов — в правом нижнем углу таблицы.

Таблица 3.14

Определение категорий Ба по данным торговых городов

Категория	Определение	Обозначение
1	Ба ≥ 10	10+
2	10 > Ба ≥ 4	4+
3	4 > Ба ≥ 2	2+
4	Ба = 0 или 1	1–

Таблица 3.15

Таблица сопряженности признаков Ба и Фр

Категории Фр	Категории Ба				Итого
	10+	4+	2+	1–	
Есть	2	5	1	1	9
Нет	4	7	13	12	36
Итого	6	12	14	13	45

Те же значения сопряженности в относительных частотах (полученные делением на общее количество объектов) представлены в табл. 3.16.

Таблица 3.16

Относительные частоты для таблицы сопряженности Фр/Ба, %

Фр / Банк	10+	4+	2+	1–	Итого
Есть	4.44	11.11	2.22	2.22	20
Нет	8.89	15.56	28.89	26.67	80
Сумма	13.33	26.67	31.11	28.89	100

Самостоятельная работа

3.9. Постройте таблицу сопряженности для признаков «Тип протокола» и «Тип атаки» по данным о компьютерных атаках (см. табл. 1.4), как в абсолютных численностях, так и относительных частотах.

3.10. Разделите ирисы в табл. 1.2 на 4 группы по признаку «Длина чашелистика» и постройте таблицу сопряженности полученного номинального признака с разбиением по таксонам как в абсолютных численностях, так и относительных частотах.

Таблица сопряженности может быть использована для исследования связи между отдельными категориями. Наиболее тесная связь — концептуальная (логическая). Концептуальная связь усматривается тогда, когда в строке k все не маргинальные величины, кроме одной, скажем в столбце l, равны 0. что означает, что если объект имеет категорию k первого признака, он заведомо будет иметь категорию l второго признака. Это означает логическую импликацию, или концептуальную связь $k \Rightarrow l$.

Вопрос 3.16. Постройте таблицу сопряженности для признаков «Тип протокола» и «Тип атаки» для данных о компьютерных атаках.

Ответ. См табл. 3.17.

Таблица 3.17

Таблица сопряженности на данных о компьютерных атаках

Категории	Apache	Saint	Smurf	Normal	Итого
Tcp	23	11	0	30	64
Udp	0	0	0	26	26
Icmp	0	0	10	0	10
Итого	23	11	10	56	100

Рабочий пример 3.11

Эквивалентности и импликации по таблице сопряженности

Рассмотрим таблицу сопряженности признаков «Тип протокола» и «Тип атаки» по данным о компьютерных атаках (табл. 3.17). В строках «Udp» и «Icmp» табл. 3.17 только один ненулевой элемент. Это значит, что таблица содержит логические импликации Udp $\Rightarrow$ Normal и Icmp $\Rightarrow$ Smurf. Более того, в столбце Smurf тоже только один ненулевой элемент! Это значит, что согласно таблице категории Icmp и Smurf эквивалентны, т. е. Icmp $\Leftrightarrow$ Smurf.

Самостоятельная работа

3.11. Рассмотрите таблицу сопряженности признаков «Сектор экономики» и «Использование интернета» по данным в табл. 1.1. Можно ли сделать вывод(ы) о наличии логических импликаций согласно этой таблице?

3.12. Сформируйте из признака «Нас» в табл. 1.5 категоризованный признак «Величина поселения» с категориями «Малая» (до 2400 жителей), «Средняя» (больше 2400, но меньше 8500 жителей) и «Большая» (более 8500 жителей). Рассмотрите таблицу сопряженности этого нового признака и признака Фр («Фермерский рынок»). Можно ли сделать вывод(ы) о наличии логических импликаций согласно этой таблице?

Рис. 3.23. **Нетривиальная импликация**

Задание 3.6. Поправки в таблице сопряженности: лучше не делать.

К сожалению, в табл. 3.15 сопряженности признаков Банки и Фр нулей нет, т. е. нет смысла говорить о концептуальной связи каких-либо категорий этих признаков. Однако, некоторые значения в таблице близки к 0, что подвергает нас соблазну немного урезать данные. Ценой удаления из выборки только двух городов, мы можем добиться того, что в строке «Да» табл. 3.15 два последних значения станут 0, а не 1. Такое преобразование будет означать, что фермерский рынок может появиться только в городе с 4 и более банками. То есть логическое правило «Если Ба ≥ 4, то в городе есть фермерский рынок» справедливо согласно модифицированной таблице сопряженности.

Таблица 3.18

Поправленная перекрестная классификация Ба/Фр (удалено 13 городов).

Фр	Банки				Итого
	10+	4+	2+	1-	
Да	2	5	0	0	7
Нет	0	0	13	12	25
Итого	2	5	13	12	32

Воспользуемся этим приемом для усиления подмеченной корреляции путем очищения таблицы от малых значений. Поправленная таким образом табл. 3.15 преобразуется в табл. 3.18. При этом удалено всего 13 городов из выборки, зато хорошо проявлена концептуальная связь: «В городе есть фермерский рынок тогда и только тогда, когда число банков в нем больше 4»! Но не будем забывать, что цена этой ясности — 13 удаленных городов. Они составляют почти 30 % исходной выборки.

Подобная поправка данных с удалением «нехарактерных» объектов, граничащая с мошенничеством, — одна из причин возникновения популярного парадоксального афоризма, приписываемого Б. Дизраэли, известному британскому политику XIX в.: «Есть три градации лжи: ложь, наглая ложь и статистика.» Здесь мы касаемся проблемы, которая до сих пор не получила в анализе данных сколь-нибудь общего решения. Ясно, что в множестве данных может присутствовать некое, обычно не очень большое, число нехарактерных объектов, так сказать, «выбросов» по отношению к остальным данным, которые следует удалить до того, как анализировать эти данные (см., например, рис. 3.24). Но как их выявить? А если такие данные характеризуют вовсе не выбросы, а наоборот, новые тенденции развития? Безотносительно к этой проблеме мы предпочитаем не урезать данные, а искать другие способы выявления концептуальных связей.

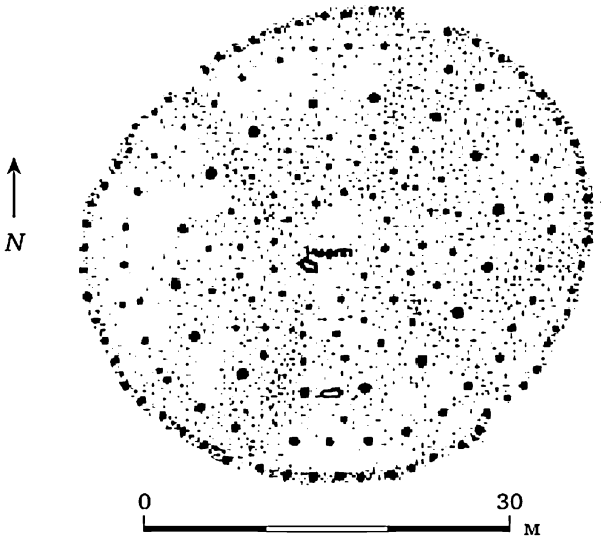


Рис. 3.24. **Нехарактерные объекты:**
схема концентрических кругов, образуемых столбиками Вудхенджа, неолитического памятника в Южной Англии. При этом срезы столбиков, принадлежащих одному и тому же концентрическому кругу, выкрашены в один и тот же цвет; несколько столбиков, находящихся вне кругов, отмечены черным — что они означают и зачем они, неизвестно

3.5.2. Исследование связей с помощью индекса Кетле

Индекс Кетле позволяет визуализировать корреляционные паттерны в таблицах сопряженности без удаления «неподходящих» объектов. Бельгийский ученый Адольф Кетле, один из основоположников современной статистики, еще в 1832 г. предложил измерять степень связи между категориями с помощью сравнения на-

блюденной условной частоты со средней частотой на всей таблице сопряженности. Коэффициент Кетле численно равен отношению этих двух величин минус единица.

Применим понятие коэффициента Кетле к анализу связи между наличием фермерского рынка и категорией «10 и более банков» в табл. 3.15. Частота совместного появления этих категорий — значение на пересечении соответствующих строки и столбца, P(Ба = 10+ & Фр = «Есть») = 2/45 = 4.44 % (частота совместного появления). В целом доля строки «Есть» — 20 %. Значит, частота категории «Ба = 10+» при условии «Фр = "Есть"» равна P(Ба = 10+/Фр = = «Есть») = P(Ба = 10+ & Фр = «Есть»)/P(Фр = «Есть») = 0.0444/0.20 = = 0.222 = 22.2 %.

Это много или мало? Трудно сказать, если не сравнить данное значение с безусловной вероятностью — частотой появления категории «Ба = 10 +» на всем множестве данных, равной P(Ба = 10+) = 13.33 %. Посчитаем теперь относительную разницу между условной и безусловной вероятностями, которую и называем индексом Кетле:

$$q(\text{Ба} = 10+/\text{Фр} = \text{«Есть»}) =$$

$$= [P(\text{Ба} = 10+/\text{Фр} = \text{«Есть»}) - P(\text{Ба} = 10+)]/P(\text{Ба} = 10+) =$$

$$= [0.2222 - 0.1333]/0.1333 = 0.6667 = 66.7\ \%.$$

Это значит, что условие «Фр = "Есть"» повышает частоту категории Ба = 10+ на 66.7 % по сравнению со средней. Такая логика полностью соответствует нашей интуиции. Рассмотрим, например, риск получения серьезного заболевания, скажем туберкулеза, частота которого очень невелика, например, около 0.1 % — одно на тысячу человек в данном регионе. При наличии же какого-либо отягчающего обстоятельства, например, «плохих жилищных условий», уровень туберкулеза будет несколько выше, например, 0.5 %, пять на тысячу человек, что тоже невелико. Но это в 5 раз выше среднего уровня. Вот эти самые «разы» — именно то, что измеряется индексом Кетле: $q(l/k) = (0.5 - 0.1)/0.1 = 400$ %, т. е. средний уровень заболеваемости повышается на 400 % для категории «плохие жилищные условия». Обратим внимание, что индекс Кетле сравнивает условную вероятность категории с безусловной, а не с той, которая получается при противоположном условии; в данном случае — с общей долей заболевших туберкулёзом, а не с долей заболевших среди тех, кто живет в хороших жилищных условиях.

Задание 3.7. Корреляция длины и ширины чашелистика в терминах категорий: Различие между условной вероятностью и индексом Кетле

Вернемся к совместному распределению длины и ширины чашелистика, анализировавшемуся в Задании 3.3. Напомним, что ши-

рина практически не связана с длиной, а если и связана, то отрицательно, так что ширина чашелистика скорее убывает с ростом его длины, чем увеличивается. Это парадоксальное заключение является следствием неоднородности выборки. В каждом таксоне зависимость правильная, возрастающая, но таксоны сильно отличаются как раз соотношением длины и ширины чашелистика.

Посмотрим, как эта связь может проявиться, если оба признака преобразовать к качественному виду. Чтобы провести категоризацию осмысленно, посмотрим на реально наблюденные распределения этих признаков. Гистограммы на Рис. 3.20 позволяют осмысленно выбрать границы категорий. Как видно, на гистограмме длины чашелистика точки минимума расположены примерно на значениях 5, 6 и 7. Это значит, что вектор $x = (4, 5, 6, 7, 8)$ может использоваться как совокупность разделителей между категориями длины, так что 4 — начало самой меньшей категории, а 8 — конец самой большой из них. Аналогично выбирается вектор $y = (2, 2.8, 3.5, 4.5)$ разделителей для признака ширины. Формирование множеств объектов, соответствующих выделенным категориям, признаков длины и ширины (точнее их индексов) осуществляется в циклах по категориям:

```
>> for k = 1:4;le{k} = find(ir(:,1)<x(k+1) & ir(:,1)> = x(k));
end
>> for l = 1:3;wi{l} = find(ir(:,2)<y(l+1) & ir(:,2)> = y(l));
end
```

Здесь:

— ir — обозначение для 150 × 4 матрицы данных об ирисах;

— find(условие) — оператор MATLAB, отбирающий индексы всех тех и только тех объектов, для которых выполнено условие в скобках;

— for … end — символика для определения цикла, автоматизирующего заполнение множеств le{k} и wi{l} объектов, попадающих в одну и ту же категорию k или l.

Таблица сопряженности состоит из чисел p(l, k), выражающих численность объектов в пересечении категорий l и k, т. е. $|wi\{l\} \cap le\{k\}|$. Получить ее можно с помощью следующей последовательности команд:

```
>> for k = 1:3;for l = 1:4;p(k,l) = ...
     length(intersect(wi{k},le{l}));end;end;
```

Здесь intersect выражает операцию взятия пересечения (общей части) множеств, а length определяет ее длину, т. е. численность. В результате получим матрицу

$$P = \begin{matrix} 3 & 18 & 11 & 1 \\ 17 & 23 & 43 & 9. \\ 2 & 20 & 0 & 3 \end{matrix}$$

В матрице P связи между созданными категориями длины и ширины чашелистика выражены в агрегированном виде. Например, 17 выражает общее число объектов, попавших в первую категорию длины и вторую категорию ширины.

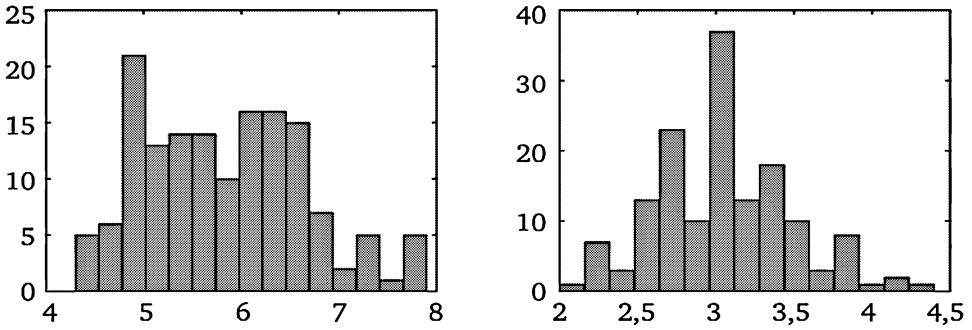


Рис. 3.25. **15-бинные гистограммы признаков чашелистика на ирисах:** длина слева, ширина справа

Парадоксальных объектов немного: первая категория длины (самые короткие) содержит только 2 объекта третьей категории (самые широкие), а третья категория длины (самые длинные) содержит только 1 объект первой категории (самые узкие). Впрочем, мало и самых коротких чашелистиков, являющихся самыми узкими (вопрос читателю: сколько?), равно как и самых длинных, являющихся самыми широкими (вопрос читателю: сколько?).

Кроме того, можно получить суммарные (маргинальные) частоты категорий по столбцам

$$\mathrm{Pc} = 22\ 61\ 54\ 13$$

и строкам

$$\mathrm{Pr} = \begin{matrix} 33 \\ 92. \\ 25 \end{matrix}$$

Для контроля вычислений можно суммировать Pr и Pc: суммы должны равняться общему числу объектов, 150. Разделив P на это число, получим матрицу относительных частот:

$$\mathrm{Pf} = \begin{matrix} 0{,}0200 & 0{,}1200 & 0{,}0733 & 0{,}0067 \\ \mathbf{0{,}1133} & \mathbf{0{,}1533} & \mathbf{0{,}2867} & \mathbf{0{,}0600}, \\ 0{,}0133 & 0{,}0133 & 0 & 0{,}0200 \end{matrix}$$

которая суммируется к 1. На самом деле эти частоты играют роль вероятностей совместной встречаемости категорий длины и шири-

ны. Мы видим, что чаще всего встречается комбинация третьей категории длины и второй категории ширины, т. е. довольно длинные и широкие чашелистики, 28.67 %.

Зададимся теперь вопросом: при заданной категории длины, какова наиболее вероятная категория ширины? Для ответа достаточно найти максимум в каждом столбце; ведь именно столбцы соответствуют категориям длины. В матрице Pf все максимумы сосредоточены во второй строке, что показано выделением этих элементов жирным шрифтом. Это означает, что знание столбца не дает никакой полезной информации — та же вторая категория ширины отвечает любой категории длины. Попробуем слегка уточнить вопрос: какова условная вероятность строк (категорий ширины) при условии, что категория длины фиксирована.

Дадим ответ на этот вопрос для ситуации, когда фиксирована первая категория длины. В этом случае частоты встречаемости категорий ширины — это элементы первого столбца матрицы P:

$$3$$
$$17$$
$$2$$

В этой категории имеется всего 22 объекта (другие объекты принадлежат другим категориям длины), так что искомые вероятности получаются делением частот столбца на это число:

$$3/22 = 0.136;$$
$$17/22 = 0.773;$$
$$2/22 = 0.091.$$

Обратим внимание на то, что полученные величины суммируются к 1, т. е. действительно соответствуют вероятностям категорий ширины при условии, что имеет место первая категория длины. Операцию деления элемента матрицы P на сумму элементов его столбца можно провести в MATLAB следующим образом:

```
>> Pcon = P./repmat(pr,1,4)
```

Здесь операция repmat(pr,m,n) создает новую матрицу путем копирования матрицы pr, *m* раз по вертикали, и *n* раз — по горизонтали.

В результате получаем:

$$\text{Pcon} = \begin{matrix} 0{,}1364 & 0{,}2951 & 0{,}2037 & 0{,}0769 \\ 0{,}7727 & 0{,}3770 & 0{,}7963 & 0{,}6923. \\ 0{,}0909 & 0{,}3279 & 0 & 0{,}2308 \end{matrix}$$

Таким образом, условная вероятность второй категории ширины равна 70 % или даже больше во всех категориях длины, кроме второй. Для второй категории длины условная вероятность второй категории ширины «всего» 37.7 %.

Условная вероятность часто применяется как способ разобраться в структуре связей между категориями столбцов и категориями строк. Но ее разрешающая способность ограничена. В ситуациях типа наблюдаемой сейчас ее использование не дает никакой новой информации по сравнению с безусловными вероятностями. Применение коэффициентов Кетле, измеряющих приростные характеристики, может оказаться значительно более информативно. По определению коэффициент Кетле равен относительной разнице между условной и безусловной вероятностями события: q = (p(строка/столбец) – p(строка))/p(строка). На языке матриц в MATLAB это может быть выражено следующим образом, в процентах:

```
>>qc = 100*(Pcon-repmat(Pr,1,4)/n)./ repmat(Pr,1,4)/n
```

что дает

$$qc = \begin{matrix} -38.02 & 34.13 & -7.41 & -65.03 \\ 25.99 & -38.52 & 29.83 & 12.88. \\ -45.45 & 96.72 & -100.00 & 38.46 \end{matrix}$$

Имеется и другая, эквивалентная, формула для вычисления индексов Кетле (см. раздел Ф. 4.3):

```
>> qq = P*n./(Pr*Pc) –1
```

Эта операция дает ту же матрицу, в процентах,

$$qq = \begin{matrix} -38.02 & 34.13 & \mathbf{-7.41} & -65.03 \\ \mathbf{25.99} & -38.52 & \mathbf{29.83} & 12.88 \\ \mathbf{-45.45} & \mathbf{96.72} & -100.00 & \mathbf{38.46} \end{matrix}.$$

Коэффициент Кетле показывает, на сколько процентов условная вероятность строки отличается от ее безусловной вероятности; чем он больше, тем выше связь. На полученной матрице максимумы столбцов — они выделены жирным шрифтом — попадают теперь не только во вторую строку, но и в третью. Конкретно, при условии, что длина выражается второй категорией, ширина попадает в категорию самых широких листьев на 96.72 % раз чаще, чем в среднем.

Рабочий пример 3.12

Индекс Кетле и таблица сопряженности

Применим приведенную выше формулу для вычисления индексов Кетле для данных в табл. 3.15 и запишем результаты в табл. 3.19.

Таблица 3.19

Коэффициенты Кетле для пары Ба/Фр, в процентах*

ФРынок	10+	4+	2+	1–
Есть	**66.67**	108.33	–64.29	–61.54
Нет	–16.67	–27.08	**16.07**	15.38

* Жирным шрифтом выделены положительные значения.

Выделяя положительные величины в таблице, получим тот же паттерн, что и в «очищенных» данных, приведенных в Проекте 3.3.

Но на этот раз мы сохранили выборку, как она есть, не пытаясь ее очистить. Выводы те же. В частности, можно отметить, что категория «Есть» признака Фр обеспечивает значительное увеличение вероятности наличия многих банков, в то время, как категория «Нет» ведет к гораздо более слабым изменениям.

Вопрос 3.17. Рассчитайте коэффициенты Кетле для табл. 3.17.

Ответ. См. табл. 3.20, где положительные значения выделены жирным шрифтом.

Таблица 3.20

Индексы Кетле для пары «Тип протокола»/«Тип атаки» из таблицы сопряженности (табл. 3.17), %

Категория	Apache	Saint	Surf	Norm
Tcp	**56.25**	**56.25**	–100.00	–16.29
Udp	–100.00	–100.00	–100.00	**78.57**
Icmp	–100.00	–100.00	**900.00**	–100.00

Задание 3.8. Правило «Останови и Обыщи» в Англии: расовые предрассудки полиции?

Рассмотрим пример реальных данных. Статистика применения полицией правила: «Не нужно никаких дополнительных разрешений, чтобы остановить и обыскать любого подозрительного индивида» (Останови-и-Обыщи, ОО) в Англии и Уэльсе в 2005 году, представлена в разрезе цвета кожи индивида (Ч — черный, А — азиат и Б — белый), в табл. 2.5 параграфа 2.8 — с подавляющим большинством индивидов, принадлежащих категории Б.

Сравнение распределения ОО с распределением всего населения по цвету кожи порождает критические замечания прессы о существовании в полиции расовых предрассудков. Попытаемся разобраться, в чем тут дело. Распределение населения по цвету кожи нетрудно найти в Интернете. Вычитая из общей численности населения той или иной категории цвета кожи количество случаев ОО для данной категории, получим табл. 3.21 (по итогам переписи 2001 года).

Таблица 3.21

Распределение случаев ОО по цвету кожи

Категория	ОО	Не ОО	Всего	Доля ОО
Ч	131 723	1 377 493	1 509 216	0.0873
А	70 252	2 948 179	3 018 431	0.0233
Б	676 178	4 683 8091	47 514 269	0.0142
Всего	878 153	51 163 763	52 041 916	0.0169

При этом неявно используется гипотеза, что никто не подвергался процедуре ОО больше одного раза. В последнем столбце указаны относительные величины. Они-то и послужили поводом для обвинений полиции в расовых предрассудках: действительно, риск подвергнуться ОО процедуре в шесть раз чаще для представителя категории Ч, чем для представителя категории Б. Коэффициенты Кетле показывают то же (см табл. 3.22). Категория Ч подвергалась ОО процедуре на 400 % больше раз, чем в среднем; а категория Б — на 15 % меньше.

Таблица 3.22

Коэффициенты Кетле для перекрестной классификации из табл. 3.21, %

Категория	ОО	Не ОО
Черный	417.2	–7.2
Азиат	37.9	–0.6
Белый	–15.7	0.3

Многие, включая автора, рассматривают подобные выводы из табл. 3.21, 3.22 как неправильные: они основаны на неявном постулате, что процедура ОО применяется к населению случайным образом. Разумно предположить, что полицейские — не идиоты, и применяют ОО только при необходимости. В таком случае распределение случаев ОО должно сравниваться с распределением цвета кожи не у всего населения, а тех, которые приговорены к тюремному заключению. Автор сделал такое сравнение: распределение множества лиц, подвергшихся ОО, по цвету кожи оказалось практически идентичным распределению лиц, находящихся в заключении. Поэтому утверждение о расовых предрассудках полицейских, столь очевидное на первый взгляд, должно быть объявлено неверным (конечно, при условии, что английская судебная система в целом свободна от них).

3.5.3. Коэффициент хи-квадрат как индекс связи и визуализация его структуры

Относительно удачная визуализация таблицы сопряженности получается с помощью индексов Кетле, взвешенных вероятностями

соответствующих ячеек таблицы, как объясняется в разделе Ф3.5.4. Сумма этих величин приводит к одному из наиболее популярных понятий анализа таблиц сопряженности, так называемому коэффициенту сопряженности хи-квадрат. Этот коэффициент был предложен К. Пирсоном (1901) как мера отклонения наблюдаемого двумерного распределения в таблице сопряженности от условия статистической независимости признаков.

Два признака считаются статистически независимыми, если все возможные пары их категорий статистически независимы, т. е. вероятность/частота их совместного появления равна произведению вероятностей этих категорий по отдельности. К. Пирсон рассматривал ситуацию, когда два признака статистически независимы в популяции, но в рассматриваемой конкретной выборке независимость не выполняется из-за случайных отклонений выборки. Он предложил взять квадрат разности наблюденной частоты и величины, полученной при выполнении предположения независимости, и разделить его на «теоретическую» вероятность, истинную для популяции. Суммарный индекс носит название коэффициент Пирсона хи-квадрат, см. формулу (3.22). Распределение суммарного коэффициента хи-квадрат, умноженного на численность выборки, при условии справедливости гипотезы независимости в популяции, сходится к так называемому распределению хи-квадрат, которое используется в статистике для проверки гипотез. К сожалению, до последнего времени никакого операционального истолкования коэффициента хи-квадрат предложено не было. Поэтому часто утверждают, что коэффициент может быть использован только для тестирования гипотезы независимости, но не как мера коррелированности. Согласно этому мнению, коэффициент хи-квадрат должен использоваться для различения только двух ситуаций: наличие статистической независимости и ее отсутствие, так что числовое значение коэффициента само по себе не может использоваться как характеристика степени связи. Впрочем, практики не всегда следуют этому мнению и все же используют хи-квадрат как меру связи между категоризованными признаками. Как свидетельствует формула (3.24), в этом нет ничего плохого или некорректного. Коэффициент сопряженности хи-квадрат по своей сути не что иное как усредненный индекс Кетле, и, значит, характеризует связь между категориями двух признаков. Какую конкретно? Вот какую: коэффициент хи-квадрат характеризует среднее повышение вероятности категории одного признака в ситуации, когда становится известной категория второго признака.

Рабочий пример 3.13

Визуализация таблицы сопряженности с использованием взвешенного индекса Кетле

Умножим коэффициенты Кетле из табл. 3.19 на частоты значений в соответствующих ячейках табл. 3.15. При этом следует выражать коэф-

фициенты Кетле в табл. 3.19 в абсолютных величинах, а не в процентах. В результате получаем табл. 3.23, элементы которой суммируются к 6.86 — коэффициенту хи-квадрат Пирсона для табл. 3.14. Заметим, что значения в табл. 3.23 могут быть как положительными, так и отрицательными; те, чье значение по модулю больше удвоенного среднего, 2*6.86/8 = 1.72, выделены жирным шрифтом — они сильно отклоняются от среднего. При этом столбец 4 + содержит как наибольший положительный, так и наибольший отрицательный вклады.

Таблица 3.23

Хи-квадрат для пары «Банк/Фермерский рынок» и его разложение по уравнению (3.19)

ФР	10+	4+	2+	1–	Итого
Есть	1.33	**5.41**	–0.64	–0.62	5.48
Нет	–0.67	**–1.90**	**2.09**	**1.85**	1.37
Итого	0.67	3.51	1.45	1.23	**6.86**

Пара категорий считается статистически независимой, если вероятность/доля совместного появления двух категорий равна произведению вероятностей этих категорий. Например, рассмотрим категорию «Есть» для Фермерского рынка и «4+» для числа банков Ба в табл. 3.16: вероятность их совместного появления равна 0.111. С другой стороны, вероятность того, что Фр = «Есть», равна 0.2, а вероятность того, что Ба = 4+, равна 0.267. Если бы две эти категории были независимы, то вместе их можно было бы наблюдать с частотой 0.2 · 0.267 = 0.053, примерно в 2 раза меньшей, чем в действительности, что говорит о том, что для этой пары говорить о статистической независимости не приходится.

Рабочий пример 3.14

Традиционное разложение коэффициента хи-квадрат

Рассмотрим общепринятый способ визуализации таблиц сопряженности, состоящий в том, что в ячейки таблицы сопряженности вписываются величины, которые удобно называть парными коэффициентами Пирсона. Коэффициент Пирсона для ячейки (k, l) — это квадратный корень $r(k, l)$ из величины, суммируемой в классической формуле (3.22) коэффициента хи-квадрат. Для рассматриваемых данных таблица коэффициентов Пирсона — в табл. 3.24.

Таблица 3.24

Коэффициенты Пирсона (квадратные корни из величин, суммируемых в традиционной формуле хи-квадрат Пирсона); сами значения указаны в скобках

ФР	10+	4+	2+	1–	Сумма
Есть	**0.73** (0.53)	**1.68 (2.82)**	–1.08 (1.16)	–0.99 (0.98)	(5.49)
Нет	–0.36 (0.13)	–0.84 (0.70)	**0.54 (0.29)**	**0.50 (0.25)**	(1.37)
Сумма	(0.67)	(3.52)	(1.45)	(1.23)	(6.86)

Эта таблица всегда показывает тот же паттерн отрицательных и положительных величин, что и разложение Кетле. Однако здесь коэффициент хи-квадрат получается суммированием не самих элементов таблицы, а их квадратов. Тот факт, что суммарные значения в маргинальных полях табл. 3.23 и 3.24 одинаковы, — не случайность: он объясняется математическим свойством, выраженным в уравнении (3.23).

Вопрос 3.18. В табл. 3.23 все маргинальные значения, суммы строк и столбцов, положительны, даже несмотря на то, что многие из элементов таблицы — отрицательные. Является ли это лишь особенностью этой таблицы или же проявлением общего свойства?

О: Проявление общего свойства: суммы элементов N_{lk} $q(l/k)$ в строке или в столбце должна быть положительны, см (3.23) далее.

Вопрос 3.19. Постройте аналогичное разложение коэффициента хи-квадрат для пары Таксон/Длина лепестка по данным об ирисах.

Подсказка: прежде всего категоризуйте количественный признак «Длина лепестка»; для этого можно использовать бины одинакового размера или любой другой разумный способ.

Вопрос 3.20. Можно ли составить какое-либо логическое правило вывода, основываясь на данных в столбцах табл. 3.17?

Ответ. Да, обе атаки, и Apache, и Saint, могут возникнуть лишь в протоколе tcp.

Вопрос 3.21. Рассмотрим следующую информацию, дополнительную к условиям Вопроса 2.14. Среди покупателей в этом вопросе каждый, кто тратит на покупки £60, это обязательно мужчина; каждый, кто тратит £100, всегда женщина; а среди оставшихся 30 человек половина женщин и половина мужчин. Постройте таблицу сопряженности двух признаков: пол и расходы на покупки. Найдите и объясните величину коэффициента Кетле для пары категорий «женщины, которые тратят по £100 каждая».

Ответ. Таблица сопряженности (численности совместного появления событий):

Пол	Расходы, £			
	60	100	150	Итого
Жен.	0	20	15	35
Муж.	50	0	15	65
Итого	50	20	30	100

В данной таблице частоты совместного появления событий совпадают со своими процентными значениями, так как число покупателей равно 100.

Рассчитаем коэффициент Кетле q(Жен/£100) по формуле (3.20′):

$$q = 100 \cdot 20/(20*35) - 1 = 2.86 - 1 = 1.86.$$

Это означает, что вероятность того, что индивид в данной категории расходов окажется женщиной, больше средней частоты женщин на 186 %.

Ф3.5.4. Анализ таблиц сопряженности: формулировки

Рассмотрим два непересекающихся множества номинальных категорий на множестве объектов I: $l = 1, \ldots, L$ (например, профессиональная принадлежность индивидов, составляющих I) и $k = 1, \ldots, K$ (скажем, тип семьи или домашнего хозяйства у этих же индивидов). Каждое множество категорий задает разбиение множества I. Рассмотрим пересечение этих разбиений, чтобы агрегировать данные и проанализировать связь между двумя множествами категорий. Для пары категорий $(k, l) \in K \cdot L$ посчитаем количество таких объектов в множестве I, которые попадают в обе категории одновременно. Обозначим через N_{kl} количество совместного появления пары (k, l). Очевидно, в сумме величины N_{kl} дадут N, общее число объектов в I, поскольку категории одного и того же множества (а) не пересекаются и (б) покрывают все множество I. Таблица, в которой записаны все N_{kl}, или относительные величины — *частоты* $p_{kl} = N_{kl}/N$, называется *таблицей сопряженности* или просто перекрестной классификацией. Суммарные значения, сумма по строке $N_{k+} = \Sigma_l N_{kl}$ и сумма по столбцу $N_{+l} = \Sigma_k N_{kl}$ (так же как и их относительные значения с учетом числа строк и столбцов, соответственно), называются *маргинальными* (поскольку находятся в крайнем столбце и крайней строке, т. е. на «полях» таблицы сопряженности).

Вероятность (эмпирическая) того, что категория l появится при наличии категории k, выражается как условная частота $P(l/k) = p_{kl}/p_{k+} = N_{kl}/N_{k+}$, т. е. частота категории l на подмножестве объектов, соответствующих категории k. Вероятность $P(l)$ категории l на всем множестве I есть $p_{+l} = N_{+l}/N$. Аналогичное обозначение используется для категорий k. Относительная разница между условной и безусловной вероятностями называется (относительным) индексом Кетле [43]:

$$q(l/k) = \frac{P(l/k) - P(l)}{P(l)}, \tag{3.20}$$

где $P(l) = N_{+l}/N$, $P(k) = N_{k+}/N$, $P(l/k) = N_{kl}/N_{k+}$. То есть индекс Кетле выражает связь между категориями k и l как относительное изменение вероятности появления l при условии k.

Используя простые алгебраические преобразования, можно получить более простое выражение:

$$q(l/k) = [N_{kl}/N_{k+} - N_{+l}/N]/(N_{+l}/N) =$$

$$= N_{kl}N/(N_{k+}N_{+l}) - 1 = \frac{p_{kl}}{p_{k+}p_{+l}} - 1 = q(k,l). \qquad (3.20')$$

Последнее обозначение, $q(k, l)$, подчеркивает тот факт, очевидный из выведенной формулы (3.20′), что коэффициент Кетле симметричен относительно индексов k и l.

Выделение наибольших положительных и отрицательных значений индекса Кетле визуализирует структуру связи между двумя множествами категорий, что проиллюстрировано табл. 3.19 и 3.22.

Это визуализированное представление может быть включено в традиционный статистический контекст. Определим интегральный индекс связи Кетле Q как сумму парных индексов Кетле, взвешенных их частотами/вероятностями:

$$Q = \sum_{k=1}^{K}\sum_{l=1}^{L} p_{kl}q(l,k) = \sum_{k=1}^{K}\sum_{l=1}^{L} p_{kl}\left(\frac{p_{kl}}{p_{k+}p_{+l}} - 1\right) =$$

$$= \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{p_{kl}^2}{p_{k+}p_{+l}} - 1. \qquad (3.21)$$

Самое правое выражение в (3.21) не является чем-то необычным; напротив, оно довольно часто встречается в статистическом анализе таблиц сопряженности. Это не что иное, как альтернативная формула для коэффициента сопряженности хи-квадрат Пирсона (1901). Коэффициент хи-квадрат был введен, и с тех пор используется, в совершенно другом контексте — в качестве меры отклонения таблицы сопряженности от статистической независимости.

Для объяснения сказанного сформулируем математическое определение понятия статистической независимости. Множества категорий k и l *статистически независимы*, если $p_{kl} = p_{k+}p_{+l}$ для всех k и l. Выполнение условия независимости в реальности маловероятно. К. Пирсон предложил использовать относительные квадратичные ошибки для того, чтобы оценить отклонение наблюдаемых частот от статистической независимости. А именно, он ввел следующий коэффициент, который и называется коэффициентом сопряжённости хи-квадрат Пирсона[1]:

$$X^2 = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}p_{+l}} = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{p_{kl}{}^2}{p_{k+}p_{+l}} - 1. \qquad (3.22)$$

[1] Традиционно под коэффициентом Пирсона понимают величину NX^2; наше обозначение позволяет избежать зависимости величины коэффициента от числа объектов N.

Уравнение справа может быть доказано с использованием элементарной алгебры. Рассмотрим, например, внутреннюю сумму из левой части выражения (3.22):

$$\sum_{l=1}^{L}\frac{(p_{kl}-p_{k+}p_{+l})^2}{p_{k+}p_{+l}}=\sum_{l=1}^{L}\frac{p_{kl}^2-2p_{kl}p_{k+}p_{+l}+(p_{k+}p_{+l})^2}{p_{k+}p_{+l}}=$$

$$=\sum_{l=1}^{L}\frac{p_{kl}^2}{p_{k+}p_{+l}}-2\sum_{l=1}^{L}p_{kl}+\sum_{l=1}^{L}p_{k+}p_{+l}=\sum_{l=1}^{L}\frac{p_{kl}^2}{p_{k+}p_{+l}}-p_{k+}.$$

Выражение в правой части получено с использованием уравнений $\Sigma_l p_{kl} = p_{k+}$ и $\Sigma_l p_{+l} = 1$. Просуммировав эти выражения по k, получим формулу (3.22). С другой стороны, последнее выведенное выше выражение, очевидно, равно $\Sigma_l p_{kl} q(l/k)$, так что

$$\sum_{l=1}^{L}\frac{(p_{kl}-p_{k+}p_{+l})^2}{p_{k+}p_{+l}}=\sum_{l=1}^{L}p_{kl}q(l/k). \qquad (3.23)$$

Сравнивая правые части (3.21) и (3.22), нетрудно заметить, что $X^2 = Q$. То же самое получается, если просуммировать уравнения (3.23) по всем k.

Популярность величины NX^2 в статистике и смежных науках опирается на теорему, доказанную К. Пирсоном. Если таблица сопряженности построена по случайной и независимой выборке объектов из популяции, в которой выполняется условие статистической независимости (так что все отклонения обусловлены лишь случайностью выборки), то вероятностное распределение величины NX^2 сходится к распределению хи-квадрат с числом степеней свободы, равным $(K-1)(L-1)$ (при стремлении N к бесконечности). Вероятностное распределение хи-квадрат с m степенями свободы определяется как распределение суммы квадратов m случайных величин, распределенных по стандартному нормальному закону (с нулевым математическим ожиданием и единичной дисперсией). Это означает, что величина NX^2 может использоваться для проверки гипотезы о статистической независимости.

Теорема Пирсона не всегда применима в анализе данных, поскольку данные могут быть не случайными, а наблюдения не обязательно независимыми. Вместе с тем коэффициент хи-квадрат Пирсона на практике иногда используется не столько для исследования независимости, сколько для оценки связи в таблицах сопряженности. Эта побочная и, в свете теоремы Пирсона, некорректная цель выглядит совершенно оправданной и корректной в свете равенства $X^2 = Q$. Данное равенство вообще придает коэффициенту X^2 другую интерпретацию — это не просто мера отклонения от независимости. Это мера взаимосвязи между категориями — усредненный коэффициент Кетле, т. е. среднее увеличение вероятности значения одного признака при условии, что стало известно значение второго признака.

Для уточнения смысла X^2 как коэффициента корреляции рассмотрим экстремальные значения X^2, и ситуации, в которых эти значения достигаются [43]. Оказывается, что при $K \leq L$, т. е. когда число строк не превышает число столбцов, X^2 изменяется в пределах от 0 до $K - 1$. X^2 равен 0, если все пары (k, l) статистически независимы, так что все $q_{kl} = 0$. С другой стороны, X^2 равен максимальному значению $K - 1$, если каждый столбец l содержит единственный ненулевой элемент — строку этого элемента обозначим $k(l)$, так что сам элемент будет $p_{k(l)l}$, который при этом, естественно, равен p_{+l}. В этом случае, очевидно, имеет место логическая импликация $k(l) \Rightarrow l$. Таким образом, X^2 действительно измеряет связь. Его максимальное значение достигается тогда и только тогда, когда имеет место логическая связь между категориями двух множеств.

Разложение коэффициента хи-квадрат через коэффициенты Кетле

$$X^2 = \sum_{k=1}^{K} \sum_{l=1}^{L} p_{kl} q(l/k). \tag{3.24}$$

позволяет представить X^2 как сумму произведений $p_{kl}q(l/k)$, называемых далее взвешенными индексами Кетле, и разместить эти произведения в соответствующих клетках таблицы сопряженности, как это сделано в табл. 3.22, где эти величины еще домножены на N, чтобы соответствовать величине NX^2 из теоремы Пирсона.

На самом деле не только общая сумма всех элементов совпадает с суммой хи-квадрат величин $(p_{kl} - p_{k+}p_{+l})^2/p_{k+}p_{+l}$, но и суммы по строкам и по столбцам также совпадают, что ясно следует из (3.23).

Тем не менее изначально все хи-квадрат величины в (3.22) положительны. Поэтому иногда используются квадратные корни из этих величин, отражающие знак связи,

$$r(k,l) = \frac{p_{kl} - p_{k+}p_{+l}}{\sqrt{p_{k+}p_{+l}}}, \tag{3.25}$$

которые принято называть индексами Пирсона. Очевидно, $X^2 = \Sigma_{k,l}\, r(k,l)^2$. Индексы Пирсона имеют те же знаки, что и $q(l/k)$, и тесно с ними связаны: $q(l/k) = r(k,l)\,[(p_{k+}p_{+l})]^{1/2}$.

Рабочий пример 3.15

Анализ связей между доходом и политическими предпочтениями

Табл. 3.25 из [44] представляет предпочтения граждан США при выборе президента по данным, собранным Исследовательским Центром Пью в 2014 году.

Они разделили все домовладения на 4 группы годового дохода: (1) менее \$30 000; (2) больше, чем \$30 000, но меньше, чем \$50 000; (3) больше \$50 000, но меньше \$100 000; (4) \$100 000 или больше. Предпочтения при выборах президента определяются как Р (предпочитающие Республ-

ликанскую партию), Д (предпочитающие Демократическую партию) или Н (Не определившиеся).

Таблица 3.25

Перекрестная классификация респондентов в США по признаку дохода (4 группы) и политическим предпочтениям на выборах президента*

Доход	Партия			Всего	Партия		
	Р	Н	Д		Р	Н	D
	Количество респондентов						
(1)	2388	2034	4423	8845	**–0.3034**	**0.4629**	0.0985
(2)	2286	938	2696	5920	–0.0037	0.0079	0.0004
(3)	3885	1126	3712	8723	**0.1491**	**–0.1788**	–0.0652
(4)	3258	695	3049	7002	**0.2005**	**–0.3686**	–0.0435

* Слева — таблица сопряженности, справа — значения индекса Кетле (жирным выделены значения, отделенные от 0 на 15 % или более).

Данные табл. 3.25 относятся к $N = 30\,490$ респондентов.

Таблица 3.26

Взвешенные индексы Кетле для данных табл. 3.25

Доход	Партия		
	Р	Н	Д
1	–0.0238	0.0309	0.0143
2	–0.0003	0.0002	0.0000
3	0.0190	–0.0066	–0.0079
4	0.0214	–0.0084	–0.0043

Взвешенные индексы Кетле из (3.24) представлены в табл. 3.26. Сумма положительных элементов этой таблицы равна 0.0859, а сумма отрицательных элементов –0.0513. Сумма этих двух дает $Q = X^2 = 0.0345$, относительно маленькое значение, едва превышающее 1 % от максимума X^2, равного в данном случае 3. Однако гипотеза независимости здесь должна быть отвергнута на уровне доверия 99.9 % из-за относительно большого значения NX^2, равного $30490 \cdot 0.0345 = 1051.9$.

Очень небольшое значение $Q = X^2 = 0.0345$ не позволяет использовать хи-квадрат для вынесения каких-либо суждений о структуре связей в табл. 3.25. Вместе с тем значения коэффициентов Кетле позволяют сделать следующие, весьма определенные, выводы:

— уровень дохода практически не сказывается на отношении населения к демократам (близкие к 0 значения в столбце Д);

— уровень дохода существенно сказывается на отношении к республиканцам (доля бедных среди голосующих за них на 30 % меньше, чем в среднем; напротив, доля богатых на 20 % больше);

— уровень дохода существенно сказывается на политической активности (доля не определившихся среди бедных на 46 % выше, чем в среднем, тогда как среди богатых — на 37 % меньше).

Вопрос 3.22. Каков смысл суммы всех положительных и суммы всех отрицательных величин в табл. 3.26?

Самостоятельная работа

3.13. В табл. 3.27 приведено распределение ответов 140 респондентов на следующие два вопроса [44]:

B1 — «Ваше отношение к снотворным таблеткам» с возможными ответами: «сильно за», «за», «нейтральное», «против», «сильно против»;

B2 — «Вы хорошо спите по ночам?» с возможными ответами:

1) никогда,
2) редко,
3) иногда,
4) часто,
5) всегда.

Таблица 3.27

Данные о распределении ответов 140 человек на вопросы B1 и B2

B2	B1					Итого
	Никогда	Редко	Иногда	Часто	Всегда	
Сильно за	15	8	3	2	0	28
За	5	17	4	0	2	28
Нейтрально	6	13	4	3	2	28
Против	0	7	7	5	9	28
Сильно против	1	2	6	3	16	28
Итого	27	47	24	13	29	140

Найдите индексы Кетле и взвешенные индексы Кетле для этой таблицы. Сделайте выводы (см. [44]).

Вопрос 3.23. Рассмотрим два бинарных признака и построим для них таблицу сопряженности (ее часто называют четырехклеточной таблицей, табл. 3.28), где символы a, b, c, d используются для обозначения частот совместного появления.

Таблица 3.28

Таблица сопряженности двух бинарных признаков.

		Признак Y		Итого
		Да	Нет	
Признак X	Да	a	b	$a+b$
	Нет	c	d	$c+d$
Итого		$a+c$	$b+d$	$N=a+b+c+d$

Докажите, что коэффициент Кетле q(Да/Да), характеризующий относительную разницу между $a/(a+c)$ и $(a+b)/N$, равен

$$q(\text{Да}/\text{Да}) = \frac{ad-bc}{(a+c)(a+b)},$$

а суммарный коэффициент Кетле Q, или X^2 Пирсона, равен

$$Q = \frac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}.$$

Вопрос 3.24. Докажите, что коэффициент корреляции двух бинарных 1/0 признаков может быть выражен в терминах четырехклеточной табл. 3.27 как $\rho = \sqrt{Q}$, т. е.,

$$\rho = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}.$$

— корень квадратный из Q в предыдущем вопросе. Это означает, что величина Q одновременно имеет смысл коэффициента детерминации в линейной регрессии бинарных признаков.

Вопрос 3.25. Рассмотрим пару категорий $k \in K$ и $l \in L$ согласно $K \times L$ таблице сопряженности P. Определим абсолютный индекс Кетле $a(l/k) = P(l/k) - P(l)$ — изменение частоты $l \in L$ на всем множестве объектов I при условии, что речь идет об объектах категории k. В соответствии с P, $P(l) = p_{+l}$ и $P(l/k) = p_{kl}/p_{+l}$. Докажите, что интегральный абсолютный индекс Кетле $A = \Sigma_{k,l} p_{kl} a(l/k) = \Sigma_{k,l} p_{kl}^2 / p_{k+} - \Sigma_l p_{+l}^2$ равен следующему выражению, асимметричному аналогу коэффициента хи-квадрат Пирсона:

$$A = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}}. \tag{3.26}$$

Величина A (3.26) является числителем известного асимметричного индекса, так называемого «тау-б» Гудмана-Крускала (см. [29]).

О. В самом деле, возведя в квадрат знаменатель, преобразуем выражение (3.26) в равносильное ему выражение

$$\Sigma_{k,l}(p_{kl}^2 - 2p_{kl}p_{k+}p_{+l} + p_{k+}^2 p_{+l}^2) / p_{k+},$$

равное

$$\Sigma_{k,l} p_{kl}^2 / p_{k+} - 2\Sigma_{k,l} p_{kl} p_{+l} + \Sigma_{k,l} p_{k+} p_{+l}^2 = \Sigma_{k,l} p_{kl}^2 / p_{k+} - 2\Sigma_{k,l} p_{+l}^2 + \Sigma_l p_{+l}^2,$$

поскольку $\Sigma_k p_{kl} = p_{+l}$ и $\Sigma_k p_{k+} = 1$. Очевидно, это можно записать как

$$\Sigma_{k,l} p_{kl}^2 / p_{k+} - \Sigma_l p_{+l}^2 = \Sigma_{k,l} p_{kl} a(l/k) = A,$$

что и доказывает утверждение.

5. Классификатор (устройство для распознавания категорий)

5.1. — Запомни, сынок, умный человек всегда во всем сомневается. Только дурак может быть полностью уверенным в чем-то.

— Ты уверен в этом, папа?

— Абсолютно.

5.2. — Свидетель, вы должны отвечать на вопросы кратко, без комментариев, только «да» или «нет». Вам понятно?

— Нет.

— Что вам не понятно?

— Да.

5.3. Покупательница спрашивает у торговки:

— Сколько у вас кур всего?

— Шесть.

— Выберите из них трех самых старых!

Торговка с готовностью быстро отбирает трех куриц.

— Вам упаковать их, уважаемая?

— Нет! Я беру трех остальных!

5.4. В ресторане сидит новый русский, подзывает официанта и говорит ему:

— Человек, ну-ка убавь кондишен, а то холодновато в натуре!

Официант уходит. Через 5 минут.

— Слышь, гарсон! (Тот подходит) Прибавь-ка, братан, кондишен, что-то жарковато стало. (Тот уходит.)

Через 5 минут.

— Эй, чувак, — официант подходит, — ну-ка убавь кондишен, опять холодно. (Официант уходит)

Сидящий за соседним столиком человек спрашивает официанта:

— Как ты все это терпишь?

Гарсон:

— Да без проблем! У нас и кондишена-то нет.

5.5. — Не зли меня, я страшна в гневе!

— Да ты и так не особо...

5.6. — Как у тебя с твоей девушкой?

— Мы расстались.

— А чего?

— Поругались. Она кричит: «Ты не любишь меня». Я ей: «Оля, да люблю я тебя!»

— А она?

— А она Лена.

5.7. Возвращается сын нового русского 1-го сентября из школы. Отец спрашивает, как ему там понравилось, на что сын отвечает:

— Да ну-у, как всегда, надули... Ни видака, ни телевизора, двадцать лохов сидят, да еще и парты со стульями деревянные. А говорили: «ПЕРВЫЙ КЛАСС!»

5.8. Муж, мрачный как туча, возвращается домой из больницы, где проведывал тещу. Жена встречает его у дверей.

Ж: — Как дела у мамы?

М: — Твоя мать здорова, как лошадь, скоро выйдет из больницы и будет жить у нас!

Ж: — Не понимаю. Вчера врач сказал мне, что она при смерти.

М: — Не знаю, что он сказал тебе, а мне он велел готовиться к самому худшему.

5.9. Изобретатель демонстрирует свое новое изобретение.

— Я разработал систему, которая позволяет установить личность человека по голосу.

— И что же я должен сделать?

— Вы должны четко и ясно назвать свои имя и фамилию...

5.10.

— Доктор, я смогу читать в этих очках?

— Конечно!

— Вот здорово! Я ведь раньше никогда читать не умел.

5.11. — Да добрая я, добрая. — бормотала Фея, вытирая кровь с волшебной палочки. — Только нервная немного.

5.12. — Доктор, помогите! Я совсем не разбираюсь в людях.

— Я не доктор.

5.13. —Что-то у тебя нездоровый вид.

— Я начал вести здоровый образ жизни.

5.14. В архивах французского министерства обороны на днях отыскалось любопытное письмецо.

Оно датировано 1960 годом и подписано человеком, призванным в армию, но вовсе не желавшим отправляться на войну в Алжир. Вот текст письма:

«Господин министр! Мне 24 года, я женат на вдове 44 лет, которая имеет 25-летнюю дочь.

Мой отец женился на этой девушке и таким образом стал моим зятем, поскольку он — муж моей дочери. Таким образом, моя падчерица стала моей мачехой, раз уж она — жена моего отца. У нас с женой родился сын. Он стал братом жены моего отца и двоюродным братом моего отца. И, соответственно моим дядей, поскольку он — брат моей мачехи. Таким образом, мой сын теперь — мой дядя. Жена моего отца тоже родила ребенка, который стал одновременно моим братом, раз уж он — сын моего отца, и моим внуком, поскольку он — сын дочери моей жены. Так как муж матери кого-либо является его отцом, получается, что я — отец своей жены, раз я — брат своего сына. Таким образом, я стал своим собственным дедом.

Учитывая вышеизложенное, господин министр, прошу вас принять необходимые меры для моей демобилизации, поскольку по закону нельзя призывать на службу одновременно сына, отца и деда. С надеждой на ваше понимание, примите, господин министр, уверения в моих искренних чувствах.»

5.15. — У вас явно завышенная самооценка.

— Вы так говорите, как будто я виноват в том, что я лучше вас.

5.16. — Мама, что это?

— Это черная смородина, доченька.

— А почему она красная?

— Потому что еще зеленая...

5.17. — Скажите, вам ведь всегда было трудно принимать однозначные решения?

— И да, и нет...

5.18. — Ты даже не заметил, что я покрасилась!

— Да заметил я, заметил!

— А я не покрасилась!

Тема 4
КОРРЕЛЯЦИЯ В МНОГОМЕРНЫХ ДАННЫХ

В данной теме будут рассмотрены примеры изучения взаимосвязей и корреляции в многомерных данных. Будут приведены и довольно подробно рассмотрены наиболее популярные методы:

— «наивный» Бейесов классификатор;

— линейная регрессия и дискриминация;

— нейронные сети и метод обратного распространения ошибки для их идентификации.

4.1. Введение: трудности коррелирования данных

Обычно при изучении связей в данных выделяют, как минимум, две группы признаков: прогнозирующие, или входные, признаки и целевые, или выходные, признаки. Обычно число целевых признаков невелико. Основные методы разработаны для случая, когда имеется только один целевой признак. Признак может выбираться в качестве целевого, если его трудно измерить или невозможно знать заранее. Поэтому хотелось бы получать такие «решающие» правила, чтобы для предсказания целевого признака было достаточно только измерения прогнозирующих признаков. Примеры областей и формулировок этой проблемы:

(а) химические соединения: входные признаки — особенности молекулярной структуры, целевые признаки — виды активности, такие как токсичность или лечебные свойства;

(б) виды зерна в сельском хозяйстве: входные признаки — свойства семян, грунта и особенности погоды, целевые признаки — урожайность или содержание клейковины;

(в) промышленные предприятия: входные признаки представляют технологии, инвестиции и персонал, в то время как целевые признаки относятся к продажам и прибыли;

(г) муниципальные районы в маркетинговых исследованиях: входные признаки связаны с демографическими, социальными и экономическими характеристиками, целевые признаки — с покупательским поведением в данных районах;

(д) банковский кредит клиентам: входные признаки характеризуют демографические факторы и доход, а целевым является факт, окажется ли клиент потенциально безнадежным должником или нет;

(е) данные экспрессии (считывания) генов: входные признаки связаны с уровнем экспрессии материалов ДНК на разных стадиях болезни, а выходные признаки характеризуют уровень болезненных проявлений.

Решающее правило предсказывает значение целевого признака по значениям входных признаков. Правило называется классификатором, если целевой признак является категориальным, и регрессией, если целевой признак имеет количественный характер. Задачи с категориальными целевыми признаками подразумевают, что множество объектов разделено на классы, соответствующие отдельным категориям. Решение задачи корреляции в таком случае заключается в построении такого решающего правила, которое для каждого объекта определит, принадлежит он данному классу или нет.

Решающее правило строится на основе подмножества объектов, на котором значения целевых признаков известны. Это подмножество часто называют обучающим. Множество объектов, на котором значения целевых признаков считаются неизвестными, используется для тестирования качества решающего правила. Идея процесса обучения заключается в том, чтобы определить такое решающее правило, которое бы минимизировало разницу между предсказанными и наблюдаемыми значениями целевого признака на обучающей выборке данных в классе допустимых решающих правил. Структура такого процесса представлена на верхней части рис. 4.1.

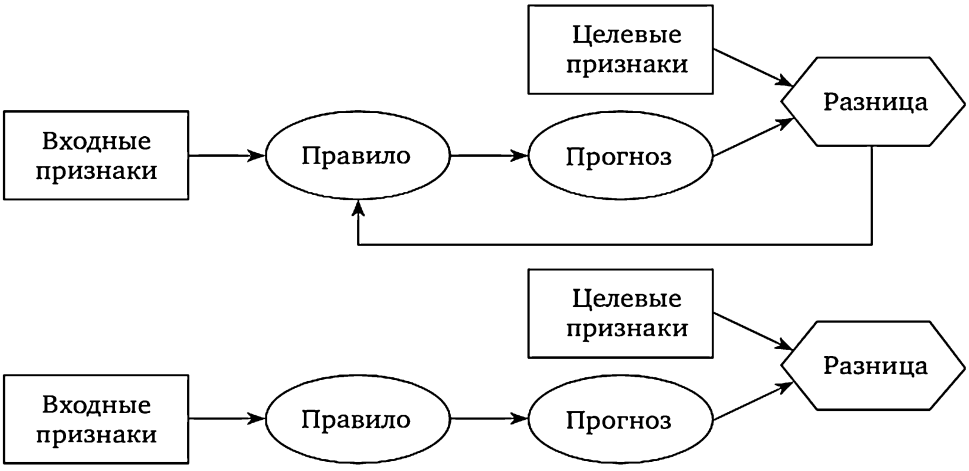


Рис. 4.1. **Двойная структура: обучение и тестирование:** наблюдаемые данные представлены прямоугольниками, вычислительные структуры представлены овалами, сравнение наблюдений с предсказаниями обозначено шестиугольником. В обучении (вверху) решающее правило подбирается так, чтобы минимизировать разницу между предсказанными и наблюденными значениями. В тестировании (внизу) сформированное на этапе обучения правило используется для прогноза; здесь нет обратной связи с правилом

Представление о том, что класс допустимых правил должен быть заранее задан, возникает из-за того, что обучающее множество конечно, так что, используя достаточное количество параметров, всегда можно «подогнать» решающее правило таким образом, чтобы ошибок на обучающем множестве не было вообще. Но, очевидно, что такое правило бы не сработало на тесте, так как подгонка захватывает все ошибки и шумы, неизбежные при сборе данных. Взгляните, например, на задачу двумерной регрессии. На рис. 4.2 изображены семь точек на плоскости (x, u), соответствующие наблюденным комбинациям входного признака x и целевого признака u. Эти семь точек — кружков на рис. 4.2 могут быть точно учтены полиномом шестой степени $u = p(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6$. В самом деле, при его оценке получим 7 уравнений $u_i = p(x_i)$ $i = 1, \ldots, 7$, так что 7 коэффициентов a_k полинома могут быть точно определены (при обычных условиях отсутствия сингулярности). Если число наблюденных точек — N, то для точной оценки потребуется полином $(N - 1)$-й степени.

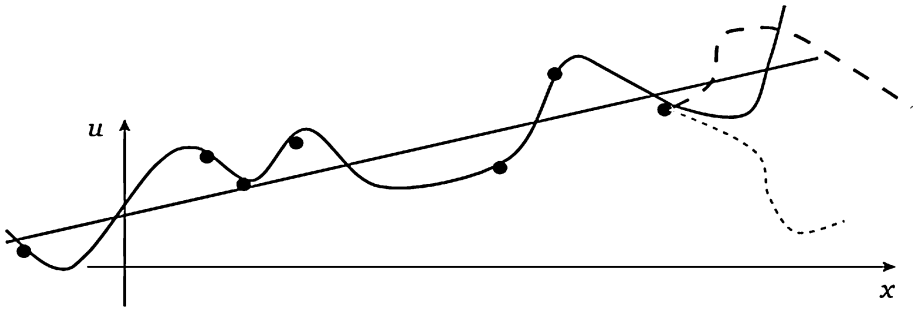


Рис. 4.2. **Возможные варианты экстраполяции графика связи *x* и *u***

Однако такой многочлен не сможет адекватно предсказать значения целевого признака ни внутри, ни за пределами диапазона. Кривая может пойти в любом направлении при самых небольших изменениях в данных.

Выбор правильного вида регрессионной функции, по-видимому, должен включать в себя понятие об «обобщающей силе» теории, определяемой этой функцией, которая для нашего случая сводится к отношению числа наблюдений к числу оцениваемых параметров: чем оно больше, тем лучше. Если это отношение относительно мало, статистики называют полученное правило избыточным или «переобученным». Переобученность обычно порождает малоинтересные (некачественные) прогнозы на вновь добавленных наблюдениях. Прямая линия не проходит ни через одну из семи точек, но выражает простую и надежную тенденцию. Следует отдать предпочтение этой прямой, поскольку она обобщает данные на более глубоком уровне: семь наблюдений обобщены здесь с использованием всего двух параметров: коэффициента наклона и свободного члена,

в то время как многочлен не дает никакого обобщения, он включает в себя столько же параметров, сколько имеется объектов. Именно по этой причине в задаче построения решающих правил в первую очередь нужно выбрать класс допустимых правил. Откуда его взять? К сожалению, на данный момент нельзя дать никакой общей рекомендации о том, как это можно сделать, кроме совета «посмотреть на форму графиков разброса». Без знаний о предметной области нет почвы для выбора класса решающих правил.

Самый популярный подход в компьютерных науках — так называемая бритва Оккама. Согласно этому подходу, решающее правило должна быть простым настолько, насколько это возможно. Считается, что данный совет реализует речение британского монаха Уильяма Оккама (ок. 1285 — 1349): «не следует множить сущности без крайней необходимости». Обычно это интерпретируется так, что при прочих равных простейшее объяснение считается наилучшим. Математически это выражается как «Принцип максимальной экономии», к которому обращаются, когда нет ничего лучшего. Будучи переформулирован как принцип «Минимальности длины описания», этот подход может быть осмысленно применен к проблеме оценки параметров статистических распределений (см. Grünwald 2007). Несколько более широкое и, возможно, более естественное толкование бритвы Оккама было предложено Вапником (2006). В несколько измененном виде, чтобы избежать смешения терминов, это толкование может быть представлено так: надо искать такое допустимое решающее правило, которое объясняет наблюденные факты, используя наименьшее число свободных параметров (Вапник, 2006, стр. 448). Однако даже в такой форме принцип не дает никаких рекомендаций о том, как выбрать адекватную функциональную форму. Например, какая из двух функций, степенная $f(x) = ax^b$ или логарифмическая $g(x) = b \cdot \log(x) + a$, более предпочтительна для суммаризации графиков на рис. 4.3: ведь обе функции имеют два параметра a и b?

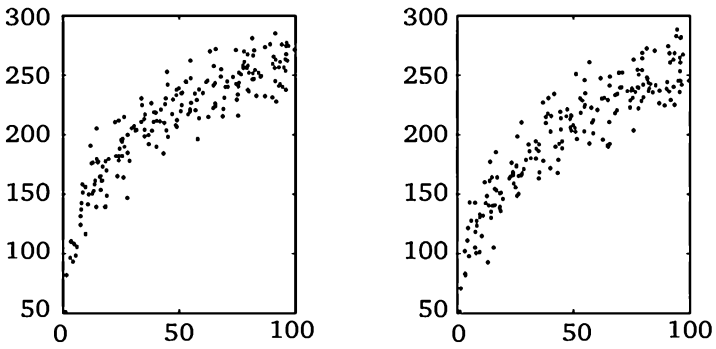


Рис. 4.3. **На рисунках представлены графики двух функций $f(x) = 65x^{0.3}$ и $g(x) = 50\log(x) + 30$, обе с добавлением шума $N(0, 15)$. Читателю предлагается попробовать определить, где какая функция**
Ответ: $f(x)$ изображена справа, $g(x)$ — слева

Другие советы подобного рода относятся к так называемому принципу опровергаемости (фальсифицируемости) К. Поппера (1902—1994), который может быть выражен следующим образом: надо объяснить факты с помощью такого допустимого решающего правила, которое проще всего опровергнуть (Вапник 2006, стр. 451). По идее, для того чтобы опровергнуть теорию, нужно привести пример, ей противоречащий. Фальсифицируемость решающего правила может быть определена в терминах так называемой *емкости* или ВЧ-сложности (*VC-complexity*, мера сложности, разработанная московскими учеными В. Н. Вапником и А. С. Червоненкисом в 1970-е гг.): чем ниже емкость, тем выше фальсифицируемость.

Рис. 4.4. **Пример сложного решающего правила**

Поясним понятие ВЧ-сложности для случая категориального целевого признака, когда решающее правило — классификатор. Говорят, что набор классификаторов Ф разбивает

обучающую выборку, если для любого разбиения обучающего множества в Ф найдется классификатор, точно воспроизводящий это разбиение. При заданном множестве допустимых классификаторов Ф ВЧ-сложность задачи классификации есть максимальное число объектов, которое можно разбить классификаторами из Ф.

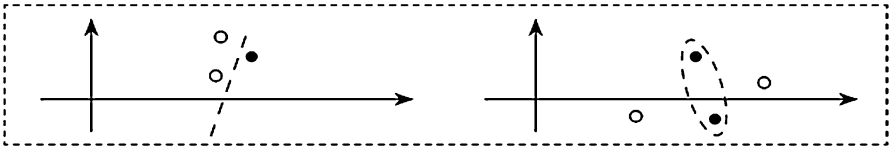


Рис. 4.5. **Любое разделение на две части множества трех точек (не находящихся на одной линии) может быть совершено с помощью прямой, но в представленном случае с четырьмя точками этого сделать невозможно**

Например, точки в двумерном пространстве имеют ВЧ-сложность, равную 3, в классе линейных решающих правил. Действительно, любые три точки, не лежащие на одной прямой, могут быть разбиты прямыми линиями. Однако не все множества из четырех точек могут быть разбиты прямыми. Последние два утверждения проиллюстрированы на правой и левой частях рис. 4.5, соответственно.

ВЧ-сложность — очень важная характеристика задачи коррелирования, особенно в рамках вероятностной парадигмы машинного обучения. В обычных условиях независимой случайной выборки данных надежный классификатор «с вероятностью a % будет точен в b % случаев, где b зависит не только от a, но также и от размера выборки и ВЧ-сложности» (Вапник, 2006).

4.2. Бейесовский подход к распознаванию

Задача изучения корреляции в таблице данных может быть сформулирована в общем виде следующим образом. Даны N пар (x_i, u_i), $i = 1, \ldots, N$, где x_i — это векторы размерности p прогнозирующих (входных) значений $x_i = (x_{i1}, \ldots, x_{ip})$, а $u_i = (u_{i1}, \ldots, u_{iq})$ — векторы размерности q целевых (выходных) значений (обычно $q = 1$). Зная эти пары, нужно построить решающее правило в предварительно заданном классе Ф допустимых правил F

$$\hat{u} = F(x) \tag{4.1}$$

такое, что в целом для наблюденных пар (x, u) разница между вычисленными $\hat{u}$ и наблюдаемыми u минимальна.

4.2.1. Бейесовское решающее правило

Рассмотрим ситуацию, когда имеется только один целевой признак. В простейшем случае этот признак является бинарным, т. е. присваивающим объектам принадлежность одному из двух классов, «положительному (да)» или «отрицательному (нет)». Согласно Бейесу[1] (1702—1761), вся значимая информация о мире должна быть представлена в виде вероятностных распределений, и поэтому любое наблюдение новых данных должно сказаться в изменении соответствующих распределений. Так возникает разница между априорными и апостериорными, до и после обновления данных, вероятностями. В частности, предположим, что вероятности классов «да» и «нет» равны $P(1) = p_1$ и $P(2) = p_2$, где p_1 и p_2 положительны и в сумме дают единицу. Это априорные вероятности двух состояний, «да» и «нет». Предположим, кроме того, что есть две функции плотности вероятности, $f_1(x_1, x_2, \ldots, x_p)$ и $f_2(x_1, x_2, \ldots, x_p)$, определя-

[1] В русскоязычной литературе обычно используется написание Байес (см. выше).

ющие распределение наблюдаемых точек объектов $x = (x_1, x_2, \ldots, x_p)$ в каждом из классов. Тогда для каждой точки $x = (x_1, x_2, \ldots, x_p)$ можно вычислить значение ее плотности вероятности $f(x)$. Для этого используем факт, что событие появления x — происходит как реализация одного из двух несовместных событий: x появляется в классе 1 (с вероятностью $p_1 f_1(x)$) и x появляется в классе 2 (с вероятностью $p_2 f_2(x)$). По правилам теории вероятности $f(x) = p_1 f_1(x) + p_2 f_2(x)$. Наблюдение объекта, представленного точкой $x = (x_1, x_2, \ldots, x_p)$, приводит к изменению вероятности классов от априорных $P(1) = p_1$ и $P(2) = p_2$ к апостериорным вероятностям $P(1/x)$ и $P(2/x)$ соответственно. Апостериорные вероятности могут быть вычислены с использованием известной теоремы Бейеса из элементарной теории вероятностей:

$$P(1/x) = p_1 f_1(x)/f(x) \text{ и } P(2/x) = p_2 f_2(x)/f(x) \tag{4.2}$$

По Бейесу, решение о том, к какому классу принадлежит объект x, определяется тем, какое из двух значений, $P(1/x)$ или $P(2/x)$, больше. Объект относится к классу «да», если $P(1/x) > P(2/x)$ или, что то же самое,

$$f_1(x)/f_2(x) > p_2/p_1 \tag{4.3}$$

или к классу «нет», если верно обратное неравенство. Это правило называется решающим правилом Бейеса. Другое представление Бейесова правила использует разность $B(x) = P(1/x) - P(2/x)$. Точка x относится к классу «да», если $B(x) > 0$, и классу «нет», если $B(x)<0$. Уравнение $B(x) = 0$ определяет так называемую разделяющую поверхность между двумя классами.

Доля ошибок при использовании правила Бейеса составляет $1 - P(1/x)$, когда принимается решение «да», и $1 - P(2/x)$, когда решение — «нет». Это минимальный достижимый процент ошибок при известных распределениях $f_1(x)$ и $f_2(x)$ и априорных вероятностях p_1 и p_2.

К сожалению, распределения $f_1(x)$ и $f_2(x)$, как правило, не известны. Поэтому вводят некоторые упрощающие предположения с тем, чтобы оценить распределение с помощью наблюдаемых данных. Наиболее популярны два постулата. Первый — что эти распределения Гауссовы (нормальные); второй — что они локально независимы.

Здесь рассматривается только предположение локальной независимости, так называемый «наивный» Бейесовский подход. Сконцентрируемся на ситуации, когда выходной признак — номинальный, причем не обязательно бинарный. Иными словами, на обучающем множестве задано разбиение на некоторое количество «целевых»

классов, а задача состоит в том, чтобы разработать решающее правило для их прогноза по входным признакам.

Локальная независимость распределения означает, что все признаки независимы в пределах каждого класса, так что распределение внутри класса k есть произведение одномерных распределений:

$$f_k(x_1, x_2, \ldots, x_p) = f_{k1}(x_1) f_{k2}(x_2) \ldots f_{kp}(x_p) \tag{4.4}$$

Это равенство значительно упрощает задачу оценки значений $f_k(x)$, поскольку обычно не составляет труда произвести достаточно точную оценку одномерной функции $f_{kv}(x_v)$ по обучающей выборке. Такая задача особенно упрощается, когда признаки x_1, x_2, ..., x_p сами являются бинарными. Во многих ситуациях постулат независимости явно не соблюдается, например, когда речь идет о категоризации текстов или анализе протеиновых цепочек — составляющие текста или белка, выступающие в качестве признаков, обязательно взаимосвязаны согласно семантической и синтаксической структуре в первом случае и согласно биохимическим реакциям во втором случае. Тем не менее решающие правила, основанные на неверных предположениях и распределениях, на практике приводят к на удивление хорошим результатам (см., например, Мэннинг и др., 2008). Бейесовское решающее правило, использующее постулат локальной независимости (4.4), называют Наивным Бейесовским классификатором.

4.2.2. Наивный Бейесовский классификатор

Рассмотрим задачу выявления корреляции по данным табл. 4.1: здесь объектами являются газетные статьи, разделенные на три категории в соответствии с темами «Феминизм», «Развлечения» и «Домохозяйство». Каждая статья характеризуется своим набором ключевых слов, представленных в соответствующей строке таблицы. Элементы каждого столбца показывают, сколько раз соответствующее ключевое слово встретилось в соответствующей статье.

Таблица 4.1

Иллюстративные данные о встречаемости 10 ключевых слов в 12 газетных статьях*

Статья	Ключевые слова									
	пить	равный	греть	играть	легкий	цена	свобода	талант	налог	женский
F1	1	2	0	1	2	0	0	0	0	2
F2	0	0	0	1	0	1	0	2	0	2
F3	0	2	0	0	0	0	0	1	0	2

Окончание табл. 4.1

Статья	Ключевые слова									
	пить	равный	греть	играть	легкий	цена	свобода	талант	налог	женский
F4	2	1	0	0	0	2	0	2	0	1
E1	2	0	1	2	2	0	0	1	0	0
E2	0	1	0	3	2	1	2	0	0	0
E3	1	0	2	0	1	1	0	3	1	1
E4	0	1	0	1	1	0	1	1	0	0
H1	0	0	2	0	1	2	0	0	2	0
H2	1	0	2	2	0	2	2	0	0	0
H3	0	0	1	1	2	1	1	0	2	0
H4	0	0	1	0	0	2	2	0	2	0

* Статьи помечены в соответствии с их главными темами — F означает тему феминизма, E — развлечений, H — домохозяйства.

Задача заключается в том, чтобы сформировать правило, с помощью которого любая статья, в том числе и не из табл. 4.1, может быть отнесена к одной из тематических категорий с помощью своего профиля — данных о частотах ключевых слов из табл. 4.1.

Сформируем наивное Бейесовское решающее правило. Каждой категории k оно присваивает условную вероятность $P(k/x)$, как это сделано в уравнениях (4.2), в зависимости от профиля x рассматриваемой статьи:

$$P(k/x) = \frac{p_k f_k(x)}{f(x)},$$

где $f(x) = \sum_l p_l f_l(x)$. По правилу Бейеса статья x относится к той категории k, для которой значение $P(k/x)$ максимально. Очевидно, что знаменатель в формуле для $P(k/x)$ не зависит от k и может быть отброшен. Таким образом, будет выбрана категория k с наибольшим значением $p_k f_k(x)$.

Условимся, что разные вхождения одного и того же ключевого слова не зависят друг от друга. Тогда для статьи с численностями появления ключевых слов, описываемых вектором $x = (x_1, x_2, \ldots, x_p)$, вероятность ее появления в k-й категории будет равна

$$f_k(x) = f_{k1}^{x_1} f_{k2}^{x_2} \ldots f_{kp}^{x_p}, \qquad (4.5)$$

где $f_{k1}, f_{k2}, \ldots, f_{kp}$ — вероятности появления соответствующих ключевых слов.

Остается договориться, о том, как можно оценить вероятности появления ключевых слов в данной категории. Это не так просто, как может показаться на первый взгляд. Например, какова вероятность появления слова «пить» в категории Н? Возможно, ее следует положить равной 1/4, поскольку это слово присутствует только в одной статье из четырех в Н. Но что тогда делать со словом «играть» в этой же категории — оно появляется трижды, но только в двух документах, поэтому вероятность его появления нельзя считать равной 3/4; однако значение 2/4 также не кажется правильным. Для оценки вероятностей частот ключевых слов используют так называемую модель «мешка слов».

Модель «мешок слов». Вместо общих определений, воспользуемся данными табл. 4.1. Прежде всего подсчитаем общее количество появлений ключевых слов в каждой целевой категории. Для категории Н по табл. 4.1 получаем 31. Идея: при расчете вероятностей относить количество появлений слов к «объему мешка». В этом и состоит модель «мешка слов» применительно к категории Н. К сожалению, этой модели самой по себе недостаточно. Дело в том, что таблица содержит слишком много нулей, — не потому, что слово не может встретиться в той или иной категории, а просто из-за случайностей в отборе статей в имеющейся выборке. Чтобы уменьшить эффект случайности отбора статей в таблицу данных, примем, что «мешок» обязательно содержит по одному появлению каждого ключевого слова, независимо от того, есть ли оно в статьях из данной категории или нет. Это добавляет 10 (по числу ключевых слов) к уже к наблюденным появлениям слов в категории Н. Данное допущение, вообще говоря, дополнительно к модели мешка слов. Оно является примером так называемого Лапласова сглаживания. Таким образом, полная емкость мешка Н становится равной 41 — сумме общего числа встречаемости ключевых слов в Н и количества ключевых слов. Вероятность каждого слова вычисляется как отношение количества его появлений в мешке к полному объему мешка. Следует отметить, что в некоторых учебниках модель мешка слов вводится с использованием теории вероятностей. Приведенный здесь вариант соответствует частному случаю вероятностной модели, связанному с «равномерными априорными вероятностями».

Таким образом, вероятности слов «пить», «греть» и «играть» в категории Н соответственно равны

$$(1 + 1)/41 = 2/41,$$

$$(6 + 1)/41 = 7/41$$

и

$$(3 + 1)/41 = 4/41.$$

Таблица 4.2

Априорные вероятности для наивного Бейесовского правила по данным из табл. 4.1, полученные применением модели «мешок слов»*

Категория	Априорная вероятность Её логарифм	Общее количество	Количества появлений слов Вероятности появления слов (в тысячных) Логарифмы «вероятностей»									
F	1/3	27	3	5	0	2	2	3	0	5	0	7
	—		108	162	27	81	81	108	27	162	27	216
	–1.099		4.6	5.1	3.3	4.4	4.4	4.7	3.3	5.1	3.3	5.4
E	1/3	32	3	2	3	6	6	2	3	5	1	1
	—		95	71	95	167	167	71	95	143	48	48
	–1.099		4.6	4.3	4.6	5.1	5.1	4.3	4.6	5.0	3.9	3.9
H	1/3	31	1	0	6	3	3	7	5	0	6	0
	—		49	24	171	98	98	195	146	24	171	24
	–1.099		3.9	3.2	5.1	4.6	4.6	5.3	5.0	3.2	5.1	3.2

* В трех строках, соответствующих каждой категории, находятся численности слов в документах этой категории, их вероятности, умноженные на 1000 и округленные до целых, а также натуральные логарифмы приведенных вероятностей.

При практическом применении наивного Бейесова классификатора удобно использовать не сами вероятности $P(k|x)$, а их логарифмы. Согласно уравнению (4.5), логарифм $log\ P(k|x)$ равен:

$$\log P(k/x) = \log p_k + x_1 \log f_{k1}(x_1) + \\ + x_2 \log f_{k2}(x_2) + \ldots + x_p \log f_{kp}(x_p). \quad (4.6)$$

Правая часть этого выражения — не что иное, как скалярное произведение вектора x и вектора логарифмов вероятностей появления соответствующих ключевых слов, $f_{k1}, f_{k2}, \ldots, f_{kp}$.

Априорные вероятности категорий считаются равными их долям в общей коллекции, 1/3 (второй столбец табл. 4.2).

Теперь мы можем применить Наивный Бейесовский классификатор к объекту, представленному в формате табл. 4.1, включая непосредственно объекты из табл. 4.1 (обучающая выборка). В табл. 4.3 приведены логарифмы оценок статьи E1 для каждой категории, рассчитанные по формуле (4.6).

Следует отметить, что при расчетах по методу наивного Бейесовского классификатора в задаче категоризации текстов, мы следовали так называемой мультиномиальной модели, в которой рассматриваются только вхождения слов в тексты. Другая популярная модель, называемая моделью Бернулли, предполагает, что слова генерируются независимо как биномиальные переменные. Вычисления, основанные на модели Бернулли, отличаются от представленных двумя моментами: во-первых, рассматриваются только действительно бинарные признаки, т. е., учитывается только бинарная информация о каждом слове (встретилось или нет); во-вторых, для каждого слова учитывается и вероятность его отсутствия (подробнее см. Мэннинг и др., 2008, Митчел 2010).

Вопрос 4.1. Применить Наивный Бейесовский классификатор в табл. 4.2 к статье, которая характеризуется вектором встречаемости ключевых слов X = (2 2 0 0 0 0 2 2 0 0), т. е. включает в себя по два вхождения слов «пить», «равный», «освободить» и «способности».

Ответ. Оценки категорий:

$$s(F/X) = 35.2,$$

$$s(E/X) = 35.6,$$

$$s(H/X) = = 29.4$$

указывают на категорию «Развлечения» или, что чуть менее вероятно, категорию «Феминизм».

Вопрос 4.2. Вычислить оценки наивного Бейесовского классификатора для всех объектов из табл. 4.1 и доказать, что он верно отнес их к их категориям.

Ответ. См. табл. 4.4.

Таблица 4.3

Вычисление оценки категории для объекта E1 (первая строка) из табл. 4.1 по логарифмам вероятностей признаков внутри каждого класса*

Категория	$\log(p_k)$	Объект E1										Оценка
		2	**0**	**1**	**2**	**2**	**0**	**0**	**1**	**0**	**0**	
		Веса признаков (логарифмы вероятностей) Скалярное произведение										
F	−1.099	4.6	5.1	3.3	4.4	4.4	4.7	3.3	5.1	3.3	5.4	35.2
		$2 \cdot 4.6$ + 0 + $1 \cdot 3.3$ + $2 \cdot 4.4$ + $2 \cdot 4.4$ + 0 + 0 + $1 \cdot 5.1$ + 0 + 0										
E	−1.099	4.6	4.3	4.6	5.1	5.1	4.3	4.6	5.0	3.9	3.9	**39.2**
		$2 \cdot 4.6$ + 0 + $1 \cdot 4.6$ + $2 \cdot 5.1$ + $2 \cdot 5.1$ + 0 + 0 + $1 \cdot 5.0$ + 0 + 0										
H	−1.099	3.9	3.2	5.1	4.6	4.6	5.3	5.0	3.2	5.1	3.2	34.5
		$2 \cdot 3.9$ + 0 + $1 \cdot 5.1$ + $2 \cdot 4.6$ + $2 \cdot 4.6$ + 0 + 0 + $1 \cdot 3.2$ + 0 + 0										

* Для каждой категории имеются две строки: верхняя повторяет логарифмы из табл. 4.2, а в нижней рассчитывается скалярное произведение из (4.6).

Таблица 4.4

Оценки методом наивного Бейесовского классификатора для объектов из табл. 4.1*

Статьи	Оценки категорий		
	F	E	H
F1	**37.7006**	35.0696	29.3069
F2	**28.9097**	25.9362	21.5322
F3	**24.9197**	20.1271	14.8723
F4	**38.2760**	34.6072	30.0000
E1	34.2349	**37.9964**	33.3322
E2	37.2440	**42.1315**	40.2435
E3	43.1957	**44.5672**	40.8398
E4	21.1663	**22.9203**	19.4367
H1	25.8505	29.3940	**34.5895**
H2	34.9290	40.4527	**42.7490**
H3	29.9582	35.3573	**38.3227**
H4	24.7518	28.8344	**34.8408**

* Максимумы по строке выделены жирным.

4.3 Меры качества классификатора

4.3.1. Точность и связанные с ней показатели

Рассмотрим общую задачу предсказания бинарного целевого признака, когда каждый объект обучения принадлежит либо классу 1, либо классу 2 этого признака. Решения классификатора могут быть как верными, так и ошибочными. Выберем один из классов, например, 1, как интересующий нас класс, скажем, некоего заболевания. Существует два вида ошибок: ложные «за» (ошибка первого рода, ЛЗ) — классификатор относит объект классу 1, хотя это не верно, и ложные «против» (ошибка второго рода, ЛП), когда классификатор отрицает принадлежность объекта классу 1, хотя на самом деле объект из этого класса.

Рассмотрим, например, устройство сканирования легких для тестирования на рак. Установленное в палате онкологического центра, устройство сканировало 200 пациентов; результаты представлены в табл. 4.5. Строки этой таблицы соответствуют диагнозу сканера, а столбцы — окончательным результатам, установленным с помощью дальнейших тестов. Эта таблица перекрестной классификации (сопряженности) по-английски часто называется *confusion*

table, что, вероятно, можно перевести как таблица ошибочности или перепутанности.

Таблица 4.5

Таблица сопряженности результатов сканирования легких

		Наличие заболевания		Всего
		Да	Нет	
Диагноз заболевания	Да	94	7	101
	Нет	1	98	99
Всего		95	105	200

Согласно табл. 4.5, есть 94 истинных «за» (ИЗ) и 98 истинных «против» (ИП), так что общая точность устройства может быть оценена как (94 + 98)/200 = 0.96 = 96 %. Соответственно, ложные «за» ЛЗ = 7 и ложные «против» ЛП = 1 в сумме дают 8, т. е. 4 % ошибок.

Однако существует значительное различие между этими двумя видами ошибок. Результаты, показанные сканером, на самом деле лучше, чем показывают итоговые оценки, потому что 7 ложных «за» не так уж и важны — с этими пациентами ничего не случится; дальнейшее исследование покажет отсутствие болезни — правда, с этим связаны определенные затраты. В то же время одно ложное «против» может привести к тому, что пациент останется без лечения, т. е. к потенциальной потере жизни из-за ошибки устройства. Это пример того, как отличаются потери, связанные с ложными «за» и ложными «против». Сканер сделал лишь одну серьезную ошибку, не установив один из 95 случаев заболевания раком. Доля истинных «за», равная доле верно установленных положительных случаев, часто называется мерой полноты или чувствительности (*recall or sensitivity*); в данном случае результат 94/95 = 98.9 % действительно впечатляет. С другой стороны, точность (precision), равная отношению 94 истинных «за» к 101 случаю диагноза «за» заболевание, несколько меньше, 93 %, что показывает также и долю ложных «за» в 7 %. Усредненное значение точности и чувствительности, равное 96 % в нашем случае, является достаточно хорошей мерой корректности (*accuracy rate*) данного устройства, и может быть выбрано для общей оценки качества.

Однако в ситуациях, когда обнаруживается большая разница между размерами положительного («да») и отрицательного («нет») классов, данная мера корректности работает не лучшим образом. Рассмотрим, например, результаты работы того же сканера, но теперь уже на другой, случайной, выборке из 200 человек «самотека», пришедших без направления врача (табл. 4.6).

Величина корректности для табл. 4.6 даже выше, чем для табл. 4.5, (2 + 195)/200 = 98.5 %. Тем не менее и чувствительность,

2/3 = 66.7 %, и точность, 2/4 = 50 %, весьма далеки от этого уровня. Высокий уровень корректности обусловлен тем, что велика специфичность (*specificity*) — доля правильно определенных случаев «нет», 195/197 = 98.9 %, а также тем фактом, что в случайной выборке очень мало случаев заболевания («да»).

Таблица 4.6

Таблица ошибок результатов сканирования легких на случайной выборке

		Наличие заболевания		Всего
		Да	Нет	
Диагноз заболевания	Да	2	2	4
	Нет	1	195	196
Всего		3	197	200

Поэтому в качестве единой меры корректности, адекватно отражающей и чувствительность, и точность, наиболее популярно не среднее арифметическое, а среднее гармоническое, так называемая F-мера, равная в данном случае F = 2/(1/(2/3)+1/(2/4)) = = 2/(3/2 + 4/2) = 4/7 = 57.1 %.

Задание 4.1. Индекс распространенности и коэффициент Кетле

Если посмотреть на результаты сканирования в табл. 4.6, где было обнаружено 4 случая заболевания, из 2 них два истинны, и сравнить этот результат с уровнем распространенности заболевания раком в выборке (всего 3 случая из 200), то разница окажется впечатляющей. Это разница и есть то, что обнаруживается коэффициентом Кетле $q(l/k)$ (см. п. 3.3.3.2) в строке $k = 1$ и столбце $l = 1$. Коэффициент равен относительной разности между условной вероятностью истинного «за» P(1/1) = 2/4 и средней долей (вероятностью) «за» на множестве, $P(1) = 3/200$. Эта последняя иногда называется индексом распространенности (prevalence):

$$q(1/1) = (2/4 - 3/200)/(3/200) =$$

$$= 2 \cdot 200/(3 \cdot 4) - 1 = 32.33 = 3233\ \%.$$

Такое высокое значение коэффициента Кетле, вероятно, и объясняет разницу характеристик чувствительности и специфичности в табл. 4.6 и 4.5.

В самом деле, такой же коэффициент Кетле для табл. 4.5 равен $q(1/1) = 94 \cdot 200/(101 \cdot 95) - 1 = 0.96 = 96\ \%$, менее чем 100%-ное увеличение. Это показывает, что табл. 4.5 более сбалансирована, чем табл. 4.6. Характеристика корректности работает хорошо на сбалансированных таблицах и не совсем удовлетворительно на несбалансированных.

Ф4.3.2. Точность и связанные с ней показатели: Формулировки

Рассмотрим общий случай таблицы ошибок распознавания одной из двух категорий, представленный в табл. 4.7. Конечно, если нас будет интересовать не класс 1, а класс 2, то ошибки останутся ошибками, но их названия изменятся: ложные «за» класс 1 станут ложными «против» относительно класса 2.

Перечислим некоторые популярные индексы уровня ошибки или точности.

Доля лз = ЛЗ/(ЛЗ+ИП) — доля ложных «за» среди объектов, не принадлежащих классу 1 («да»). Величина (1 – ЛЗ) иногда называется специфичностью: она показывает долю правильно предсказанных объектов среди всех объектов, не принадлежащих классу 1.

Таблица 4.7

Статистическое представление ошибок распознавания класса «да»

		Истинный класс		Всего
		Да	Нет	
Предсказанный класс	Да	Истинные «за» (ИЗ)	Ложные «за» (ЛЗ)	ИЗ+ЛЗ
	Нет	Ложные «против» (ЛП)	Истинные «против» (ИП)	ЛП+ИП
Всего		ИЗ+ЛП	ЛЗ+ИП	*N*

Чувствительность, она же полнота, ИЗ/ (ИЗ + ЛП) — доля истинных «за» в «реальном» классе 1.

Точность Из = ИЗ/(ИЗ + ЛЗ) — доля истинных «за» в предсказанном классе 1.

Эти индексы отражают каждую из возможных ошибок в отдельности.

Рис. 4.6. **Ложное «против»**

Из комплексных индексов следует отметить корректность и *F*-меру.

Корректность (ИЗ + ИП)/N — общая доля точных предсказаний. Очевидно, что 1 – Корректность — это общая доля ошибок.

F-мера = 2/(1/Точность + 1/Чувствительность) — гармоническое среднее показателей Точности и Чувствительности.

Как отмечалось выше, последняя мера становится все более популярной, потому что среднее гармоническое более адекватно, чем среднее арифметическое, в тех частых случаях, когда ошибки одного рода обходятся дороже ошибок другого рода. Вспомним, например, случай с медицинской диагностикой в табл. 4.5—4.6: ситуация, когда доброкачественная опухоль диагностируется как злокачественная, гораздо менее серьезна, чем та, в которой, наоборот, злокачественная опухоль диагностируется как безвредная. Величина F-меры до некоторой степени является консервативной оценкой, потому что, во-первых, она использует доли, а не абсолютные величины встречаемости, и, во-вторых, применяет среднее гармоническое, которое значительно ближе к минимуму из двух, как видно из ответов на вопросы 4.3 и 4.4.

Вопрос 4.3. Рассмотрим два положительных действительных числа a и b и предположим, например, что a меньше, чем b. Докажите, что среднее гармоническое $h = 2/(1/a + 1/b)$ лежит в интервале от a до $2a$, независимо от величины разности $b - a$.

Ответ. Пусть b представляется в виде $b = ka$ с некоторым коэффициентом $k > 1$. Тогда $h = 2/(1/a + 1/(ka)) = 2ka/(1 + k)$. Коэффициент при a, равный $2k/(1 + k)$, меньше 2, что доказывает утверждение.

Вопрос 4.4. Рассмотрим два положительных действительных числа a и b. Докажите, что их среднее, $m = (a + b)/2$, и гармоническое среднее, $h = 2/(1/a+1/b)$, удовлетворяют уравнению $mh = ab$.

Ответ. Запишем произведение $mh = [(a + b)/2][2/(1/a + 1/b)]$ и проделаем элементарные алгебраические преобразования.

Более сложное представление ошибок двух типов может быть достигнуто с помощью анализа графиков ROC, так называемой рабочей характеристики приемника (*Receiver Operating Characteristic*) — графиков зависимости чувствительности от доли ложных «за» (см. [40]). Кривые ROC особенно хорошо подходят для случаев классификаторов с непрерывным выходом, таких как Бейесовские классификаторы. Кривая ROC в двумерной декартовой системе координат показывает зависимость доли истинных «за» (по оси y) от доли ложных «за» (по оси x), рис. 4.7.

Для определенности возьмем правило Бейесовского классификатора из (4.4) и изменим отношение p_2/p_1 на произвольный порог $d > 0$. Возьмем теперь $d = d_1$ для конкретного d_1, так что правило теперь предсказывает класс 1, если $f_1(x)/f_2(x) > d_1$. Посчитаем доли истинных и ложных «против», $tp1$ и $fp1$, при заданном пороговом значении и отметим точку ($fp1$, $tp1$) на кривой ROC. Теперь поменя-

ем d на d_2 и сосчитаем доли $tp2$ и $fp2$ при этом пороговом значении. Если, скажем, $d_2 > d_1$, то доля истинных «за» может только уменьшиться, поскольку число положительных предсказаний может только уменьшиться. Доля ложных «за», вообще говоря, должна бы возрастать при $d_2 > d_1$, т. е. точка ($fp2$, $ft2$) будет двигаться вправо и вверх относительно предыдущей точки на графике ROC. Таким образом, шаг за шагом, меняя пороговое значение d, можно получить экспериментальную кривую ROC, такую как a и b на рис. 4.7. Такая кривая может быть использована в качестве характеристики рассматриваемого классификатора для выбора, например, подходящих уровней долей истинных и ложных «за». В случае, показанном на рис. 4.7, можно с уверенностью утверждать, что классификатор a лучше классификатора b, потому что для каждого значения доли ложных «за» доля истинных «за» у классификатора a больше, чем у классификатора b.

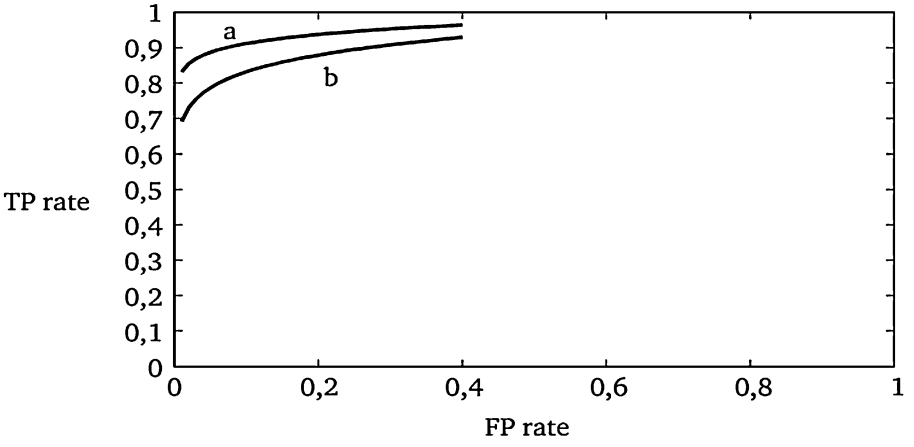


Рис. 4.7. **Кривые ROC для двух классификаторов; классификатор *a* лучше классификатора *b***

4.4. Нейронные сети для представления отображения «вход-выход»

4.4.1. Искусственные нейроны и нейронные сети

Нейронная сеть — одна из наиболее популярных структур, используемых для предсказания целевых признаков по входным. Ее элементами являются искусственные нейроны, моделирующие нейроны живых организмов. Нейронная клетка в живом организме испускает сигнал, когда сумма входных сигналов становится больше определенного порога. Входные сигналы поступают по дендри-

там, выходные — по аксонам, а сигнал формируется в синапсе — месте контакта между нейронами, которое и порождает этот порог (см. рис. 4.8).

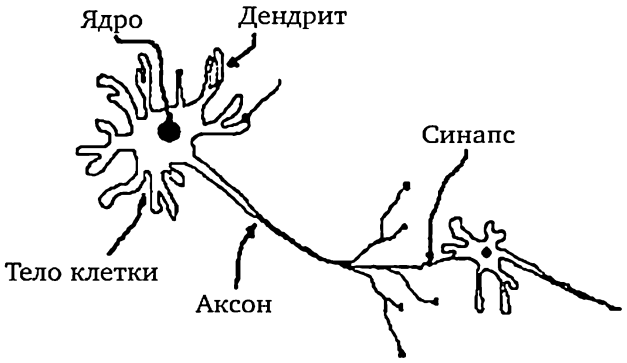


Рис. 4.8. **Схема нейронной клетки**

Искусственный нейрон, соответственно, представляет собой блок, соединяющий входные сигналы с выходным сигналом. Каждому соединению приписан определенный вес (рис. 4.9).

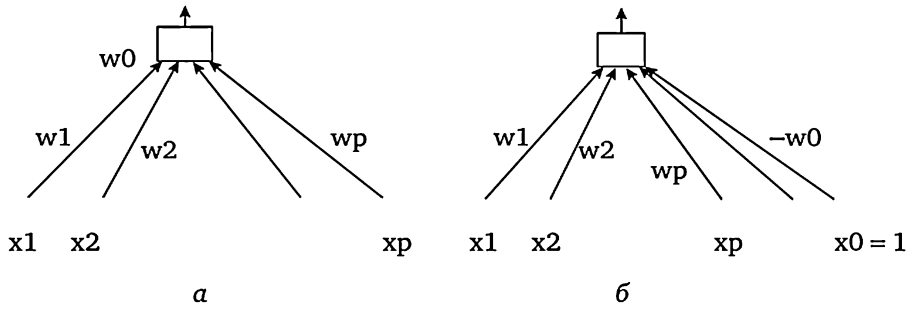


Рис. 4.9. **Искусственный нейрон:**
а — схема искусственного нейрона; *б* — этот же нейрон с нулевым порогом: исходный порог переведен в вес ребра, соединяющего выход с добавленным искусственным входным сигналом, всегда равным 1

Входные сигналы — признаки данных или же выходы других нейронов. На выходной элемент искусственного нейрона подается сумма значений признаков, помноженных на веса соединений. Выходной элемент осуществляет так называемую функцию активации искусственного нейрона. В простейшем случае он сравнивает получаемую сумму с значением порога (называемого также *смещение*). Выходной сигнал равен 1, если сумма выше порога, и –1, если ниже. Таким образом, в данном случае функция активации — это то, что называется «сигнум» функция, sign(x), значение которой равно 1, 0 или –1, если x положителен, равен нулю, или отрицателен, соответ-

ственно. Эта функция не очень удобна для математических манипуляций. Как правило, ее заменяют так называемой сигмоидой или даже симметрической сигмоидой, представляющими собой «мягкие» аналоги функции $sign(x)$. Графики этих активационных функций показаны на рис. 4.10. Иногда функция активации осуществляет тождественное преобразование $f(x) = x$ — тогда ее называют линейной. Очень популярна так называемая очищенная линейная функция активации (Relu) $f(x) = max(0, x)$, которая не дает сигнала, если вход отрицательный.

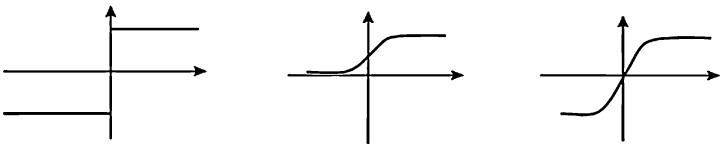


Рис. 4.10. **Графики функций:**
а — сигнум (а), *б* — сигмоида, *в* — симметрическая сигмоида

Порог срабатывания нейрона, смещение, спрятан в выходном блоке рис. 4.9, *б*. Его можно сделать явным, если добавить дополнительный входной сигнал. Этот фиктивный вход всегда равен 1, так что его вес $-w0$ всегда добавляется к сумме входов нейрона (см. рис. 4.9, *а*). Далее предполагается именно такая конфигурация (она соответствует конструкциям линейной алгебры).

Нейронная сеть — это соединение некоторого количества искусственных нейронов. Существует значительное количество типов структур нейронных сетей (называемых архитектурами). Самая, пожалуй, распространенная — это трехслойная структура без обратных связей типа той, что представлена на рис. 4.11 ниже. В ней два крайних слоя, входной и выходной, а также промежуточный слой, называемый скрытым, поскольку его заслоняют два крайних слоя. Эту структуру часто называют «архитектура прямого распространения с одним скрытым слоем».

Нейронная сеть на рис. 4.11 представляет собой как раз такую архитектуру, созданную для предсказания размеров лепестка по размерам чашелистика на таблице данных Ирис. Напомним, что каждый из 150 экземпляров Ириса характеризуется четырьмя признаками, из которых первые два характеризуют длину и ширину чашелистика (признаки w1 и w2), а вторые два (признаки w3 и w4) — длину и ширину лепестка. Ожидается, что размеры чашелистика и лепестка коррелируют.

На самом деле нижеследующий материал может использоваться для формирования нейронной сети данной архитектуры для моделирования корреляции между любыми входами и выходами. Един-

ственная разница — в количествах входных и выходных признаков — не играет особой роли в последующих вычислениях.

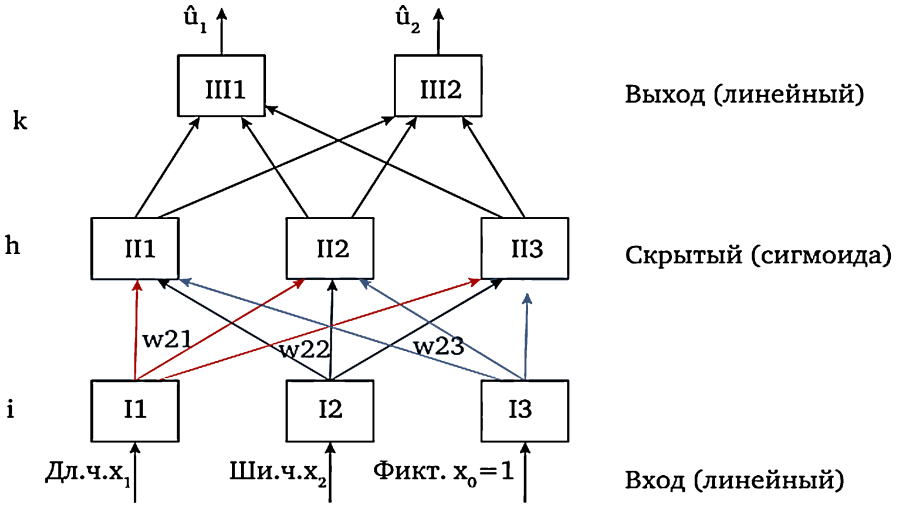


Рис. 4.11. **Нейронная сеть прямого распространения с одним скрытым слоем, двумя входами и двумя выходами. Слои:** Входной (I, с индексом *i*); Выходной (III, с индексом *k*); Скрытый (II, с индексом *h*)

Эта нейронная сеть имеет три слоя.

1. Входной слой с тремя входами: смещение $x_0 = 1$, как объясняется выше (см. рис. 4.9, *б*), а также длина и ширина чашелистика, взвешенные суммы которых поступают на вход каждого нейрона скрытого слоя.

2. Выходной слой, нейроны которого производят оценки длины и ширины лепестка с помощью линейной функции активации. Входные сигналы здесь — комбинации выходов скрытого слоя. Причем отсутствует фиктивный сигнал $x_0 = 1$, поскольку никакого порога не предполагается.

3. Скрытый слой содержит три нейрона. Функция активации каждого нейрона преобразует входной сигнал от входного слоя помощью сигмоидной функции активации. Выходные сигналы этих трех нейронов комбинируются в входные сигналы выходного слоя. На самом деле эта архитектура позволяет использовать в скрытом слое любое количество нейронов без какого-либо изменения вычислений (в матричном виде).

Эта структура в определенном смысле универсальна. Например, было доказано, что ее можно обучить распознавать любое подмножество объектов. Более того, любое отображение входов в выходы может быть аппроксимировано с заранее заданной точностью функцией, реализуемой структурой прямого распространения с од-

ним скрытым слоем, если увеличить число нейронов в скрытом слое достаточным образом. Это свойство почти 30 лет служило основанием того, что исследователи не стремились рассматривать более многослойные структуры. Последнее десятилетие увидело взрывной, и очень успешный, рост исследований по многослойным сетям, содержащим до 10 и более слоев, — так называемое глубокое обучение (см. [4, 16]). Оказалось, глубокие сети способны решать тонкие высокоинтеллектуальные задачи самообучения, существенно продвинувшие практические исследования в области машинного перевода текстов, распознавания речи, изображений и других семантически-нагруженных задач. Это происходит за счет резкого роста количества параметров многослойной (глубокой) сети, оценка которых требует соответствующего увеличения объемов данных для обучения. Кроме того, появляется возможность усреднения определенных фрагментов сети (сверточные нейронные сети), а также учета предыдущих состояний скрытых слоев (рекуррентные сети). Такое ощущение, что обученные на огромных массивах «примеров» глубокие сети способны «увидеть» сложные нелинейные признаки объектов, которые позволяют существенно улучшить точность принятия решений. Эти нелинейные признаки, однако, по-настоящему скрыты от исследователя, будучи рассыпанными на многие тысячи индивидуальных весовых коэффициентов. Возникает «ловушка», описанная писателем-фантастом А. Азимовым в его цикле «Я, робот», когда машина, предоставленная сама себе, может повести себя непредсказуемым, и даже опасным образом. Эту ловушку в настоящее время пытаются преодолеть на пути разработки «объяснимых» систем глубокого обучения.

4.4.2. Функция активации и преобразование, задаваемое нейронной сетью

Функция активации сигмоида определяется формулой:

$$s(x) = (1 + e^{-x})^{-1} \tag{4.6}$$

задающей гладкий аналог сигнум-функции, хотя значения (4.6) лежат в интервале между 0 и 1, тогда как сигнум-функция меняется от –1 до 1. Чтобы «растянуть» интервал значений до [–1, 1], нужно умножить $s(x)$ в (4.6) на 2 и вычесть 1 из результата. При этом получается так называемая симметрическая сигмоида, или гиперболический тангенс:

$$\text{th}(x) = 2s(x) - 1 = 2(1 + e^{-x})^{-1} - 1. \tag{4.7}$$

Эта функция, называемая симметрической сигмоидой или гиперболическим тангенсом, кососимметрична, $\text{th}(-x) = -\text{th}(x)$, как и сигнум.

Сигмоидные функции активации обладают хорошими математическими свойствами. Они не только гладкие, но и имеют производные, выражаемые через сами функции (см. **Вопрос** 4.5 далее).

Выразим преобразование входов в выходы, задаваемое сетью на рис. 4.11, аналитически. Обозначим матрицу весов соединений входного и скрытого слоев через $W = (w_{ih})$, где i — индекс входа, а h — индекс скрытого нейрона, $h = 1, 2, \ldots, H$ (H — число нейронов скрытого слоя). Веса соединений скрытого и выходного слоев образуют матрицу $V = (v_{hk})$, где h — индекс скрытого нейрона, а k — выхода.

Предполагается, что выходные элементы слоев I и III дают сигналы, совпадающие с входами (линейная функция активации), а все нейроны скрытого слоя имеют симметрическую сигмоиду в качестве функции активации.

Рассмотрим последовательность преобразований, осуществляемую слоями сети рис. 4.11. Входом нейрона h скрытого слоя является комбинация

$$z_h = w_{1h}x_1 + w_{2h}x_2 + w_{3h}x_0,$$

т. е. h-я компонента вектора $z = \sum_i x_i \cdot w_{ih} = x \cdot W$, где x — это 1×3 входной вектор. Тогда на выходе этого нейрона будет $\mathrm{th}(z_h)$. Значит, на выходе скрытого слоя имеем вектор $\mathrm{th}(z) = \mathrm{th}(x \cdot W)$, который подается на вход выходного слоя. Его k-й элемент получает комбинированный сигнал $\sum_j v_{jk} \cdot \mathrm{th}(z_j)$ — k-я компонента матричного произведенията $\mathrm{th}(z) \cdot V$, которое и является выходом выходного слоя $\hat{u}$. Таким образом, нейронная сеть на рис. 4.11 преобразует входной вектор x в выходной вектор $\hat{u}$ согласно формуле:

$$\hat{u} = \mathrm{th}(x \cdot W) \cdot V, \tag{4.8}$$

в которой присутствуют линейные операции матричного умножения и нелинейное преобразование симметрической сигмоиды. Если матрицы W, V известны, то (4.8) определяет преобразование сети $u = F(x)$. Проблема состоит в том, чтобы найти эти матрицы, используя данные, т. е. таблицу Ирис.

4.4.3. Обучение нейронной сети

При заданных весах W и V соединений между входным и скрытым слоями и скрытым и выходным слоями, соответственно, а также при заданных функциях активации нейронная сеть (см. рис. 4.11) преобразует входные сигналы, длину и ширину чашелистика, в выходные сигналы, оценки длины и ширины лепестка.

Для оценки качества этих оценок используется среднеквадратическая ошибка. Чем лучше оценены матрицы W и V, тем меньше эта ошибка. Где же взять величины весов W и V? Из имеющихся дан-

ных. Возможны разные подходы. Парадигма машинного обучения предполагает, что машина обучается путем постепенного получения информации об объектах, один за одним. Считается, что полная таблица машине неизвестна, т. е. глобальные решения типа ортогонального проецирования в линейной дискриминации не применимы. В этой ситуации мы должны ограничиться таким алгоритмом минимизации, который обрабатывает данные об объектах не одновременно, а последовательно. Такой метод есть. Это так называемый градиентный, точнее, анти-градиентный, метод, именуемый также методом наискорейшего спуска.

Метод использует градиент минимизируемой функции. Градиент — это вектор, показывающий направление наибольшего подъема функции, рассматриваемой как многомерная поверхность, в той точке, в которой градиент вычислен или оценен. Его компонентами в данном случае являются частные производные этой функции. Понятно, как использовать градиент для максимизации функции: двигаться от решения к решению в направлении градиента. А как использовать градиент для минимизации? Общее мнение — двигаться в противоположном направлении, т. е. используя минус градиент.

Пусть известны какие-то оценки матриц W и V и их градиентов, выраженных матрицами gW и gV. Компоненты этих матриц выражают направление наибольшего подъема для изменения матриц W и V. Согласно методу наискорейшего спуска, следует двинуть V и W в направлении $-gW$ и $-gV$, контролируя длину шага коэффициентом, называемым скоростью обучения. Уравнения, выражающие переход от старого состояния матриц W и V к новому:

$$V' = V - \mu \cdot gV,\ W' = W - \mu \cdot gW, \tag{4.9}$$

где μ — скорость обучения (длина шага), а штрих обозначает новое состояние.

Рис. 4.12 иллюстрирует значимость правильного выбора длины шага.

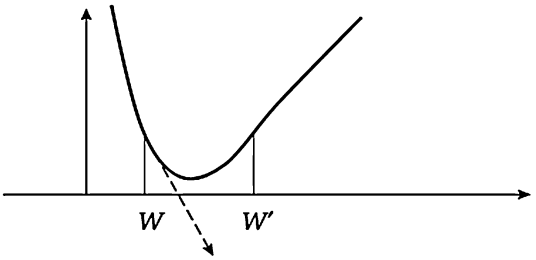


Рис. 4.12. **Значимость правильного выбора длины шага в градиентном процессе:** при очень длинном прыжке новое состояние, W' может оказаться хуже, чем старое W

Как будет показано далее, градиент квадратичной ошибки определяется: (a) матрицами *W* и *V*, (b) самим значением ошибки, (c) значениями входных признаков. Поэтому удобно применять этот подход, подавая объекты последовательно (в случайном порядке). Тогда каждый отдельный объект дает новую оценку градиента и, соответственно, новые значения элементов матриц *W* и *V* согласно уравнениям (4.9). Поскольку количество объектов относительно невелико (всего 150 экземпляров Ириса, например), а градиентный процесс довольно медленный, обычно одноразовой обработки множества объектов недостаточно. Одноразовую обработку называют *эпохой*. Определенное, обычно заранее задаваемое количество эпох, скажем, 5000, задается пользователем. Порядок поступления объектов в каждую данную эпоху определяется заново случайным образом. Полученные матрицы *W* и *V* рассматриваются в качестве результата.

Задание 4.2. Обучение размерам лепестка

Таблица 4.8

Относительные значения ошибок в размерах лепестков после 5000 эпох (в процентах к размаху признака)

Количество нейронов в скрытом слое	Относительная ошибка, %	
	Длина	Ширина
3	5.36	8.84
6	3.99	8.40
10	3.98	8.15

Примените алгоритм обратного распространения ошибки к нейронной сети рис. 4.11 для прогноза значений признаков длины и ширины лепестка по признакам длины и ширины чашелистика по таблице данных Ирис. MATLAB-программа nnn.m, реализующая этот алгоритм, размещена в Приложении № 4.

Определите количество элементов в матрицах V и W как функцию числа p нейронов скрытого слоя. (Ответ: (3 + 2)p = 5p).

Убедитесь, что уровень ошибок после 5000 эпох примерно соответствует результатам в табл. 4.8. В частности, что дальнейшее увеличение числа нейронов скрытого слоя практически не сказывается на уровне ошибки.

Самостоятельная работа

4.1. Посмотрите, что произойдет, если не нормализовать данные.

4.2. Посмотрите, что произойдет, если скорость обучения уменьшится (увеличится) в пять раз.

4.3. Попробуйте поменять местами входы и выходы, т. е. займитесь прогнозом размеров чашелистика по размерам лепестков.

4.4. Дальнейшее изменение задачи: рассмотрите задачу прогноза ширины лепестка по значениям длины и ширины чашелистика, а также длины лепестка.

Задание 4.3. Контроль качества алгоритма машинного обучения

В отличие от анализа данных, имеющего целью компактное описание данных, машинное обучение должно производить алгоритмы, работающие не только на данной таблице, но и на других аналогичных данных. Поэтому протокол машинного обучения должен предусматривать случайное разделение обучаемой выборки на части. Какие-то из этих частей используются для «обучения», т. е. настройки параметров алгоритмов, так чтобы по возможности наилучшим образом решить решаемую задачу, а другие, дополнительные, части — для проверки качества этих решений. Два основных метода формирования такого разделения — это (а) контроль на отложенных данных (*hold-out method*) и (б) контроль по k блокам (*k-fold cross-validation*).

Согласно методу контроля на отложенных данных, при заданной вероятности p (обычно $p = 0.2$ или $p = 0.3$) имеющаяся выборка случайно разбивается на две части в пропорции $p/(1-p)$, т. е. обычно 20 %/80 % или 30 %/70 %, так что $(1-p)$-часть используется для обучения, а p-часть — для тестирования. Применительно к нейронным сетям это означает, что при $p = 0.2$ 80 % выборки используется для подбора весовых матриц W и V, а оставшиеся 20 % — для вычисления среднеквадратической ошибки при этих W и V.

Согласно методу контроля по k блокам вся имеющаяся выборка разбивается на k равных (по возможности) частей. Обычно k берется равным 2 или 10. Тогда при $N = 367$ выборка разбивается на части, содержащие 183 и 184 объекта (для $k = 2$) или содержащие по 36 (3 части) и 37 объектов (7 частей), для $k = 10$. Затем для каждой из полученных k частей производится следующее. Эта часть используется для тестирования, а остальные $k - 1$ частей объединяются в единое множество, на котором производится обучение модели.

Обученная модель применяется к данной части и рассчитывается получаемая ошибка. Результирующая ошибка — среднее арифметическое всех полученных на k частях ошибок.

Самостоятельная работа

4.5. Применить методы контроля на отложенных данных и контроля по k блокам на данных Ирис по 10 раз. Рассчитать средние ошибки и их стандартные отклонения. Сравнить полученные результаты.

4.4.4. Настройка нейронной сети и градиентная оптимизация

Ф4.4.4.1. Метод наискорейшего спуска для критерия квадратичной ошибки

Машинное обучение исходит из предположения, что решающее правило настраивается постепенно, используя объекты поочередно. Предполагается, что глобальные методы, использующие всю

доступную выборку, неприменимы. В этой ситуации применяют такие оптимизационные алгоритмы, которые обрабатывают объекты по мере их поступления. Таков градиентный метод, называемый также методом наискорейшего спуска.

Для минимизации функции $f(x)$ относительно x из подпространства D n-мерного векторного пространства R^n, можно использовать ее градиент gf. Градиент gf в точке $x \in D$ — это вектор, компонентами которого являются частные производные f по отдельным компонентам x, в предположении, что существует полная производная, соответствующая касательной гиперплоскости к поверхности f в этой точке. Этот вектор показывает направление наибольшего подъема $f(x)$, так что противоположный вектор $-gf$ показывает противоположное направление, принимаемое за направление скорейшего спуска $f(x)$. Метод наискорейшего спуска производит последовательность точек $x(0)$, $x(1)$, $x(2)$, ..., начиная с произвольного $x(0)$, используя рекуррентное соотношение

$$x(t+1) = x(t) - \mu_t \cdot gf(x(t)), \tag{4.10}$$

где параметр μ_t характеризует длину шага от $x(t)$ в направлении скорейшего спуска и называется скоростью обучения. Последовательность $x(t)$ гарантированно сходится к точке минимума при постоянном $\mu_t = \mu$, если $f(x)$ строго выпукла, так что существует сфера конечного радиуса такая, что $f(x)$ всегда больше, чем нижняя часть этой сферы, как показано на рис. 4.13 (см. Поляк 1987).

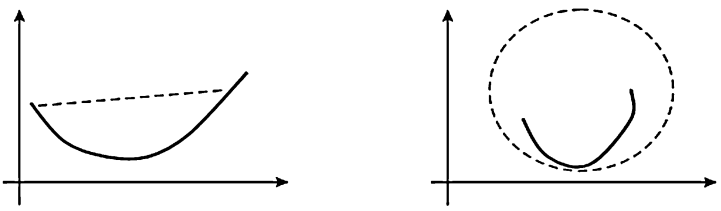


Рис. 4.13. **Выпуклая функция (*а*) и строго выпуклая функция (*б*)**

Если $f(x)$ — выпуклая (не обязательно строго) функция, то сходимость процесса можно гарантировать только, если μ_t стремится к 0 при увеличении t, но не слишком быстро, так что сумма ряда $\Sigma_t \mu_t$ бесконечна. Эти условия гарантируют, что шаги от $x(t)$ к $x(t+1)$ достаточно малы для того, чтобы не «перепрыгнуть» точку минимума, но не слишком малы, так чтобы движение не заглохло просто из-за малости шага.

Если, однако, $f(x)$ не выпуклая, то эта последовательность сходится к какой-либо точке локального минимума, зависящей от выбора начальной точки $x(0)$ (см. рис. 4.14). На этом основан алгоритм, описанный в следующем разделе.

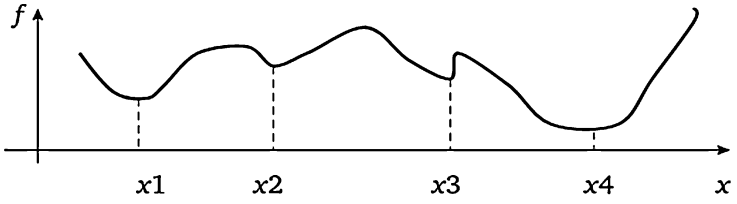


Рис. 4.14. **Точки *x1*, *x2*, *x3*, *x4* — точки локальных минимумов функции, график которой представлен на рисунке. Глобальный минимум достигается в одной из них, x4**

Ф4.4.4.2 Метод обратного распространения ошибки

Проблема обучения нейронной сети состоит в том, чтобы найти весовые матрицы W и V, минимизирующие квадратичную разницу между наблюденным выходом u и тем $\hat{u}$, которое рассчитано по преобразованию, реализованному сетью,

$$E = d(u, \hat{u}) = <u - \text{th}(x \cdot W) \cdot V,\ u - \text{th}(x \cdot W) \cdot V>/2, \qquad (4.11)$$

по всем элементам таблицы данных. Деление на 2 в (4.11) сделано, чтобы избежать удвоения в производных функции E.

Для случая двух выходных нейронов (см. рис. 4.11) функция ошибки имеет вид

$$E = [(u_1 - \hat{u}_1)^2 + (u_2 - \hat{u}_2)^2]/2, \qquad (4.12)$$

где $u_1 - \hat{u}_1$ и $u_2 - \hat{u}_2$ — разности между наблюденными и вычисленными значениями двух выходных сигналов.

Уравнения (4.10) для покомпонентного пересчета V и W имеют следующий вид:

$$v_{hk}(t+1) = v_{hk}(t) - \mu \cdot \partial E / \partial v_{hk},$$
$$w_{ih}(t+1) = w_{ih}(t) - \mu \cdot \partial E / \partial w_{ih} \quad (i \in \text{I}, h \in \text{II}, k \in \text{III}). \qquad (4.13)$$

Найдем явные выражения для производных, участвующих в (4.13). Сначала продифференцируем выходы, по v_{hk}:

$$\partial E / \partial v_{hk} = -(u_k - \hat{u}_k) \cdot \partial \hat{u}_k / \partial v_{hk}.$$

Заметим, что $\partial \hat{u}_k / \partial v_{hk} = \text{th}(zh)$, так как $\hat{u}_k = \sum_j \text{th}(zh) \cdot v_{hk}$. Подставив это в выражение для производной, получим

$$\partial E / \partial v_{hk} = -(u_k - \hat{u}_k) \cdot \text{th}(zh). \qquad (4.14)$$

Теперь можно перейти к производным второго уровня, по W:

$$\partial E / \partial w_{ij} = \sum_k [-(u_k - \hat{u}_k) \cdot \partial \hat{u}_k / \partial w_{ij}].$$

Подставляя сюда $\hat{u}_k = \sum_j \text{th}(\sum_i x_i \cdot w_{ij}) \cdot v_{jk}$, получаем:

$$\partial \hat{u}_k / \partial w_{ij} = v_{jk} \cdot \text{th}'(\sum_i x_i \cdot w_{ij}) \cdot x_i.$$

Производная $\text{th}'(z)$ выражается через $\text{th}(z)$, как объясняется в **Вопросе** 4.5 и формуле (4.17) далее. Окончательное выражение для этих производных имеет вид:

$$\partial E / \partial w_{ij} = -\sum_k [(u_k - \hat{u}_k) \cdot v_{jk}] \cdot (1 + \text{th}(z_j))(1 - th(z_j)) \cdot x_i / 2. \quad (4.15)$$

Уравнения (4.13), (4.14) и (4.15) приводят к следующим правилам обработки данных объект-за-объектом для пересчета матриц V и W применительно к нейронной сети на рис. 4.11.

1. **Прямое вычисление (выхода $\hat{u}$ и ошибки).** При заданных V и W, по получении информации (x, u) об объекте, вычисляется оценка $\hat{u}$ вектора u по формуле (4.8), после чего рассчитывается ошибка $e = u - \hat{u}$.

2. **Обратное распространение ошибки (для оценки элементов градиентов).** Каждый нейрон получает релевантную ему информацию, т. е.

$-e_k = -(u_k - \hat{u}_k)$, в (4.14) для выходных нейронов k ($k = III1, III2$);

$-\sum_k [(u_k - \hat{u}_k) \cdot v_{hk}]$, в (4.15) для нейронов скрытого слоя h ($j = II1, II2, II3$) (можно сказать, что на вход нейрона поступает сумма ошибок выходного слоя, взвешенных весами соответствующих соединений).

Эта информация используется для вычисления производных (4.14) и (4.15).

3. **Пересчет весовых коэффициентов.** Матрицы V и W пересчитываются по формуле (4.13).

Необходимо отметить, что все вычисления могут быть произведены локально каждым нейроном, используя только ту информацию, которая ему доступна. В частности, алгоритм не изменится, если количество нейронов в скрытом слое изменить с $h = 3$ на рис. 4.11, на любое другое значение h. Более того, алгоритм остается тот же и при изменении количества входов и выходов.

4.4.4.3. Вычислительная схема для метода обратного распространения ошибки

Прежде, чем применять метод, сделаем замечание о предварительном преобразовании данных. В машинном обучении обычно все признаки нормализуются так, чтобы минимальное значение было –1, а максимальное +1. Тем самым входные данные переводятся в формат выходных, определяемых функциями активации $\text{th}(x)$ и $\text{sign}(x)$, с их границами в –1 и 1. Для этого используется традиционная формула, включающая сдвиг и масштабирование: любой признак x_v может быть преобразован по формуле $y_v = (x_v - a_v)/b_v$,

где b_v равно половине размаха $b_v = (M_v - m_v)/2$, а сдвиг a_v, — середине размаха $a_v = (M_v + m_v)/2$. Здесь M_v обозначает максимум, а m_v — минимум признака v. Практика показывает, что машинные вычисления накапливают меньше ошибок, когда обрабатываемые числа находятся в интервале [–10, 10], что предполагает дальнейшее умножение всей таблицы данных на 10.

Применение метода градиентного спуска для обучения нейронных сетей осложняется тем, что выбор допустимых точек ограничен объектами, имеющимися в выборке. Именно наблюденные пары (x_i, u_i) используются на каждом шаге процесса, имитирующего скорейший спуск для изменения матриц V и W. Точнее говоря, при заданных V W наблюденные пары поступают одна за одной в случайном порядке. Для каждой измеряются ошибки между наблюденным u и вычисленным $\hat{u}$. После того, как все наблюденные пары обработаны, их порядок случайно меняется, и они вводятся снова, уже в измененном порядке. Этот повторный процесс называется эпохой. Матрицы V и W меняются либо после каждой пары (x_i, u_i), используя разности $\hat{u} - u$, как объяснялось, локально; либо же в конце каждой эпохи, используя накопленную ошибку.

Алгоритм обратного распространения ошибки с локальными изменениями матриц V и W можно сформулировать следующим образом.

Алгоритм обратного распространения ошибки

1. Инициализируйте весовые матрицы $W = (w_{ih})$ и $V = (v_{hk})$, заполняя их случайными числами согласно нормальному распределению $N(0, 1)$ с нулевым средним и единичной дисперсией.
2. Стандартизируйте данные к интервалу [–10, 10], как описано выше.
3. Сформулируйте критерий остановки процесса, как объяснено ниже, и применяйте алгоритм эпоха за эпохой, до достижения критерия остановки.
4. Для каждой эпохи рандомизируйте заново порядок предъявления объектов и применяйте нижеописанный алгоритм обратного распространения ошибки для изменения матриц V и W в этом порядке.
5. Если критерий остановки достигнут, заканчивайте вычисления и выдайте результаты: W, V, $\hat{u}$, e, and E. В противном случае, переходите к выполнению 4.

Критерий остановки процесса:

1) матрицы V и W стабилизировались. К сожалению, в реальных вычислениях этот критерий очень трудно достижим;

2) разница между средними значениями (по эпохе) функции ошибки становится меньше, чем заранее выбранное число, например, 0.0001;

3) количество эпох достигает заранее обусловленного порога, скажем, 5000.

Процедура обработки отдельного наблюдения

Особенности структуры и функции нейронной сети приводят к простым и эффективным правилам обработки индивидуальных наблюдений согласно процедуре скорейшего спуска. Перед тем как применить формулу (4.13) для пересчета весов, следует выполнить два следующих шага.

1. **Вычисление выходного сигнала сети и его ошибки.** Получив на вход данные об отдельном объекте, нейронная сеть перерабатывает входной сигнал, получая на выходе оценку выходного сигнала. Затем вычисляется ошибка как разница наблюденных и оцененных значений компонент выхода.

2. **Обратное распространение ошибки для оценки градиента.** Вычисленная ошибка выходного сигнала распространяется по сети «обратным ходом». Каждый нейрон выходного слоя соответствует какому-то выходному признаку и, значит, получает значение ошибки в этом признаке. Каждый нейрон скрытого слоя получает сумму этих ошибок, взвешенных весами соединений. Эти величины используются для корректировки элементов градиента, вычисляемого по функциям активации скрытых нейронов, как описано выше.

Программа nnn.m для MATLAB реализует алгоритм обратного распространения ошибки для обучения сети, представленной на рис. 4.11 (см. Приложение). Два параметра этого алгоритма, количество нейронов во внутреннем слое и скорость обучения, — входные параметры этой программы. Выход программы — минимальный уровень ошибки, достигнутый программой, и соответствующие матрицы V и W.

Эта программа включает в себя следующие шаги.

1. **Загрузка данных.** Предполагается, что матрица данных находится в папке Data. Это может быть либо iris.dat, либо stud.dat, либо любой другой подобный файл.

2. **Стандартизация данных.** Каждый столбец сдвигается в середину интервала значений и делится на половину размаха. Затем вся таблица данных умножается на 10. Это переводит все признаки к шкале [–10, 10], описанной выше.

3. **Подготовка подматриц входа и выхода для обучения.** Сначала принимается решение о том, какие признаки — входные (предиктивные), а какие — выходные (целевые). Для данных Ирисы в качестве целевых выбраны признаки лепестка (w3 и w4), а в качестве входных — измерения чашелистика (w1 и w2). Для данных «Студенты» в качестве целевых могут быть выбраны их оценки по всем трем предметам (CI, SP и OOP), а другие признаки (категории профессии, возраст и количество детей), в качестве предикторов.

4. **Инициализация сети.** Для этого надо: (a) выбрать количество нейронов скрытого слоя H, (b) заполнить матрицы W и V случайными значениями из нормального распределения N(0,1), и (c) назначить количество прогонов через эпохи со счетчиком, инициализированным нулем.

5. **Организация цикла по объектам множества данных.** Случайный порядок на объектах может быть организован с помощью команды MATLAB randperm(n), где n — количество объектов в обрабатываемом множестве.

6. **Проход вперед.** Для каждого данного объекта, при заданных матрицах V, W и функциях активации, вычисляются оцениваемые значения выходных признаков и их ошибки. В программе используется симметрическая сигмоида (гиперболический тангенс) как функция активации нейронов скрытого слоя.

7. **Обратное распространение ошибки.** Градиенты для матриц V и W вычисляются по формулам (4.14) и (4.15).

8. **Пересчет V и W.** Используя вычисленные значения градиентов и заданную скорость обучения, матрицы W и V обновляются по формуле (4.13).

9. **Условие остановки.** Оно использует как заданный уровень точности, скажем, 0.01, и порог на количество эпох, скажем, 5,000. Вычисления останавливаются, если какое-либо из них достигнуто.

Вопрос 4.5. Докажите, что производные сигмоиды (4.6) и гиперболического тангенса (4.7) могут быть представлены как простые полиномы от самих себя. Точнее,

$$s'(x)=((1+e^{-x})^{-1})'=(-1)(1+e^{-x})^{-2}(-1)e^{-x}=s(x)(1-s(x)), \quad (4.16)$$

$$\begin{aligned}\mathrm{th}'(x)&=[2s(x)-1]'=2\cdot s(x)'=2s(x)\cdot(1-s(x))=\\&=(1+\mathrm{th}(x))\cdot(1-\mathrm{th}(x))/2.\end{aligned} \quad (4.17)$$

Вопрос 4.6. Предложите способы улучшения сходимости процесса (4.13) — например, путем изменения величины шага.

Кстати говоря

6. Коррелирование (отыскание связей между признаками)

6.1. — Ты кем работаешь?
— Ландшафтным дизайнером!
— Ух, ты! На компьютере?
— Нет... На бульдозере...

6.2. — А как же ты понял, что этот медведь — людоед?
— По глазам. Взгляд тот же, что и у жены...

6.3. — Что-то наш баран сегодня грустный какой-то. Может, его зарежем?
— Ну если думаешь, что это его развеселит....

6.4. Дни рождения — вещь очень полезная. Как утверждает статистика, чем больше их у человека, тем дольше он живет.

6.5. По сообщениям Госкомстата, за последний месяц цена на бензин в среднем по стране упала на 0,7 процента. Объем литра уменьшился на 1,2 процента.

6.6. — Я дрессировала своего пса, чтобы он лаял, когда захочет есть. Сотни раз я показывала ему, как это надо делать.

— Ну и как? Лает он теперь, когда голоден?

— Наоборот, он теперь ничего не ест, пока я не начну лаять!

7. Логический вывод

7.1. Оказывается, у выражения «Если ты такой умный, то почему такой бедный?» есть строгое математическое обоснование.

Начнем с известных постулатов:

Постулат 1: Знание = сила

Постулат 2: Время = деньги

Все знают, что:

$$\text{путь} = \text{скорость} * \text{время} = \text{работа} / \text{сила},$$

откуда:

$$\text{работа}/\text{время} = \text{сила} \cdot \text{скорость}. \quad (1)$$

Подставив значение для времени и силы из обоих постулатов в (1), получим:

$$\text{работа}/(\text{знание} \cdot \text{скорость}) = \text{деньги}. \quad (2)$$

Из полученного равенства (2) видно, что устремляя знание или скорость к нулю, мы можем получить за любую работу сколь угодно большие деньги. Вывод: чем глупее и ленивее человек, тем больше денег он может заработать.

7.2. Мать сыну:

— Каждая твоя выходка — это ещё один седой волос на голове!

Мальчик, глядя на седую бабушку:

— Я смотрю ты в молодости тоже чудила помаленьку.

7.3. — Почему у слона глаза красные?

— Чтобы в помидорах мог прятаться.

— Видели когда-нибудь слона в помидорах?

— Хорошо прячется, да?

7.4. Физик: Почему у поезда колеса круглые, а когда он едет, они стучат?

Математик: — Это элементарно. Формула круга —пи эр квадрат, так вот этот квадрат как раз и стучит.

7.5. Стоит человек на дороге, навстречу ему едет автомобиль. Человек думает:

— Не трамвай, объедет.

Водитель думает:

— Не столб, отойдет!

7.6. Спускается профессор логики в лифте, лифт останавливается, человек, который хочет войти, спрашивает:

— Этот лифт едет вверх или вниз?

Профессор: — Да.

Тема 5
СУММАРИЗАЦИЯ ДАННЫХ

В данной теме будут рассмотрены основные методы суммаризации многомерных данных в количественном и качественном виде: метод главных компонент (МГК) и метод к-средних для кластерного анализа, а также вопросы их применения к реальным данным.

Проблема суммаризации, или агрегации, данных охватывает различные задачи, такие как измерение ненаблюдаемых факторов, построение кластеров, формирование аннотаций текстовых документов и пр. В отличие от задач коррелирования, признаки объектов здесь не разделяются на входные и выходные параметры некоторого процесса. Скорее можно полагать все имеющиеся признаки данных целевыми признаками, а выделенные агрегаты — кластеры или ненаблюдаемые признаки — «скрытыми» входными признаками (рис. 5.1, *а*).

Таким образом, проблема агрегации концептуально может рассматриваться как аналог проблемы коррелирования; только надо включить правило восстановления исходных данных из их агрегированного представления. Тогда исходные данные будут играть роль целевых, а восстановленные, «модельные» данные — роль предсказания. Такой подход приводит нас к необходимости установления не одного правила, как в задачах исследования корреляции, а двух — 1) правило агрегации данных (кодировщик, или «кодер»), 2) правило восстановления исходных данных из их агрегированного представления (расшифровщик, или «декодер»).

В отличие от проблемы коррелирования признаков, в проблеме агрегации правило порождения исходных данных должно быть специфицировано именно для восстановления исходных данных по их агрегированному представлению, а не прогноза новых данных. Именно поэтому мы говорим о расшифровщике (декодере), а не правиле предсказания. В литературе по машинному обучению проблеме агрегации не уделяется должного внимания. Поэтому зачастую проблему агрегирования понимают упрощенно, без дальнейшего восстановления данных, как это показано на рис. 5.1, *в*.

Такое осмысление структуры проблемы агрегации данных приводит к пониманию необходимости иметь обратную связь через

сравнение восстановленных данных с самими исходными данными, что делает задачу агрегации похожей на задачу коррелирования (см. рис. 5.1, *а* в сравнении с рис. 5.1, *б*). При этом расшифровщик выступает инструментом отображения пространства агрегированных данных в пространство исходных данных. Для сравнения «целевых» данных, полученных из агрегированного представления с помощью расшифровщика, с исходными данными в пространстве исходных данных проще всего использовать критерий минимизации разницы между ними. В данном учебнике рассмотрены методы именно такого рода (см. рис. 5.1, *а*)

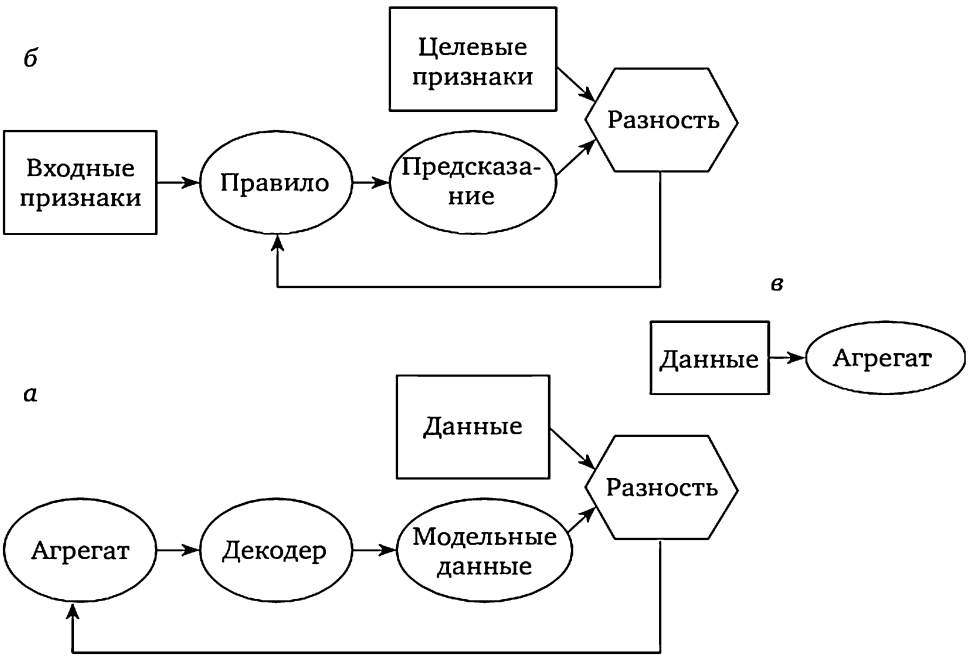


Рис. 5.1. **Диаграмма, описывающая задачу агрегации данных с расшифровщиком (*а*), задачу выявления корреляции между переменными (*б*), и задачу агрегации данных без декодера (*в*). Наблюдаемые данные представлены прямоугольниками, вычислительные структуры представлены овалами, сравнение наблюдений с предсказаниями обозначено шестиугольником.**

5.1. Метод главных компонент

Метод главных компонент (МГК) возник в связи с исследованиями «наследования таланта», проводившихся на стыке XIX и XX веков Френсисом Гальтоном (*F. Galton*, 1822—1911) и Карлом Пирсоном (*K. Pearson*, 1857—1936). В настоящее время это наиболее популярный метод суммаризации и визуализации данных. Математи-

ческая структура и свойства метода основаны на так называемом сингулярном разложении матриц (англ. *Singular Value Decomposition*, SVD). Поэтому в некоторых публикациях термины МГК и SVD используются как синонимичные. Однако в США и Великобритании термин МГК относится только к эвристической методологии анализа матрицы ковариаций или корреляций между признаками, которая эквивалентна модельному представлению МГК в случае, когда данные были предварительно центрированы. В этой теме мы излагаем оба варианта метода, а также рассказываем о двух направлениях применения МГК — для оценки скрытого фактора и для визуализации данных.

5.1.1. Модель и метод для измерения скрытого фактора

Рассмотрим матрицу данных X с элементами x_{iv} и стандартизуем ее в $Y = (y_{iv})$ ($i = 1, 2, \ldots, N$; $v = 1, 2, \ldots, V$), как обычно, путем вычитания некоего центрального значения с последующим делением на характеристику разброса. Модель метода главных компонент (МГК) предполагает, что значения признаков v на объектах i — это внешние проявления некоего скрытого фактора с значениями z_i^* (иногда их называют факторными баллами), которые определяются «нагрузками» на признаки c_v^*, так что произведение $z_i^* c_v^*$ является декодером величины y_{iv}, максимально ее аппроксимирующим. Иными словами, предполагается, что произведения $z_i^* c_v^*$ должны совпадать с y_{iv} с точностью до аддитивных невязок e_{iv},

$$y_{iv} = c_v^* z_i^* + e_{iv}, \tag{5.1}$$

которые должны минимизироваться через критерий суммарных квадратичных невязок

$$D^2 = \sum_{i \in I} \sum_{v \in V} (y_{iv} - c_v^* z_i^*)^2. \tag{5.2}$$

Декодер (5.1)—(5.2), как математическая модель для вычисления z_i^* и c_v^*, имеет существенный недостаток: решение не может быть определено единственным образом! Допустим, например, что мы каким-то образом получили значение фактора z_i^* для объекта i и нагрузку c_v^* на признак v, и можем вычислить произведение $z_i^* c_v^*$ как оценку значения признака на объекте. Поменяем теперь z_i^* и c_v^*: разделим первое на 2, а второе умножим на 2 — произведение не изменится: $z_i^* c_v^* = (z_i^* / 2)(2c_v^*)$. Любое другое число, взятое в качестве множителя — делителя, очевидно, приведет к тому же эффекту.

Прибегнем к традиционному способу избавления от множественности решений — зафиксируем нормы векторов z^* и c^*, взяв их равными, скажем, единице, а величину произведения выделим отдельно, обозначив, скажем, через $\mu > 0$. Теперь в (5.1) и (5.2)

вместо $z_i^* c_v^*$ будет стоять произведение $\mu z_i c_v$, где z и c — нормированные версии z^* и c^*, а μ — их общий эффект, произведение норм векторов z^* и c^*. (Евклидова) норма $\|x\|$ вектора $x = (x_1, \ldots, x_N)$ определяется как его длина, т. е. корень квадратный величины $\|x\|^2 = x^T x = x_1^2 + x_2^2 + \ldots + x_N^2$. При этом вектор называется нормированным, если его длина равна 1, так что $\|x\| = 1$. После того, как μ, z и c, минимизирующие (5.2), найдены, вернуться к вектору оценок способностей z^* и вектору нагрузок c^* можно с помощью формул: $z^* = \mu^{1/2} z$, $c^* = \mu^{1/2} c$. Любая несимметричная формула восстановления такая, как $z^* = \mu^a z$, $c^* = \mu^{1-a} c$, при $0 < a < 1$, работать не будет, т. е. a может быть только $a = 1/2$ (см. вопрос 2.18). Следует также заметить, что любое другое нормирование, скажем, такое, как $|x_1| + |x_2| + \ldots + |x_N| = 1$ будет вести к другому решению. Данное решение наиболее популярно, так как именно Евклидово нормирование ведет к элегантному сингулярному разложению (*Singular Value Decomposition*, SVD), рассматриваемому ниже.

Итак, надо найти μ, z и c, минимизирующие

$$D^2 = \sum_{i \in I} \sum_{v \in V} (y_{iv} - \mu c_v z_i)^2 \qquad (5.2')$$

при условиях, что $\|c\|^2 = 1$ и $\|z\|^2 = 1$. Как известно, для этого достаточно проанализировать необходимые условия оптимальности соответствующей функции Лагранжа:

$$L^2 = \sum_{i \in I} \sum_{v \in V} (y_{iv} - \mu c_v z_i)^2 + \alpha \left(1 - \sum_v c_v^2\right) + \beta \left(1 - \sum_i z_i^2\right),$$

получаемые приравниванием нулю соответствующих частных производных этой функции. Выпишем формулы для частных производных:

$$\frac{\partial L^2}{\partial \mu} = -2 \sum_{i \in I} \sum_{v \in V} (y_{iv} - \mu c_v z_i) c_v z_i,$$

$$\frac{\partial L^2}{\partial c_v} = -2 \sum_{i \in I} (y_{iv} - \mu c_v z_i) \mu z_i - 2\alpha c_v,$$

$$\frac{\partial L^2}{\partial z_i} = -2 \sum_{v \in V} (y_{iv} - \mu c_v z_i) \mu c_v - 2\beta z_i.$$

Приравнивая эти выражения нулю, мы первым делом можем определить числовые величины μ, α и β. Первое выражение дает $\sum_{i \in I} \sum_{v \in V} y_{iv} c_v z_i - \sum_v c_v^2 \sum_i z_i^2 = 0$. В силу нормированности c и z получаем выражение для μ:

$$\mu = \sum_{i \in I} \sum_{v \in V} y_{iv} c_v z_i = z^T Y c. \qquad (*)$$

Теперь нетрудно доказать, что $\alpha = \beta = 0$.

Действительно, приравнивая нулю второе выражение, получаем:

$$\alpha c_v = -\sum_{i \in I}(y_{iv} - \mu c_v z_i)\mu z_i = -\mu\ (\sum_{i \in I} y_{iv} z_i - \mu c_v).$$

Умножим обе части полученного уравнения на c_v и суммируем результаты по v.

$$L^2 = \sum_{i \in I}\sum_{v \in V}(y_{iv} - c_v z_i)^2 + \alpha\left(1 - \sum_v c_v^2\right) + \beta\left(1 - \sum_i z_i^2\right),$$

получаемые приравниванием нулю соответствующих частных производных этой функции. Выпишем формулы для частных производных:

$$\frac{\partial L^2}{\partial \mu} = -2\sum_{i \in I}\sum_{v \in V}(y_{iv} - c_v z_i)c_v z_i,$$

$$\frac{\partial L^2}{\partial c_v} = -2\sum_{i \in I}(y_{iv} - c_v z_i)\mu z_i - 2\alpha c_v,$$

$$\frac{\partial L^2}{\partial z_i} = -2\sum_{v \in V}(y_{iv} - c_v z_i)\mu c_v - 2\beta z_i.$$

Приравнивая эти выражения нулю, мы первым делом можем определить числовые величины μ, α и β. Первое выражение дает

$$\sum_{i \in I}\sum_{v \in V} y_{iv} c_v z_i - \sum_v c_v^2 \sum_i z_i^2 = 0.$$

В силу нормированности c и z получаем выражение для μ:

$$\mu = \sum_{i \in I}\sum_{v \in V} y_{iv} c_v z_i = z^T Y c. \qquad (*)$$

Теперь нетрудно доказать, что $\alpha = \beta = 0$.

Действительно, приравнивая нулю второе выражение, получаем:

$$\alpha c_v = -\sum_{i \in I}(y_{iv} - c_v z_i)\mu z_i = -\mu\ (\sum_{i \in I} y_{iv} z_i - c_v).$$

Умножим обе части полученного уравнения на c_v и суммируем результаты по v.

С учетом того, что $\sum_v c_v^2 = 1$, получим

$$\alpha = -\mu \sum_{i \in I, v \in V} y_{iv} z_i c_v - \mu^2 = 0.$$

Последнее равенство вытекает из (*). Равенство $\beta = 0$ доказывается аналогично.

С учетом того, что $\alpha = \beta = 0$, приравнивание $\frac{\partial L^2}{\partial c_v}$ и $\frac{\partial L^2}{\partial z_i}$ нулю приводит, при $\mu \neq 0$, к равенствам

$$\sum_{i \in I} y_{iv} z_i = \mu c_v \text{ и } \sum_{v \in V} y_{iv} c_v = \mu z_i,$$

или, на языке матриц,

$$Y^T z = \mu c \text{ и } Yc = \mu z. \tag{5.3}$$

Как известно из линейной алгебры, уравнения вида (5.3) означают, что число μ и векторы c, z образуют так называемую сингулярную тройку матрицы Y. Если тройка (μ, c, z), удовлетворяет уравнениям (5.3), то число μ называется сингулярным значением матрицы Y, а z, c — сингулярными векторами матрицы Y, соответствующими сингулярному значению μ. Сингулярные тройки матрицы Y тесно связаны с собственными векторами соответствующих квадратных матриц $A = YY^T$ и $B = Y^TY$. А именно, имеет место следующее свойство.

Свойство 1. Тройка (μ, c, z) сингулярная для матрицы Y тогда и только тогда, когда вектор z является собственным вектором матрицы $A = YY^T$, соответствующим ее собственному числу μ^2, а вектор c — собственным вектором матрицы $B = Y^TY$, соответствующим тому же собственному числу μ^2.

Прежде, чем доказывать это свойство, напомним, что это такое — собственные значения квадратных матриц.

Вектор a называется собственным для квадратной матрицы G, если $Ga = \lambda a$ для некоторого, возможно, комплексного числа λ, которое называется собственным значением G, соответствующим собственному вектору a. Для случая матриц A и B, в некотором роде аналогичных числовым квадратам, нетрудно доказать, что собственные числа не только вещественны, но и неотрицательны. Количество ненулевых собственных значений G равно рангу G, причем собственные векторы, соответствующие различающимся собственным значениям, взаимно ортогональны. В анализе данных собственные значения предполагаются различающимися, потому что вероятность их совпадения на эмпирически полученной таблице — нулевая.

Вернемся к доказательству Свойства 1. Действительно, если (μ, c, z) сингулярная тройка для матрицы Y, выразим из первого уравнения в (5.3), $c = Y^Tz/\mu$ и подставим его во второе уравнение в (5.3). Получим $YY^Tz/\mu = \mu z$, так что $YY^Tz = \mu^2 z$, т. е., действительно, μ^2 — собственное число, а z — соответствующий собственный вектор матрицы $A = YY^T$. Аналогично доказывается, что μ^2 — собственное число, а c — соот-

ветствующий собственный вектор матрицы $B = Y^TY$. Пусть теперь, наоборот, z — собственный вектор матрицы $A = YY^T$, соответствующий ее собственному числу λ. Это значит, что справедливо равенство $YY^Tz = \lambda z$. Докажем, что $\mu = \lambda^{1/2}$ — сингулярное число матрицы Y, соответствующее сингулярным векторам z и $c = Y^Tz/\mu$. Действительно, равенство $Y^Tz = \mu c$ вытекает из определения c. Так же по определению c имеем $Yc = YY^Tz/\mu = \mu z/\mu$, так как z — собственный вектор матрицы YY^T. Но это и значит, что (μ, c, z) — сингулярная тройка для матрицы Y. Утверждение доказано.

Свойство 1 показывает, что известные свойства собственных чисел и векторов можно без особых изменений перенести на сингулярные числа и векторы. Прежде всего, их количество конечно и равно рангу матрицы[1], который можно принять равным минимуму числа ее строк и столбцов, поскольку это — матрица наблюденных данных, которую трудно заподозрить в специальной «заботе» о наличии внутренней линейной связи.

Какому же из сингулярных чисел соответствует решение задачи на минимум D^2 в (5.2′)? Раскроем скобки:

$$D^2 = \sum_{i\in I}\sum_{v\in V} y_{iv}^2 - 2\mu\sum_{i\in I}\sum_{v\in V} y_{iv}z_ic_v + \mu^2\sum_{i\in I} z_i^2\sum_{v\in V} c_v^2.$$

С учетом нормированности векторов, а также равенства (*), получаем:

$$D^2 = \sum_{i\in I}\sum_{v\in V} y_{iv}^2 - \mu^2 \qquad (5.2'')$$

Отсюда ясно, что минимум D^2 достигается на максимальном сингулярном значении и соответствующих сингулярных векторах матрицы Y. Это решение и называется главной компонентой.

Мы доказали следующее утверждение.

Свойство 2. Оптимальное решение задачи (5.1)—(5.2) достигается на сингулярной тройке (μ, c, z) матрицы Y, соответствующей ее максимальному сингулярному значению. Эта главная компонента удовлетворяет уравнениям $z^* = \mu^{1/2}z$ и $c^* = \mu^{1/2}c$, так что, в силу (5.3), вектор факторных баллов z^* является взвешенной комбинацией столбцов матрицы Y, причем веса задаются компонентами вектора нагрузок c^*. Аналогично, вектор нагрузок c^* — не что иное, как взвешенная комбинация строк матрицы Y, в которой веса определяются компонентами вектора z^*.

В процессе доказательства мы вывели еще одно свойство главной компоненты, заслуживающее отдельной формулировки.

---

[1] Ранг матрицы в линейной алгебре — это максимальное количество ее линейно независимых строк.

Свойство 3. Имеет место пифагоровское разложение квадратичного разброса данных $T(Y) = \sum_{i,v} y_{iv}^2$, связывающее критерий наименьших квадратов D^2 и максимальное сингулярное значение:

$$T(Y) = \mu^2 + D^2. \tag{5.4}$$

Разложение (5.4) показывает, что μ^2 выражает ту часть квадратичного разброса данных, которая «объяснена» найденной главной компонентой.

Почему выражение $T(Y) = \sum_{i,v} y_{iv}^2$ называется «квадратичным разбросом» данных? Потому, что $T(Y)$ — это сумма выражений $dd(i) = \sum_{v} y_{iv}^2$ по всем объектам i, а каждое $dd(i)$ — не что иное, как квадрат Евклидова расстояния от объекта $y(i) = (y_{i1}, y_{i2}, \ldots, y_{iV})$ до точки $0 = (0, 0, \ldots, 0)$ начала координат. То есть $T(Y)$ — это сумма квадратов длин отрезков, соединяющих точки y_i, $i \in I$, данного множества с началом координат (рис. 5.2).

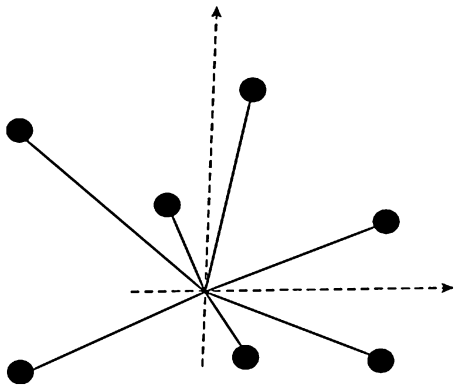


Рис. 5.2. **Разброс данных — сумма квадратов длин отрезков, соединяющих данные точки с нулем**

Следует отметить, что выражения (5.3) позволяют использовать полученную главную компоненту как теоретический конструкт, дающий возможность включать в анализ объекты и признаки, не участвовавшие в расчете главной компоненты. Например, новый объект $yn = (yn_1, yn_2, \ldots, yn_V)$ получает значение фактора (факторный балл) $z^*(yn) = \langle c^*, yn \rangle / \mu$ согласно второму уравнению в (5.3). Аналогично, дополнительный признак, представленный N-мерным вектором y значений на объектах, стандартизованный так же, как и другие признаки, получит величину нагрузки $c^*(y) = \langle z^*, y \rangle / \mu$. В этом смысле модель (5.1) может рассматриваться как «генеративная», т. е. постулирующая природу данных как тех, что порождены в соответствии с моделью (5.1), с точностью до невязок

5.1.2. Метод главных компонент (МГК): случай нескольких скрытых факторов

Модель (5.1) предполагает, что все $N \times V$ чисел в матрице данных Y можно восстановить с помощью $N + V$ значений в векторах z^* и c^*. Например, при $N = 1000$, $V = 10$ в матрице Y 10 000 наблюденных чисел, а в векторах — всего 1010 значений. Поэтому обычно не получается аппроксимировать данные моделью (5.1) единственного фактора с достаточной точностью. Естественно желание увеличить количество аппроксимирующих факторов до двух, трех или даже K штук. Будем считать, что существует относительно небольшое число K скрытых факторов z_k^* и соответствующих векторов нагрузок на признаки c_k^* ($k = 1, 2, \ldots, K$; $K < V$), так что стандартизованная матрица данных представляет собой сумму факторных баллов, взвешенных соответствующими нагрузками.

Это можно выразить нижеследующими уравнениями декодера (5.5) для матрицы данных $Y = (y_{iv})$, предполагающими, что векторы факторных баллов z_k^* и векторы нагрузок c_k^* могут быть найдены путем минимизации невязок e_{iv}. Чтобы избавиться от математической неопределенности, опять считаем, что $z_k^* = \mu^{1/2} z_k$ и $c_k^* = \mu^{1/2} c_k$, где z_k и c_k — нормированные векторы.

Тогда уравнения (5.5) будут иметь вид:

$$y_{iv} = \sum_{k=1}^{K} \mu_k c_{kv} z_{ik} + e_{iv}. \qquad (5.5')$$

Обозначим ранг матрицы Y через r. Будем считать, что $K < r$. Отсортируем сингулярные числа матрицы Y в порядке убывания: $\mu_1 > \mu_2 > \ldots > \mu_r > 0$. Так как матрица получена прямым наблюдением, мы принимаем, что вероятность совпадения каких-то сингулярных значений — нулевая. Можно доказать, что критерий суммы квадратов невязок в (5.5′) минимизируется первыми K сингулярными тройками

$$y_{iv} = \sum_{k=1}^{r} \mu_k c_{kv} z_{ik}, \qquad (5.6)$$

т. е. максимальными сингулярными значениями μ_k и соответствующими нормированными версиями сингулярных векторов z_k и c_k ($k = 1, 2, \ldots, K$).

Это объясняется тем математическим фактом, что любая прямоугольная матрица Y разлагается в сумму слагаемых (5.6), определяемых ее сингулярными тройками. Равенство (5.6) называется сингулярным разложением матрицы Y, а на английском языке – *Singular Value Decomposition* (SVD). Переведенное на язык матриц, разложение (5.6) имеет вид

$$Y = \sum_{k=1}^{r} \mu_k z_k c_k^T = ZMC^T, \qquad (5.6')$$

где Z это матрица размера $N \times r$ со столбцами z_k, C — матрица размера $V \times r$ со столбцами c_k, а M — диагональная матрица размера $r \times r$ со значениями μ_k на главной диагонали и нулями — вне ее[1].

Напомним, что сингулярные векторы одновременно являются и собственными для производных матриц Y^TY и YY^T. Поэтому они взаимно ортогональны. Отсюда вытекает, с учетом (5.6), что квадратичный разброс $T(Y) = \sum y_{iv}^2$ матрицы Y разлагается в сумму квадратов ее сингулярных значений:

$$T(Y) = \mu_1^2 + \mu_2^2 + \ldots + \mu_r^2. \qquad (5.7)$$

Отсюда вытекает, что решение по методу главных компонент в уравнении (5.5) приводит к разложению квадратичного разброса данных на вклады отдельных компонент и минимизируемый критерий наименьших квадратов $D^2 = \sum_{i,v} e_{iv}^2$ (пифагоровское разложение):

$$T(Y) = \mu_1^2 + \mu_2^2 + \ldots + \mu_K^2 + D^2. \qquad (5.8)$$

Следовательно, относительный вклад решения по методу главных компонент в разброс равен $(\mu_1^2 + \mu_2^2 + \ldots + \mu_K^2) / T(Y)$.

Рис. 5.3. **Геометрическая иллюстрация Пифагоровской связи между наблюденными данными (звездочка), рассчитанными «модельными» данными (квадрат) и невязками (соединяющий отрезок)**

Необходимо отметить, что разложение квадратичного разброса данных (5.8) на две части, одна — объясненная, другая — не объясненная моделью, так же, как и его частный случай (5.4), может рассматриваться как многомерная реализация известной теоремы Пифагора. Пифагор, полулегендарный древнегреческий философ, живший в VI-м веке до нашей эры, считается одним из создателей научной картины мира, включающей сам термин «философия» и понятие математического доказательства, а также глубокие параллели между числами, музыкой и космосом, лежащие в основе

[1] В математике разложение (5.6) обычно записывается как $Y = V\Sigma W^T$, где V и W — матрицы сингулярных векторов, а Σ — диагональная матрица сингулярных значений σ_k, подчеркивая тем самым глубокое сходство правых и левых сингулярных векторов и намекая на первую букву в слове «сингулярный». Наши обозначения подчеркивают глубокое сходство между факторным и кластерным анализом, а буква «сигма» в науках о данных, увы, уже давно «зарезервирована» для матрицы ковариаций.

современного классического образования. Теорема Пифагора, как известно, утверждает, что в любом прямоугольном треугольнике квадрат длины гипотенузы равен сумме квадратов длин катетов. На рис. 5.3, иллюстрирующем (5.8), квадратичный разброс данных представлен гипотенузой, тогда как катеты олицетворяют суммарную квадратичную невязку D^2 и сумму вкладов отдельных компонент — квадратов соответствующих сингулярных значений.

5.1.3. Традиционная формулировка МГК через ковариационную матрицу

В англоязычной литературе до последнего времени доминировал совсем другой способ определения МГК — не через простецкие кодер-декодер «модельки» (5.1) и (5.2), а через эвристическое формирование скрытого фактора как такой линейной комбинации наблюденных признаков, которая дает наибольший вклад в дисперсию данных, для многих более убедительное.

Существенную роль в этом традиционном подходе играет так-называемая матрица ковариаций. Эта матрица определяется как квадратная матрица размера $V \times V$ (по числу признаков) $C = Y^TY/N$, где Y — центрированная версия исходной матрицы данных X, в которой каждый столбец центрирован, т. е. его среднее значение вычтено из всех элементов. Каждый (v', v'')-й элемент матрицы C — это коэффициент ковариации между признаками v' и v''; так что ее диагональные элементы — не что иное как дисперсии соответствующих признаков. Матрица ковариации является матрицей корреляции в том случае, если данные в матрице Y стандартизованы преобразованием z-скоринг, т. е. если после сдвига вычитанием среднего столбец признака был нормализован делением на его стандартное отклонение. Напомним, что произведение матрицы на свой транспонированный вариант, Y^TY, используется для оценки ковариационной матрицы нормального распределения, функция плотности которого имеет вид (3.10). Для этой цели данное произведение делится не на N, а на $N - 1$, чтобы получить несмещенную оценку. Как хорошо известно, подставляя арифметическое среднее вместо математического ожидания, мы вносим в данные дополнительную связь, что уменьшает на 1 количество степеней свободы в данных. Поскольку вычисления по методу главных компонент не имеют никакого отношения к цели оценки параметров Гауссова распределения, мы оставляем N в C в качестве делителя.

Обычная формулировка МГК при традиционном подходе имеет следующий вид. При заданной центрированной матрице данных Y размера $N \times V$, найти нормированный V-мерный вектор $c = (c_v)$ такой, что сумма столбцов матрицы Y, взвешенных элементами вектора c, $f = Yc$, имеет максимальную дисперсию. Этот вектор и есть главная компонента, так названная потому, что показывает направление максимальной дисперсии в данных.

Проанализируем эту формулировку. Вектор f центрирован при любом c, поскольку все столбцы Y центрированы. Следовательно, его дисперсия равна $s^2 = \langle f, f \rangle / N = f^T f / N$. Последнее равенство определяется соглашением, что V-мерный вектор — это матрица размера $V \times 1$, т. е. столбец. Подставив Yc вместо f, получаем $s^2 = c^T Y^T Yc / N$. Задача максимизация этой величины по всевозможным нормированным, т. е. удовлетворяющим условию $c^T c = 1$, векторам c эквивалентна задаче максимизации отношения

$$q(c) = \frac{c^T Y^T Yc}{c^T c}. \tag{5.9}$$

по всевозможным V-мерным c. Выражение (5.9) хорошо известно в линейной алгебре как отношение Рэлея (*Rayleigh quotient*) для матрицы $Y^T Y$. Доказано, что его максимум $q(c^*)$ равен максимальному собственному числу матрицы $Y^T Y$ и достигается на соответствующем собственном векторе c^*. Остается заметить, что матрица $Y^T Y$ пропорциональна матрице ковариации $A = Y^T Y / N$. Таким образом, максимум отношения Рэлея (5.9) достигается на векторе c^*, являющемся собственным вектором матрицы A, соответствующим ее максимальному собственному значению $q(c^*)/N$.

Мы показали, что первая главная компонента по традиционному подходу — это вектор $f = Yc$, где c — это нормированный собственный вектор матрицы ковариации A, соответствующий ее максимальному собственному значению. Вторая главная компонента в традиционном подходе определяется как такая линейная комбинация столбцов матрицы Y, которая максимизирует дисперсию при условии, что эта комбинация ортогональная первой главной компоненте. Очевидно, вторая главная компонента определяется вторым по величине собственным значением матрицы A и соответствующим собственным вектором. Следующие главные компоненты определяются аналогично — это такие линейные комбинации признаков, которые ортогональны всем предыдущим главным компонентам и максимизируют дисперсию при этом условии. По свойствам собственных чисел и векторов, каждая главная компонента определяется соответствующими собственными значением и вектором.

Данное построение на первый взгляд кажется никак не связанным с вышерассмотренной моделью кодер-декодер для МГК. На самом же деле два определения вычислительно эквивалентны. Рассмотрим уравнение $Yc = \mu z$ из (5.3) и выразим z как $z = Yc/\mu$, после чего подставим это z во второе уравнение (5.3): $Y^T z = \mu c$. Получаем $Y^T Yc/\mu = \mu c$. Значит, μ^2 и c, определяемые мультипликативной моделью кодер-декодер, удовлетворяют уравнению

$$Y^T Yc = \mu^2 c, \tag{5.10}$$

так что c — собственный вектор квадратной матрицы Y^TY, соответствующий ее максимальному собственному значению $\lambda = \mu^2$. В случае, когда Y центрирована, Y^TY совпадает с матрицей ковариации A с точностью до постоянного сомножителя $1/N$. То есть c в (5.10) — собственный вектор матрицы A, соответствующий ее максимальному собственному значению. Мы доказали, что два определения эквивалентны в случае, когда матрица данных Y центрирована. При этом установлено простое соответствие между максимальным собственным значением λ матрицы A и сингулярным значением μ матрицы Y: $\lambda = \mu^2$.

Вычислительная эквивалентность сопровождается существенными концептуальными различиями между двумя подходами. Первый подход основан на модели, тогда как второй намеренно эвристический. Второй подход имеет смысл только для предварительно центрированных данных, тогда как модельный допускает любое предварительное преобразование данных. Более того, тот факт, что скрытые факторы являются линейными комбинациями признаков, не предполагается, но выводится из свойств оптимальности решения модели. Этот факт, напротив, постулируется при традиционном подходе. Табл. 5.1 подытоживает концептуальные различия двух подходов.

Таблица 5.1

Основные отличия между двумя эквивалентными версиями МГК

Особенности МГК	Модельный подход	Традиционный подход
Отношение к данным	Модельное	Эвристическое
Линейная комбинация признаков	Выводится	Постулируется
Стандартизация данных	Любая	Центрирование
Оценка вклада в разброс данных	Сумма квадратов сингулярных значений	Нет
Факторные баллы	Масштабированные сингулярные векторы	Нормированные собственные векторы
Соответствие интуиции	Не всегда	Хорошее

5.1.4. Применения МГК

5.1.4.1. Измерение внутреннего фактора

Рабочий пример 5.1

Измерение размера цветков Ириса по данным табл. 1.9

Признаки длины и ширины чашелистика и лепестка в табл. 1.9 можно рассматривать как «внешние» характеристики неизмеримого признака

«размер цветка». Применим МГК для суммаризации этих признаков в фактор «размер». Прежде всего, стандартизуем данные.

Таблица 5.2

Данные о трех представителях каждого таксона данных Ирис до и после стандартизации

Объекты	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп
	До стандартизации			
1	5.10	3.50	1.40	0.30
2	4.40	3.20	1.30	0.20
3	4.40	30	1.30	0.20
51	6.40	3.20	4.50	1.50
52	5.50	2.40	3.80	1.10
53	5.70	2.90	4.20	1.30
191	6.30	3.30	6.00	2.50
192	6.70	3.30	5.70	2.10
103	7.20	3.60	6.10	2.50
Мин	4.30	2.00	1.00	0.10
Размах	3.60	2.40	5.90	2.4
После стандартизации				
1	22.222	62.5	6.7797	8.3333
2	2.778	50	5.0847	4.1667
3	2.778	41.67	5.0847	4.1667
51	58.333	50.00	59.3220	58.3333
52	33.333	16.67	47.4576	41.6667
53	38.889	37.50	54.2373	50.0000
191	55.556	54.17	84.7458	100.0000
192	66.667	54.17	79.6610	83.3333
103	80.556	66.67	86.4407	100.0000
Мин	0	0	0	0
Размах	100	100	100	100

Для этого преобразуем шкалы всех признаков так, чтобы они менялись от 0 до 100. При этом минимальное значение признака переводится в 0, а максимальное — в 100. Эта стандартизация определяется формулой

$$y_{iv} = 100\frac{x_{iv} - \min_i x_{iv}}{\max_i x_{iv} - \min_i x_{iv}}, \qquad (5.11)$$

где в знаменателе стоит, очевидно, размах признака v. Такое преобразование помогает объективировать ранжирование данных по агрегированному фактору.

В верхней части табл. 5.2 представлены 9 объектов матрицы данных Ирис, по три из каждого таксона, а в нижней части — те же объекты в стандартизованном по формуле (5.11) виде.

Первый сингулярный вектор нагрузок на признаки имеет вид

$$c_1 = \begin{pmatrix} -0.4866 \\ -0.3910 \\ -0.5531 \\ -0.5517 \end{pmatrix},$$

что выглядит странновато: все его компоненты отрицательны, а это противоречит интуиции. В чем? Да в том, что каждый МГК фактор — не что иное как сумма признаков столбцов матрицы Y, взвешенных этими компонентами. Веса не должны быть отрицательны!

Беде легко помочь, если вспомнить, что одновременно с сингулярной тройкой (μ, c, z) является сингулярной и тройка $(\mu, -c, -z)$. Справедливость этого утверждения легко проверить, просто умножив уравнения (5.3) на -1. То есть достаточно в качестве сингулярных векторов взять столбцы вычисленных матриц факторных баллов и нагрузок, умноженные на -1. В данном случае приходим к равенству

$$c_1 = \begin{pmatrix} 0.4866 \\ 0.3910 \\ 0.5531 \\ 0.5517 \end{pmatrix}.$$

Эти значения можно использовать для интерпретации полученного «синтетического» фактора. Он пропорционален взвешенной сумме

$$Z = \alpha \cdot (0.4866 \cdot SL + 0.3910 \cdot SW + 0.5531 \cdot PL + 0.5517 \cdot PW). \quad (5.12)$$

Величина α характеризует масштаб шкалы. Ее следует выбирать из внешних соображений. Требование нормированности здесь не имеет никакого смысла и должно быть заменено каким-то другим принципом.

Как видим, веса вполне положительны — факт их участия в «размере» налицо. Веса признаков лепестка почти одинаковы и максимальны, а вот признак ширины чашелистика «проседает», его вес, 0.391, относительно мал, что соответствует вышеотмеченным его особенностям (негативная корреляция с другими признаками).

Для определения качества фактора обратимся к его вкладу, пропорциональному μ_1^2. Сингулярные значения стандартизованной матрицы: 1206, 384.6, 122.9, 60. Отношение квадрата первого сингулярного значения к сумме квадратов всех сингулярных значений равно 89.72 % — вполне внушительная доля квадратичного разброса, объясненная построенным фактором.

Остается определиться с масштабом шкалы. Данный автор придерживается принципа, что шкала измерения фактора должна соответствовать шкалам использованных признаков, т. е. меняться от 0 до 100. Очевидно, что если все исходные признаки, *SL*, *SW*, *PL*, *PW*, равны нулю, то значение Z по формуле (5.12) тоже равно 0. Потребуем, чтобы значение фактора на «идеальных» объектах, имеющих оценку 100 по всем признакам, тоже равнялось 100. Применяя это к уравнению (5.12), получим

$$100 = \alpha \cdot (0.4866 \cdot 100 + 0.3910 \cdot 100 + 0.5531 \cdot 100 + 0.5517 \cdot 100),$$

что дает $1 = \alpha \cdot (0.4866 + 0.3910 + 0.5531 + 0.5517)$, т. е. сумма весов признаков должна равняться единице.

Таблица 5.3

Значения фактора «Размер цветка» на представителях таксонов Ирис

№	9 объектов Ириса после стандартизации. Веса признаков				Фактор размера
	0.2455	0.1972	0.279	0.2783	
1	22.222	62.5	6.7797	8.3333	21.9926
2	2.778	50	5.0847	4.1667	13.1217
3	2.778	41.67	5.0847	4.1667	11.4781
51	58.333	50	59.322	58.3333	56.9656
52	33.333	16.67	47.4576	41.6667	36.3061
53	38.889	37.5	54.2373	50	45.9894
191	55.556	54.17	84.7458	100	75.794
192	66.667	54.17	79.661	83.3333	72.4648
103	80.556	66.67	86.4407	100	84.8691

Математически получаем: $1 = 1.9824 \cdot \alpha$. Отсюда $\alpha = 0.5044$. Получаемые веса признаков представлены в табл. 5.3.

В ее правом столбце содержатся значения вычисленного нами фактора «Размер цветка». Как видно, таксоны, определенные на основе генетической информации, определяются также и размером: от малых цветков в первом таксоне через средний размер во втором таксоне до большого размера в третьем таксоне.

Самостоятельная работа

5.1. Примените метод главных компонент к построению скрытого фактора «уровень развития города» по таблице 1.11, включая все основные этапы:

— «ранжировочную» стандартизацию данных;

— вычисление первой сингулярной тройки и ее вклада в разброс данных;

— интерпретацию главной компоненты;

— шкалирование главной компоненты, при котором значение внутреннего фактора равно 100 на объектах, имеющих значение 100 по каждому использованному признаку.

5.1.4.2. Визуализация данных

Визуализация данных по методу главных компонент — это проекция данных на плоскость двух первых сингулярных векторов (z_1, z_2). Эта плоскость лучше всего аппроксимирует данные в смысле модели (5.5). Каждый i-й объект получает координаты (z_{i1}, z_{i2}). При этом данные аппроксимируются моделью (5.5) при $K = 2$:

$$y_{iv}^* \approx z_{i1}^* c_{1v}^* + z_{i2}^* c_{2v}^*. \qquad (5.6'')$$

Это уравнение отражает долю $100 \cdot (\mu_1^2 + \mu_2^2) / T(Y)\,\%$ разброса данных.

Отметим, что данная визуализация позволяет не только отобразить объекты — точками (z_{i1}^*, z_{i2}^*), но и признаки — пары (c_{1v}^*, c_{2v}^*), представляемые не только точками плоскости, но и отрезками, соединяющими их с точкой начала отсчета 0. Смысл отрезков в том, что проекции точек на такой отрезок отражают значения соответствующего признака на объектах как по длине, так и знаку.

Для получения адекватного портрета данных при визуализации данные следует вначале центрировать, так чтобы значения признаков рассматривались не сами по себе, а в сравнении с общим средним. После этого структура данных просматривается значительно лучше, чем до вычитания средних. Этот эффект хорошо иллюстрируется рис. 5.4. На этом рисунке представлено облако данных до вычитания средних (вверху) и после этого (внизу). Поскольку главная компонента обязана проходить через 0, в первом случае она оказывается длинной, а во втором — короткой. Соответственно меняется ее вклад в разброс данных: большой в первом случае и маленький — во втором.

Рабочий пример 5.2

Визуализация данных Ирис

Опишем применение МГК к визуализации таблицы данных Ирис (табл. 1.9) размера 150 × 4. Напомним, что прежде всего таблицу следует центрировать — вычесть из элементов каждого столбца егосреднее.

Для выравнивания вкладов столбцов, разделим каждый из них на его стандартное отклонение (средние и стандартные отклонения приведены в верхней половине табл. 4.12), т. е. осуществим стандартизацию данных по методу *z*-скоринга.

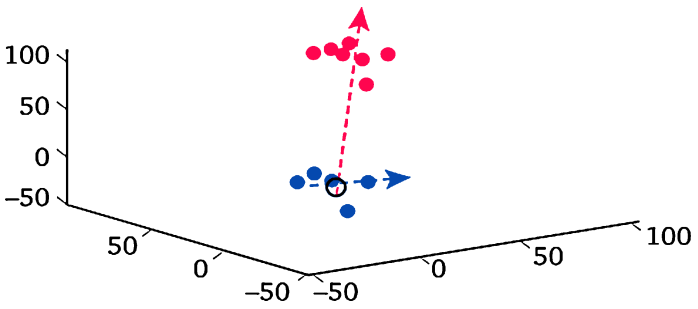


Рис. 5.4. **Главная компонента при разных стандартизациях:** облако точек до центрирования (вверху) и после центрирования (внизу). Вверху длина и вклад главной компоненты большие, внизу — маленькие

Таблица 5.4

Данные о трех представителях каждого таксона данных Ирис до и после стандартизации по методу *z*-скоринг

	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп
До стандартизации				
1	5.10	3.50	1.40	0.30
2	4.40	3.20	1.30	0.20
3	4.40	3.00	1.30	0.20
51	6.40	3.20	4.50	1.50
52	5.50	2.40	3.80	1.10
53	5.70	2.90	4.20	1.30
191	6.30	3.30	6.00	2.50
192	6.70	3.30	5.70	2.10
103	7.20	3.60	6.10	2.50
Среднее	5.84	3.06	3.76	1.20
Ст.откл.	0.83	0.44	1.77	0.76
После стандартизации				
1	–0.8977	1.0156	–1.3358	–1.1799
2	–1.7430	0.3273	–1.3924	–1.3111
3	–1.7430	–0.1315	–1.3924	–1.3111
51	0.6722	0.3273	0.4203	0.3945

Окончание табл. 5.4

	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп
52	–0.4146	–1.5081	0.0238	–0.1303
53	–0.1731	–0.3610	0.2504	0.1321
191	0.5515	0.5567	1.27	1.7064
192	1.0345	0.5567	1.1001	1.1816
103	1.6384	1.245	1.3267	1.7064
Среднее	0	0	0	0
Ст. откл.	1	1	1	1

Могут сказать, что нормализация необязательна, так как все признаки измерены в одной и той же шкале и выражают соизмеримые элементы цветка. Предлагаем читателю провести визуализацию на центрированной матрице Ирис без деления столбцов на стандартные отклонения самостоятельно.

Образец данных Ирис после z-стандартизации приводится в нижней части табл. 5.4.

Получаемая матрица, чьи столбцы — нормированные векторы нагрузок:

$$c = \begin{matrix} 0.5211 & 0.3774 & -0.7196 & -0.2613 \\ -0.2693 & 0.9233 & 0.2444 & 0.1235 \\ 0.5804 & 0.0245 & 0.1421 & 0.8014 \\ 0.5649 & 0.0669 & 0.6343 & -0.5236 \end{matrix}.$$

Четверка сингулярных значений матрицы Y: μ = (20.85, 11.67, 4.68, 1.76); ее квадратичный разброс: 596. Вклады первых двух компонент равны $20.85^2/596 = 72.96$ % и $11.67^2/596 = 22.85$ %, давая суммарный вклад 95.81 % — вполне комфортабельный уровень точности визуализации.

Осуществим интерпретацию факторов $z1$ и $z2$, используемых для визуализации. Для этого рассмотрим соответствующие векторы нагрузок в табл. 5.5 — именно их используют для интерпретации.

Таблица 5.5

Нормированные векторы нагрузок на признаки в первых двух факторах по методу МГК

	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп	Сингулярное значение	Вклад, %
c_1	0.5211	**–0.2693**	0.5804	0.5649	20.8532	72.96
c_2	0.3774	**0.9233**	0.0245	0.0669	11.6701	22.85

Примечание. Жирным выделены «странные» значения. В столбце «Вклад» — вклады факторов в квадратичный разброс данных, выраженные в процентах.

Вопрос 5.1. Правда ли, что c_1 и c_2 в табл. 5.13 ортогональны?

Подсказка: рассчитайте скалярное произведение этих векторов и убедитесь, что оно составляет число порядка 10^{-15}. Такая величина при вычислениях на MATLAB — несомненный машинный нуль, что доказывает ортогональность.

Качество интерпретации зависит от удачи. В отличие от ситуации в (5.12), первом сингулярном векторе, возникшем при стандартизации ранжирования, не все знаки в c_1 здесь положительны. Признак «Ширина чашелистика» получил в c_1 отрицательную нагрузку, что отражает его отрицательную корреляцию с «Длиной чашелистика», отмечавшуюся в п. 3.2.4.1, (и на самом деле и с другими признаками).

В остальном же первый фактор вполне хорош, каждый из трех оставшихся признаков одинаково высоко проявлен в нем, на уровне 0.5—0.6. Таким образом, с учетом относительно небольшой величины отрицательной компоненты, первый фактор можно интерпретировать как «размер цветка, проявленный в длине и ширине лепестка, а также длине чашелистика».

При интерпретации второго фактора важно отметить, что значения двух последних его компонент пренебрежимо малы. Напротив, вторая компонента существенно больше остальных. Поэтому второй фактор можно считать «выразителем ширины чашелистика».

Имея в виду объяснения в параграфе 5.1 (см. Свойство 2), перейдем от нормированных векторов в матрице сингулярных векторов z к их модельным версиям $z1 = z(:,1)\cdot m_1^{1/2}$ и $z2 = z(:,2)\cdot m_2^{1/2}$, после чего осуществим операции в среде MATLAB по получению визуализации, даваемой решением по методу МГК:

```
subplot(1,2,1); plot(z1,z2,'b.'); h = axis; axis(1.1*h)
% это команда открытия левой части рис. 5.5, затем представления объектов, заданных вышеопределенными координатами z1 и z2, синими точками, и последние две команды немного переформатируют картинку, умножая поле на 1.1, чтобы все точки оказались внутри его границ, задаваемых осями координат.

subplot(1,2,2);
% команда открытия поля на Рис. 5.5 справа

plot(z1(1:50),z2(1:50),'k*',z1(51:100),z2(51:100),'ko', ...
  z1(101:150),z2(101:150),'k^');
% эта команда помещает в поле первые 50 объектов, представляя их черными звездочками; вторые 50 объектов, представляя их кружками; и третьи 50 объектов, представляя их треугольниками.

h = axis;axis(1.1*h)
% перемасштабирование поля, как и в левой части
```

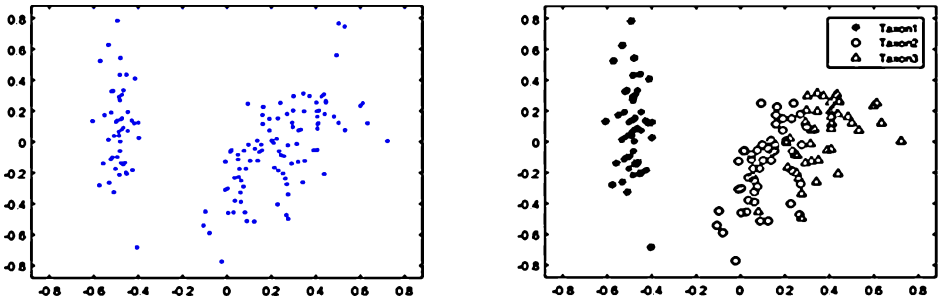


Рис. 5.5. **Визуализация данных Ирис:**
слева — как точек декартовой плоскости,
справа — как элементов трех разных таксонов

Как видим, таксон 1 (звездочки на рисунке справа) существенно отделен от остальных по первому фактору «размер без ширины чашелистика». Остальные два таксона тесно переплетены между собой, что является основой популярности базы данных Ирис. Несмотря на относительную простоту этих данных, пока что не удалось построить разумных алгоритмов, надежно разделяющих все таксоны.

Рабочий пример 5.3

Визуализация данных Ирис по традиционному МГК

Согласно традиционному подходу, главные компоненты ищутся с помощью собственных векторов матрицы ковариаций $C = Y^TY/N$. Расчет дает

$$C = \begin{matrix} 0.9933 & -0.1168 & 0.8659 & 0.8125 \\ -0.1168 & 0.9933 & -0.4256 & -0.3637 \\ 0.8659 & -0.4256 & 0.9933 & 0.9564 \\ 0.8125 & -0.3637 & 0.9564 & 0.9933 \end{matrix}.$$

Что-то странное есть в полученной матрице. Что именно? По диагонали должны стоять единицы, а здесь какие-то «недомерки». Почему? Потому что для z-стандартизованной матрицы данных матрица ковариации должна совпадать с матрицей корреляции, а корреляция признака с самим собой равняется единице. В чем дело?

Дело в методе расчета ковариаций в MATLAB, да и других пакетах. В этой среде коэффициенты ковариации рассчитываются не сами по себе, а как оценки ковариаций в Гауссовой функции плотности. Поскольку при этом вместо математического ожидания используется арифметическое среднее, то для получения несмещенной оценки надо делить не на N, а на $N-1$. Действительно,

$$C = Y^TY/(N-1) = \begin{matrix} 1.0000 & -0.1176 & 0.8718 & 0.8179 \\ -0.1176 & 1.0000 & -0.4284 & -0.3661 \\ 0.8718 & -0.4284 & 1.0000 & 0.9629 \\ 0.8179 & -0.3661 & 0.9629 & 1.0000 \end{matrix}.$$

Эта матрица, действительно, содержит коэффициенты корреляции между признаками данных Ирис. Нелишне бросить взгляд на элементы этой матрицы. Мы увидим, что три признака связаны высокими корреляциями, тогда как один из них, номер два, имеет отрицательные корреляции с остальными.

Это, очевидно, объясняет структуру первого собственного вектора нагрузок в табл. 5.5: вторая компонента в нем просто обязана быть отрицательной. Причина отрицательности второй компоненты фактора c_1 здесь — как на ладони, тогда как тот же феномен казался значительно более загадочным в рамках анализа сингулярного разложения. Как видим, расчет матрицы корреляций может оказаться полезным — наша интуиция легче справляется с парными сравнениями, чем с интегральными.

Проведем спектральный анализ этой матрицы:

```
[cc, la] = eig(C)
% команда MATLAB, выполняющая спектральный анализ матрицы
C; столбцы матрицы cc — нормированные собственные векторы
матрицы C; матрица la — диагональная, содержит соответствующие
собственные знаения.
```

Матрица собственных значений:

$$\text{la} = \begin{matrix} 0.0207 & 0 & 0 & 0 \\ 0 & 0.1468 & 0 & 0 \\ 0 & 0 & 0.9140 & 0 \\ 0 & 0 & 0 & 2.9185 \end{matrix}$$

располагает их по возрастанию (в отличие от расположения сингулярных значений). Матрица соответствующих собственных векторов:

$$\text{cc} = \begin{matrix} -0.2613 & 0.7196 & 0.3774 & 0.5211 \\ 0.1235 & -0.2444 & 0.9233 & -0.2693 \\ 0.8014 & -0.1421 & 0.0245 & 0.5804 \\ -0.5236 & -0.6343 & 0.0669 & 0.5649 \end{matrix}.$$

Как видно, четвертый столбец (первый собственный вектор) совпадает с сингулярным вектором c_1, а третий столбец (второй собственный вектор) совпадает с сингулярным вектором c_2 (см. табл. 5.13), как и должно быть.

Проверим связь между собственными и сингулярными значениями. Согласно теории, изложенной выше, собственные значения равны квадратам сингулярных значений, деленным на делитель в матрице C (в данном случае не N, а $N - 1$). Первое сингулярное значение равно 20.8532 (см. табл. 5.5), его квадрат — 434.8562, а частное от деления этого последнего на $N - 1 = 149$ равно 2.9185 — в точности максимальное собственное значение в матрице la выше.

Для визуализации по традиционному методу определяем векторы координат по первому и второму факторам: $y_1 = Yc_1, y_2 = Yc_2$. (Обратите внимание на то, что здесь явно не фигурируют ни сингулярные, ни собственные значения.) Визуализации по обоим методам МГК представлены на рис. 5.6. Бросается в глаза легкая несогласованность масштабов. Оба поля намеренно представлены квадратными. Это позволяет заметить еще

одну разницу в изображениях: на правом облако данных относительно шире, чем на левом. Это отражает принцип масштабирования по методу, основанному на сингулярном разложении данных, где масштабы пропорциональны корню квадратному из соответствующих сингулярных значений.

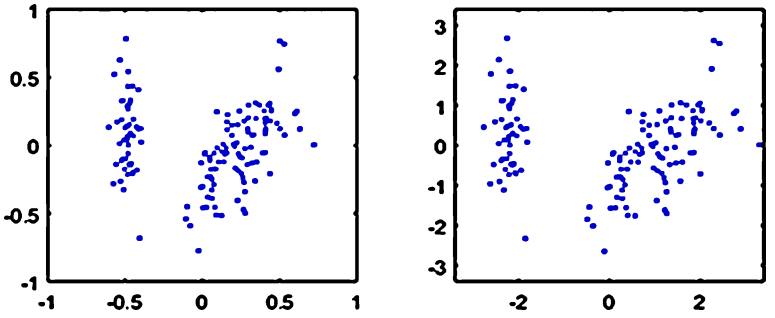


Рис. 5.6. **Визуализация данных Ирис:**
слева — по методу сингулярного разложения,
справа — по традиционному методу

Вопрос 5.2. Рассмотрим подмножество объектов S. Обозначим через $y(S)$ вектор средних значений признаков на S. Докажите, что визуализация вектора $y(S)$ на плоскости сингулярных векторов z как точки $z^* = \sqrt{\mu}z = Y \cdot y(S) / \sqrt{\mu}$, определяемой уравнением (5.3), эквивалентна представлению категории S значениями двумерных точек z^*_{1i} и z^*_{2i} по $i \in S$.

Ответ. Действительно, при вычислении среднего используются только операции сложения и деления на константу, которые сохраняются в линейных пространствах при применении линейных операций, включая операцию умножения на матрицу.

Вопрос 5.3. Почему для визуализации данных мы рекомендуем умножать нормированные сингулярные векторы на корни квадратные из соответствующих сингулярных значений?

Ответ. Потому что таковы величины в уравнениях (5.5) и (5.6″), определяющих модель МГК для двумерной аппроксимации.

Самостоятельная работа

5.2. Повторить работу, проделанную выше для данных Ирис, для визуализации компаний — элементов табл. 1.1 и 1.2 . Мы предоставляем читателю самому повторить все шаги метода на табл. 1.2, стандартизованной по методу z-скоринг, и сравнить результаты с теми, что представлены в учебнике [32, стр. 119—120].

Задание 5.1. Визуализация данных о компаниях.

Здесь мы рассмотрим стандартизацию, использующую деление на размах после вычитания среднего. Более того, мы дополни-

тельно пронормируем последние три столбца. Дело в том, что эти три столбца возникли при квантизации номинального признака «Сектор экономики» и как бы утроили вклад последнего в разброс данных. Деление этих столбцов на $\sqrt{3}$ оборачивается делением их на 3 как слагаемых квадратичного разброса данных. То есть данная операция как бы возвращает признаку «Сектор экономики» его исходный единичный вес. Результаты этой стандартизации представлены в табл. 5.6.

Таблица 5.6

Данные о Компаниях из табл. 1.2 после стандартизации, включающей центрирование признаков и нормирование их на размах с последующим делением последних трех столбцов на $\sqrt{3}$

Названия	Income	MShare	MainC	Internet	Chemistry	Metal	Retail
Аве	–0.1994	0.233	–0.33	–0.625	0.3608	–0.2165	–0.1443
Ант	0.4017	0.0456	0	–0.625	0.3608	–0.2165	–0.1443
Аст	0.0838	0.0943	0	–0.625	–0.2165	0.3608	–0.1443
Бма	–0.2341	–0.1515	–0.33	0.375	0.3608	–0.2165	–0.1443
Бре	0.1879	–0.2877	0	0.375	–0.2165	0.3608	–0.1443
Бум	–0.5983	–0.4191	–0.33	0.375	–0.2165	0.3608	–0.1443
Виж	0.0838	–0.0955	0.33	0.375	–0.2165	–0.2165	0.433
Вур	0.2746	0.5809	0.67	0.375	–0.2165	–0.2165	0.433

В табл. 5.7 представлены основные характеристики первых двух элементов сингулярного разложения данных табл. 5.6. Следующие вопросы предназначены для того, чтобы помочь студенту в освоении метода.

Вопрос 5.4. Разброс данных. Докажите, что квадратичный разброс данных в табл. 5.6 равен $T = 9.4457$.

Вопрос 5.5. Рассчитайте первые две сингулярные тройки.

Ответ. См. табл. 5.7.

Вопрос 5.6. Нормализация. Какой смысл в делении последних трех столбцов на $\sqrt{3}$?

О. См. абзац над табл. 5.6.

Первые две компоненты в табл. 5.7 учитывают 76.75 % разброса данных, что считается достаточно высоким уровнем. По нашему опыту, интерпретируемые результаты получаются, когда визуализация по МГК учитывает 50 % разброса данных или более.

Для интерпретации факторов МГК посмотрим на величины нагрузок в последнем столбце. Какие величины? Самые большие (по абсолютной величине). В первом столбце это Интернет (0.79), ОснПот (0.38) и Химия (–0.35).

Таблица 5.7

Элементы сингулярного разложения для матрицы в табл. 5.6

Объекты	Главн. компоненты		Факторные баллы		Признаки	Нагрузки	
	z_1	z_2	$\mu^{1/2}z_1$	$\mu^{1/2}z_2$		c_1	c_2
Аве	–0.5183	–0.0668	–0.6480	–0.0806	Доход	0.0240	–0.4792
Ант	–0.4255	–0.3124	–0.5320	–0.3769	ДоляРын	–0.0299	–0.5007
Аст	–0.2937	–0.0876	–0.3672	–0.1057	ОснПот	0.3752	–0.4897
Бма	–0.0051	0.2971	–0.0064	0.3584	Интернет	0.7916	0.3206
Бре	0.2216	0.2298	0.2771	0.2772	Химия	–0.3505	–0.0326
Бум	0.1321	0.6458	0.1651	0.7792	Металл	0.0222	0.3125
Виж	0.4094–	0.1492	0.5119	–0.1800	Торговля	0.3283	–0.2799
Вур	0.4794–	0.5567	0.5994	–0.6717			
Сингулярные значения	1.5632	1.4558			Общий вклад в разброс данных, %	76.75	
Вклады в разброс данных, %	41.10	35.65					

Таким образом, первый фактор определяется использованием Интернета и наличием большого количества потребителей. Напротив, большие нагрузки во втором факторе — это отрицательные корреляции с Доходом (–0.48), Долей Рынка (–0.50) и количеством основных потребителей (–0.49) — чем их меньше, тем лучше. Такое впечатление, что знак c_2 — неправильный. Если поменять знак, то интерпретация второго фактора ясна — «Рыночный успех». Но тогда надо не забыть поменять и знак второй главной компоненты (см. рис. 5.6. справа)[1]!

Визуализация полученного решения представлена на рис. 5.7, причем изображение на правом поле скорректировано так, чтобы значимые нагрузки второго фактора — на Доход, Долю рынка и Количество потребителей — были положительны. Мы видим, что компании, начинающиеся на букву В — лидеры рыночного успеха. Компании, начинающиеся на Б, в этом плане — неудачники, хотя и пользуются Интернетом.

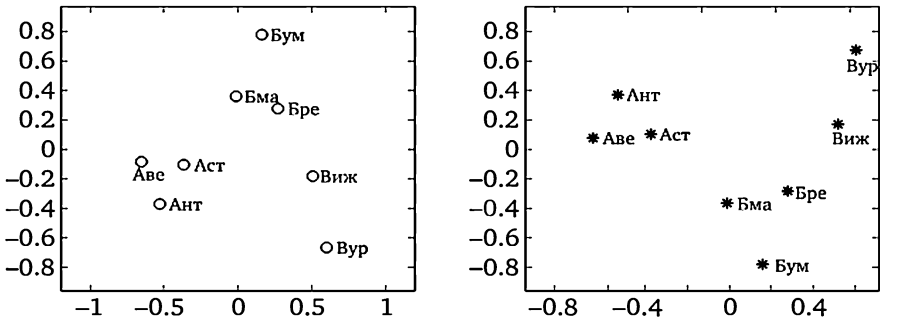


Рис 5.7. **Визуализация данных о компаниях из табл. 5.6:** как есть в столбце «Факторные баллы» табл. 5.7 — слева, и с отрицательным знаком второго фактора — справа

Обратим внимание, что на обоих рисунках компании с названиями, начинающимися на одну и ту же букву, образуют кластеры — группы близких друг к другу точек. Это положительно отвечает на вопрос о том, отражают ли признаки в таблице данных тот факт, что компании с названиями, начинающимися на одну и ту же букву, производят похожую продукцию.

5.2. Модель и метод К-средних для кластерного анализа

5.2.1. Параллельный метод *K*-средних

Рассмотрим различные виды структур данных на рис. 5.8: отчетливая структура кластеров на рис. 5.8, *а*, «капля» на рис. 5.8, *б* и неопределенное «облако» на рис. 5.8, *г*.

[1] Имеется в виду, что вместе с тройкой (μ, z, c) является сингулярной и тройка $(\mu, -z, -c)$.

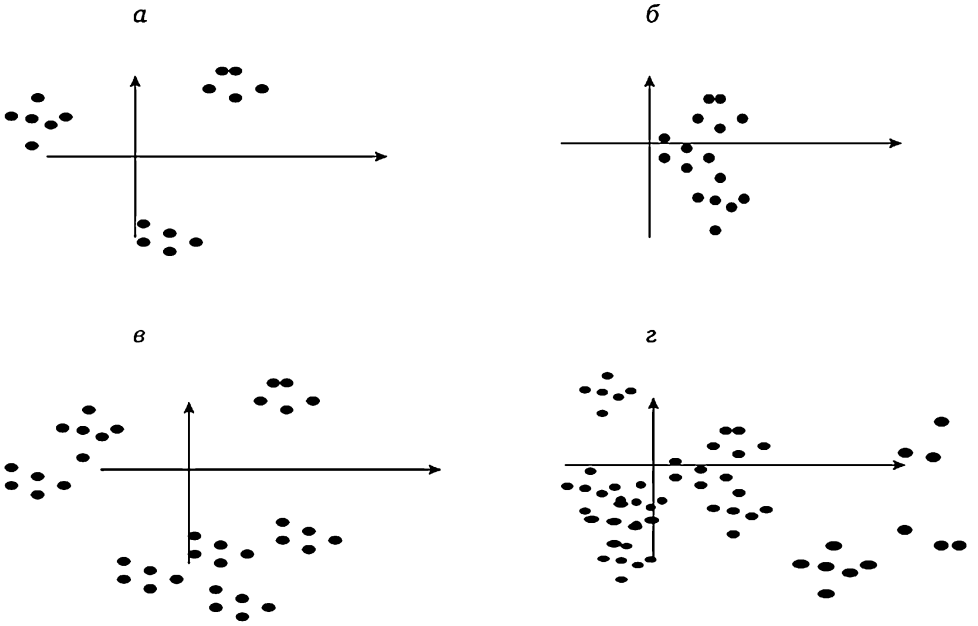


Рис. 5.8. **Четкая структура кластеров на *а* и *в*; данные без четкой структуры *б* и *г***

Есть мнение, что термин «кластеризация» применяется исключительно к структурам данных, представленным на рис. 5.8, *а* и *в*, хотя на рис. 5.8, *в*, можно увидеть 3 или 7 кластеров в зависимости от уровня гранулярности. На рис. 5.8, *б* нет «естественных» кластеров, тогда как на рис. 5.8, *г* часть объектов организована в кластеры, а часть — нет.

Чтобы кластеры служили моделями натуральных классов и категорий, они должны быть не просто найдены, но также и концептуально объяснены. Действительно, говоря о классе как элементе какой-либо классификации, мы всегда имеем в виду двоякую структуру. С одной стороны, класс — это понятие, встроенное в соответствующий фрагмент знания (в логике это называют *интенсиональной* интерпретацией), а с другой стороны, класс может быть представлен множеством соответствующих ему предметов реального мира (это называют *экстенсиональной* интерпретацией). Например, «береза» — это дерево — элемент биологической таксономии, обладающий такими-то признаками. С другой стороны, объем этого понятия хорошо представлен всеми экземплярами березы, растущими в лесах, лугах, городах и пр. Подобным же образом для эмпирических классов — т. е. кластеров, эти два подхода, построение и описание, должны сосуществовать.

Как показано на рис. 5.9 слева, кластер можно описать без больших ошибок, если он отделён от остальных объектов. Использова-

ние такого подхода может быть отражено в разделении всех методов нахождения кластеров на следующие категории:

(а) кластеры получают непосредственно в терминах признаков (данный способ часто называется концептуальной кластеризацией);

(б) кластеры получают одновременно с трансформацией пространства признаков, что делает кластеры более четкими; данное направление совсем молодое и пока хорошо не изучено;

(в) сначала получают кластеры как подмножества объектов, а затем уже производят их описание — такой способ является самым распространенным в настоящее время.

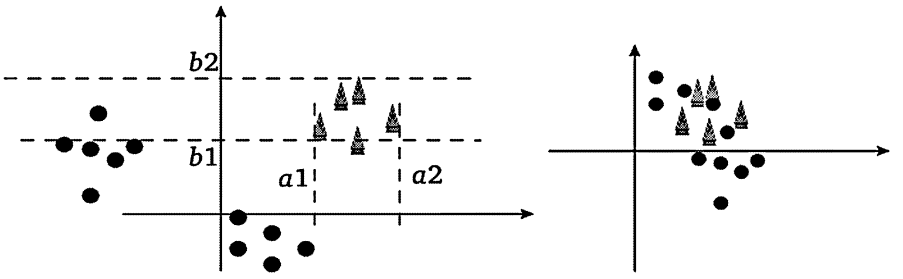


Рис. 5.9. **Кластер, состоящий из треугольников на левом рисунке, хорошо описывается предикатом «$a1 < x < a2$ и $b1 < y < b2$». Аналогичный кластер на правом рисунке не может быть хорошо, т. е. без ошибок первого и второго рода, описан с помощью интервальных предикатов**

Метод *K*-средних — наиболее популярный метод кластеризации (типа (в)), который в разных формах представлен во всех основных статистических пакетах, таких, как SPSS и SAS, а также в пакетах анализа и майнинга данных, таких, как Clementine, Weka и DBMiner. Метод очень популярен во многих приложениях, например, в анализе изображений, маркетинговых исследованиях, биоинформатике и медицинской информатике.

Далее мы будем говорить о K кластерах с номерами $k = 1, 2, \ldots, K$ и соответственно, называть метод *K*-средних. Процесс нахождения кластеров по методу *K*-средних стартует с K центров — обычно в качестве таковых берутся какие-либо случайные объекты из анализируемого множества. Затем последовательно выполняются итерации, каждая из которых состоит из двух шагов:

1) обновление кластеров (вокруг центров),
2) обновление центров (внутри кластеров).

Итерации повторяются, пока процесс не сойдется.

Рис. 5.10 иллюстрирует одну итерацию процесса формирования кластеров. Согласно этой логике, как бы центры не были выбраны изначально, они заменяются в процессе вычислений на те, которые представляют места скопления объектов.

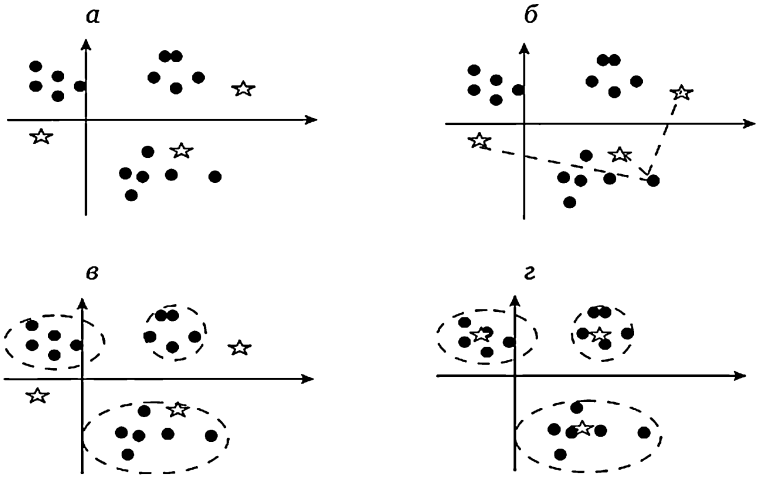


Рис. 5.10. **Итерации метода *K*-средних:**
а — инициализация центров, представленных звездочками;
б — обновление кластеров с помощью правила минимального расстояния;
на рисунке, к примеру, пунктирными линиями показаны расстояния
от центров до каждого объекта; *в* — кластеры сформированы;
г — новые центры сформированы как центры масс кластеров

Теперь опишем процесс формирования кластеров более точно.

Алгоритм *K*-средних

0. Инициализация: пользователь выбирает число K кластеров и назначает K гипотетических центров, см. рис. 5.10, *а*;

1. Обновление кластеров: При заданных K центрах c_k ($k = 1, 2, \ldots, K$), каждый объект $i \in I$ приписывается одному из центров по правилу минимального расстояния: вычисляются расстояния от i до каждого c_k; объект i приписывается ближайшему центру c_k, см. рис. 5.10, *б*. Те объекты, которые приписаны центру c_k, образуют кластер S_k ($k = 1, 2, \ldots, K$), см. рис. 5.10, *в*.

2. Обновление центров: Вычисляется арифметический центр (центр масс) каждого кластера S_k, который и назначается новым центром c'_k ($k = 1, 2, \ldots, K$), см. рис. 5.10, *г*. Компоненты центра вычисляются как средние арифметические соответствующих компонент объектов из S_k.

3. Правило остановки: Новые центры c'_k сравниваются со старыми. Если $c'_k = c_k$ для каждого $k = 1, 2, \ldots, K$, то вычисления останавливаются и выдаются результаты: центр c'_k и кластер S_k для каждого $k = 1, 2, \ldots, K$. Если же хотя бы одно из равенств не верно, то каждый центр c_k заменяется вновь полученным центром c'_k, и процесс возвращается к шагу 1, «Обновление кластеров».

Модель суммаризации, лежащая в основе метода, предполагает, что каждый кластер представлен своим центром, иногда также на-

зываемым стандартной точкой кластера или прототипом кластера [32, 33]. Прототип как бы концентрирует в себе всю информацию о кластере.

Таблица 5.8

Нормализованные данные компаний из табл. 1.2

Ав	–0.20	0.23	–0.33	–0.63	0.36	–0.22	–0.14
Ан	0.40	0.05	0	–0.63	0.36	–0.22	–0.14
Ас	0.08	0.09	0	–0.63	–0.22	0.36	–0.14
Бм	–0.23	–0.15	–0.33	0.38	0.36	–0.22	–0.14
Бр	0.19	–0.29	0	0.38	–0.22	0.36	–0.14
Бу	–0.60	–0.42	–0.33	0.38	–0.22	0.36	–0.14
Ви	0.08	–0.10	0.33	0.38	–0.22	–0.22	0.43
Ву	0.27	0.58	0.67	0.38	–0.22	–0.22	0.43
Вклад	0.74	0.69	0.89	1.88	0.62	0.62	0.50
Вклад в процентах	12.42	11.66	14.95	31.54	10.51	10.51	8.41

Примечание: данные компаний из табл. 1.2, нормализованные: (i) вычитанием из столбцов их средних значений, (ii) затем делением столбцов на их размахи и (iii) дополнительным делением трех последних столбцов, отвечающих трем качественным категориям признака «Сектор экономики», на $\sqrt{3}$. Вклады признаков в исходные данные, вычисляемые как суммы квадратов элементов соответствующего столбца, представлены ниже.

Рабочий пример 5.4

Кластеризация данных «Компании» методом *K*-средних

Рассмотрим нормализованные данные о компаниях из табл. 1.2 (см. табл. 5.8).

Данные могут быть визуализированы в пространстве двух главных компонент, см. рис. 5.11.

Возьмем для примера объекты «Ан», «Бр» и «Ви» в качестве центров трех кластеров. Теперь мы можем сравнить каждый объект с центром, чтобы понять, какой центр ближе к каждому конкретному объекту. Для сравнения любых двух точек использован квадрат Евклидовой меры вычисления расстояний (см. табл. 5.9).

Таблица 5.9

Вычисление Евклидовой меры между строками «Ав» и «Ан» в табл. 5.8 как суммы квадратов разностей между соответствующими ячейками

Точки	Координаты							d(Ан, Ав)
Ан	0.40	0.05	0.00	–0.63	0.36	–0.22	–0.14	
Ав	–0.20	0.23	–0.33	–0.63	0.36	–0.22	–0.14	
Ан-Ав	0.60	–0.18	0.33	0.00	0.00	0.00	0.00	
Квадраты	0.36	0.03	0.11	0.00	0.00	0.00	0.00	0.50

Каждый объект приписывается ближайшему центру согласно правилу минимального расстояния (см. табл. 5.10, в которой представлены все

расстояния между объектами и центрами; жирным выделены те из расстояний, которые вычислены с помощью правила минимального расстояния.)

Таблица 5.10

Расстояния между тремя компаниями — центрами и всеми остальными компаниями

Точка	Ав	Ан	Ас	Бм	Бр	Бу	Ви	Ву
Центр 1	**0.22**	**0.19**	**0.31**	1.31	1.49	2.12	1.76	2.36
Центр 2	1.58	1.84	1.36	**0.33**	**0.29**	**0.25**	0.95	2.3
Центр 3	2.5	2	1.95	1.69	1.2	2.4	**0.15**	**0.15**

Примечание: каждый столбец матрицы показывает три расстояния между компанией и центром; значения матрицы, выделенные жирным, показывают наилучшие пары «центр — компания».

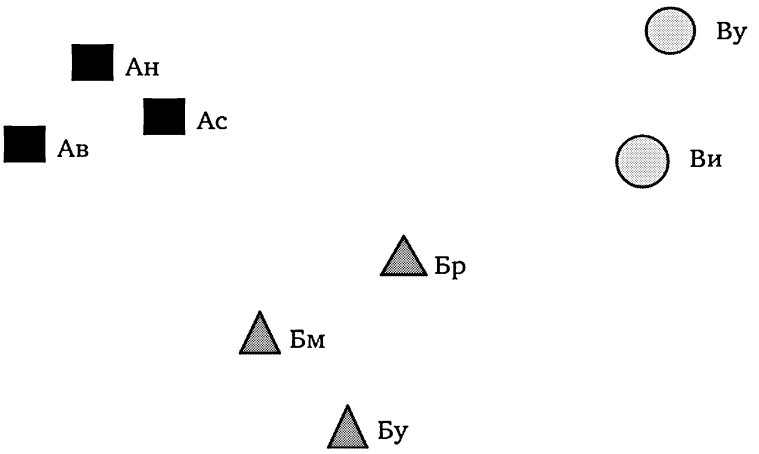


Рис. 5.11. **Строки табл. 5.9 на плоскости двух первых главных компонент:** достаточно просто отделить кластеры, сформированные продуктами: расстояния в группах А, Б и В (отображенных квадратами, треугольниками и кругами соответственно) меньше, чем между ними

Объекты, которые приписаны одному и тому же центру, формируют текущий кластер. Кластеры, найденные по табл. 5.10: $S1$ = {Ав, Ан, Ас}, $S2$ = {Бм, Бр, Бу}, и $S3$ = {Ви, Ву}. Приведенные кластеры соответствуют производимым продуктам. Метод K-средних подразумевает, что новые предполагаемые кластеры будут определяться теми же центрами.

Дальше необходимо обновлять центры, используя информацию из предопределенных кластеров. Новые центры определены как центры предполагаемых кластеров, компонентами которых являются средние соответствующих компонент внутри кластеров; данные представлены в табл. 5.11.

Обновленные центры отличаются от центров предыдущего шага.

Таблица 5.11

Предполагаемые кластеры из табл. 5.10 и их центры

Ав	–0.20	0.23	–0.33	–0.63	0.36	–0.22	–0.14
Ан	0.40	0.05	0	–0.63	0.36	–0.22	–0.14
Ас	0.08	0.09	0	–0.63	–0.22	0.36	–0.14
Центр 1	0.10	0.12	–0.11	–0.63	0.17	–0.02	–0.14
Бм	–0.23	–0.15	–0.33	0.38	0.36	–0.22	–0.14
Бр	0.19	–0.29	0	0.38	–0.22	0.36	–0.14
Бу	–0.60	–0.42	–0.33	0.38	–0.22	0.36	–0.14
Центр 2	–0.21	–0.29	–0.22	0.38	–0.02	0.17	–0.14
Ви	0.08	–0.10	0.33	0.38	–0.22	–0.22	0.43
Ву	0.27	0.58	0.67	0.38	–0.22	–0.22	0.43
Центр 3	0.18	0.24	0.50	0.38	–0.22	–0.22	0.43

Следовательно, необходимо обновить кластеры, используя расстояния между обновленными центрами и объектами; расстояния приведены в табл. 5.12. После следующей итерации получаем те же центры с теми же наборами объектов по правилу минимального расстояния. Следовательно, процесс стабилизировался: если запустить еще несколько итераций, ничего нового не произойдет и центры останутся на своем месте. Процесс останавливается на данной итерации и метод возвращает найденные кластеры с центрами (в стандартизованной форме они представлены в табл. 5.11).

Таблица 5.12

Расстояния между тремя обновленными центрами и всеми компаниями; наилучшие пары «центр — компания» выделены жирным

Точка	Ав	Ан	Ас	Бм	Бр	Бу	Ви	Ву
Ан	**0.50**	**0.00**	**0.77**	1.55	1.82	2.99	1.90	2.41
Бр	2.20	1.82	1.16	**0.97**	**0.00**	**0.75**	0.83	1.87
Ви	2.30	1.90	1.81	1.22	0.83	1.68	**0.00**	**0.61**

Очевидно, что данный результат сильно зависит от способа стандартизации данных, сделанной перед началом вычислений, так как метод основан на суммировании разностей значений (возведенной в квадрат) различных признаков, которые сильно зависят от выбранных масштабов.

Самостоятельная работа

5.3. Проведите кластеризацию компаний, начиная с тех же объектов Ан, Бр и Ви в качестве центров трех кластеров, при условии, что стандартизация данных выполнена путём вычитания средних арифметических с последующим делением на стандартные отклонения признаков.

5.4. Проведите кластеризацию компаний на два кластера, начиная с объектов Ан и Ви в качестве центров кластеров, не меняя стандартизацию признаков табл. 4.9.

5.2.2. Критерий, минимизируемый методом *K*-средних

Метод *K*-средних оперирует с разбиением множества объектов на *K* непересекающихся кластеров, $S = \{S_1, S_2, \ldots, S_K\}$, представленных в виде списков номеров объектов S_k, и центрами этих кластеров $c_k = (c_{k1}, c_{k2}, \ldots, c_{kV})$, $k = 1, 2, \ldots, K$.

В основе работы алгоритма *K*-средних находится математическая модель представления данных с помощью кластерной структуры. Согласно этой модели, каждый объект i, заданный соответствующей строкой матрицы Y как $y_i = (y_{i1}, y_{i2}, \ldots, y_{iV})$, принадлежит к какому-то из кластеров S_k, и равен, с точностью до малых ошибок, центру этого кластера:

$$y_{iv} = c_{kv} + e_{iv} \text{ для всех } i \in S_k \text{ и всех } v = 1, 2, \ldots, V. \qquad (5.13)$$

Проблема кластер-анализа ставится так: найти такое разбиение $S = \{S_1, S_2, \ldots, S_K\}$ и такие центры кластеров $c_k = (c_{k1}, c_{k2}, \ldots, c_{kV})$, $k = 1, 2, \ldots, K$, которые минимизировали бы суммарную квадратичную ошибку в равенствах (5.13), равную

$$L^2 = \sum_{i \in I} \sum_{v \in V} e_{iv}^2 = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} (y_{iv} - c_{kv})^2. \qquad (5.14)$$

Критерий (5.14) может быть переформулирован в терминах евклидовских расстояний. Величина (5.14) есть не что иное как сумма квадратов Евклидовских расстояний между объектами и центрами их кластеров (см. рис. 5.12).

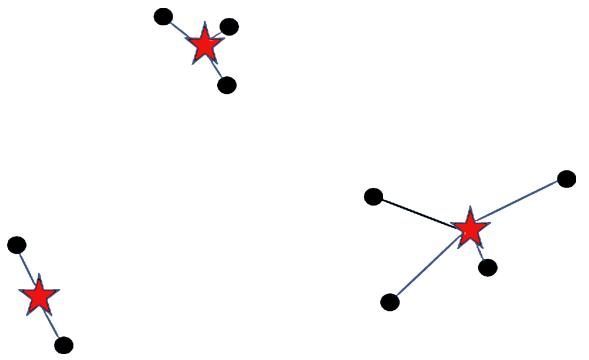


Рис. 5.12. **Расстояния между объектами и центрами в критерии *W(S, c)***

Нужно иметь в виду, что число расстояний в сумме равно *N* и не зависит от количества кластеров. Вот явное выражение критерия через расстояния между векторами (c_{kv}) и (y_{iv}):

$$L^2 = W(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k). \qquad (5.15)$$

Здесь расстояние d между любыми векторами $x = (x_v)$ и $y = (y_v)$ определяется формулой

$$d(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_V - y_V)^2,$$

что в точности совпадает с внутренней суммой в (5.15) при $x = y_i$ и $y = c_k$.

Данный критерий зависит от двух наборов переменных, S и c, и поэтому может быть минимизирован с помощью метода альтернативной минимизации, который заключается в регулярном повторении одного и того же шага минимизации по той или иной группе переменных. При заданных значениях одной группы признаков нужно оптимизировать критерий по другой группе признаков. Потом, при найденных значениях этой второй группы признаков следует оптимизировать критерий по первой группе признаков, и так продолжать, пока процесс не сойдется.

При заданном наборе центров $c = (c_1, c_2, \ldots, c_K)$ можно легко найти разбиение S, минимизирующее суммарное расстояние (5.15). Для этого достаточно определить для каждого объекта $i \in I$ K расстояний $d(y_i, c_1)$, $d(y_i, c_2)$, ..., $d(y_i, c_K)$ до всех центров. Минимизация (5.15) по S при заданных центрах происходит с использованием правила минимального расстояния: для каждого $i \in I$, выбираем минимум из $d(y_i, c_k)$, $k = 1, \ldots, K$, и приписываем объект i ближайшему центру. В случае, когда минимум достигается на нескольких центрах, можно взять любой из них. При этом может оказаться, что некоторым центрам не приписано ни одного объекта. Чтобы избежать такой дегенерации решения, достаточно взять начальные центры из объектов анализируемого множества.

Следующим шагом альтернативной минимизации будет минимизация (5.15) относительно центров c при заданном S, найденном на предыдущем шаге. Аддитивная форма критерия (5.15) делает кластеры и их компоненты независимыми друг от друга. Как хорошо известно (см. табл. 2.1 в теме 2), решением задачи минимизации квадрата ошибки является среднее, поэтому векторы средних внутри-кластерных значений минимизируют (5.15) по c при данном S.

Таким образом, метод альтернативной минимизации критерия (5.15) состоит в следующем. Сначала каким-то образом выбираются центры $c = (c_1, c_2, \ldots, c_K)$, после чего начинаются повторяющиеся итерации, состоящие из двух шагов: (а) обновление кластеров — определение кластеров S по правилу минимального расстояния и (б) обновление центров как векторов внутри-кластерных средних значений по каждой компоненте отдельно. Вычисления останавливаются, когда новые кластеры совпадают с кластерами, полученными на предыдущем шаге. Очевидно, этот метод совпадает с параллельным методом K-средних.

Сходимость метода достигается благодаря двум следующим фактам: (i) на каждом шаге значение критерия (5.15) может только

уменьшаться, и (ii) количество разбиений S конечно. Конечно, достижение глобального оптимума не гарантировано.

Вопрос 5.7. (1) Сколько расстояний суммируются в $W(S, c)$?

Ответ. Ровно столько, сколько объектов, N. (2) Зависит ли это число от количества кластеров K?

Ответ. Нет.

Вопрос 5.8. Верно ли, что имеет место следующий факт: чем больше K, тем меньше минимальное значение $W(S, c)$?

Ответ. Да.

Подсказка: так как число расстояний в критерии одно и то же, то их можно сделать в среднем меньше путем увеличении числа кластеров. Например, возьмем любой кластер оптимального разбиения S и разобьем его пополам, взяв два центра вместо одного. Сумма расстояний до двух новых центров станет меньше, чем была до одного центра.

Вопрос 5.9. Почему гарантирована сходимость метода K-средних?

*Ответ. П*отому что метод K-средних является альтернативным процессом минимизации, в котором критерий $W(S, c)$ может только уменьшаться на каждом шаге. Сходимость следует из того, что на конечном множестве объектов I может быть только конечное число разбиений.

Вопрос 5.10. Допустим, что $d(y_i, c_k)$ в $W(S, c)$ это не квадрат евклидова расстояния, а расстояние городских кварталов. Можно ли изменить метод K-средних таким образом, чтобы сделать его методом альтернативной минимизации для модифицированного критерия $W(S, c)$?

Ответ. Да, при обновлении кластеров надо использовать расстояние городских кварталов, а центры формировать не из средних, а из медиан признаков внутри кластеров.

Вопрос 5.11. Приведут ли действия в Вопросе 5.10 к каким-либо отличиям?

Ответ. Обычно да, особенно в случае, когда распределения признаков асимметричны.

Вопрос 5.12. Продемонстрировать, что в данных о компаниях значение $W(S, c)$ на разбиении {1—2—3, 4—5—6, 7—8} по продукту меньше, чем на разбиении {1—4—6, 2, 3—5—7—8}, найденном, начиная с центров в объектах 1, 2 and 3.

Ответ. Действительно, суммы внутрикластерных расстояний до центров кластеров по продукту равны 0.7193, 0.8701, 0.3070, соответственно, что в сумме дает 1.8964. Во втором же разбиении суммарные расстояния до центров — 1.4411, 0, 2.1789, что в итоге дает 3.62, чуть ли не в два раза больше.

Вопрос 5.13. Доказать, что в данных о компаниях значение $W(S, c)$ на разбиении {1—2—3, 4—5—6, 7—8} по продукту меньше, чем на разбиении {1—2—3, 4—6, 5—7—8}, найденном, исходя из начальных центров в объектах 1, 4 и 7.

Ответ. Действительно, суммы внутри-кластерных расстояний до центров кластеров по продукту равны 0.7193, 0.8701, 0.3070 соответственно, т. е. в сумме 1.8964, тогда как внутри-кластерные суммы расстояний во втором разбиении равны 0.7193, 0.4413, 1.1020, т. е. в сумме 2.2626.

Вопрос 5.14. Сформулируем шаги алгоритма *K*-средних.

Ответ. Выбираем начальные данные.

1. Нормализуем данные.
2. Выбираем количество кластеров *K*.
3. Определяем *K* потенциальных центров.
4. Обновляем кластеры, приписывая объекты центрам по правилу минимальных расстояний.
5. Обновляем центры, определяя центры как центры масс кластеров.
6. Повторяем шаги 4 и 5, пока не сойдутся.

Псевдокод (в кодах MATLAB) для самых сложных шагов 4 и 5 представлен ниже.

4. Обновление кластеров: По правилу минимальных расстояний присваиваем точки центрам:

Вход: матрица данных *X* и *K*×*V* матрица центров, выход: вектор размерности *N* меток кластеров на объектах, а также и суммы расстояний до центров кластеров (переменная wc в коде):

```
function [labelc,wc] = clusterupdate(X,cent)
[K,m] = size(cent);
[N,m] = size(X);
for k = 1:K
  cc = cent(k,:); %центр кластера k
  Ck = repmat(cc,N,1);
  dif = X-Ck;
  ddif = dif.*dif; %Nxm матрица квадратов разностей
  dist(k,:) = sum(ddif'); % это расстояния от всех объектов
                          % до соответствующих центров
end
[aa,bb] = min(dist); % правило минимального расстояния
wc = sum(aa);
labelc = bb;
return
```

5. Обновление центров: Определяем центры масс кластеров, заданных вектором номеров кластеров *labelc* соответственно данным в матрице *X*. Псевдокод вычисления *K*×*V* матрицы центров:

```
function centres = ceupdate(X,labelc)
  K = max(labelc);
  for k = 1:K
    clk = find(labelc = = k);
    elemk = X(clk,:);
    centres(k,:) = mean(elemk);
  end
return
```

Параллельный метод *K*-средних реализуется включением в вычисление шагов 3—6, описанных выше, и выдачей массива, названного *Clusters*, а также суммарных расстояний до центров кластеров в векторе uds:

```
function [Clusters,uds] = k_means(X,cent)
[N,m] = size(X);
[K,m1] = size(cent);
flag = 0;
%-- переменная для останова
membership = zeros(N,1);
dd = sum(sum(X.*X));
%-- разброс данных
%--- обновление кластеров и центров
while flag = = 0
  [labelc,wc] = clusterupdate(Y,cent);
  if isequal(labelc,membership)
  %--критерий останова
    flag = 1;
    centre = cent;
    w = wc;
  else
    cent = ceupdate(Y,labelc);
    membership = labelc;
  end
end
%-----подготовка выдачи результатов--------------
uds = w*100/dd;
Clusters{1} = membership;
Clusters{2} = centre;
return
```

Вопрос 5.15. Показать, что если метод К-средних применить к данным об ирисах, стандартизованным путем вычитания среднего из каждого признака с последующим делением признака на размах, при $K = 3$ и объектах 1, 51, и 101 в качестве начальных центров, то итоговые результаты кластеризации и перекрестной классификации будут, как в табл. 5.13.

Таблица 5.13

Таблица сопряженности разбиения Ирисов на таксоны и результатов применения методов к-средних (с начальными центрами 1, 51 и 101). Кластер 1 совпадает с таксоном *Iris Setosa*, но остальные два кластера неправильно классифицируют 14 + 3 = 17 объектов двух оставшихся групп

Кластер	Setosa	Versicolor	Virginica	Всего
S1	50	0	0	50
S2	0	47	14	61
S3	0	3	36	39
Всего	50	50	50	150

5.2.3. Особенности метода *K*-средних

Метод *K*-средних имеет сильные стороны, обеспечивающие его популярность. Концептуально он моделирует процесс создания типологии человеком, при этом типы характеризуются центрами c_k и кластерами S_k. Кроме того, алгоритм имеет хорошие вычислительные характеристики: вычисления просты и интуитивны, метод сходится быстро и не требует много памяти.

Метод имеет и слабые стороны:

(а) неясно, где брать информацию о том, как инициализировать число кластеров *K* и начальные центры;

(б) результаты сильно зависит от инициализации и нормализации исходных данных;

(в) метод не защищен от присутствия «случайных», «проходных» объектов или признаков, наличие которых может сильно исказить результаты;

(г) не проработаны вопросы интерпретации кластеров в тех случаях, когда центры не слишком отличаются от вектора средних на всем множестве.

Существуют модификации метода *K*-средних, позволяющие частично преодолеть указанные проблемы (см. [33]). Одна из них, нацеленная на решение проблемы (а), будет рассмотрена далее в подпараграфе 5.2.5.

Рабочий пример 5.5

Зависимость метода *K*-средних от начальных данных: преимущества и недостатки

Очевидный минус метода *K*-средних заключается в том, что его результаты зависят от начальных данных — центров и количества кластеров *K*. Действительно, если на старте выбрать неправильные центры, то результаты метода будут неутешительными.

В некоторых пакетах, таких как SPSS (Green, Salkind 2003), в качестве центров выбираются первые *K* объектов. Если применить этот прием к данным о компаниях, начальные центры будут в объектах Ав, Ан и Ас. В этом случае метод выдаст не слишком удачные кластеры *S*1 = {Ав, Бм, Бр}, *S*2 = {Ан}, и S3 = {Ас, Бу, Ви, Ву} (см. вопрос 5.12). Может показаться, что этот результат определяется неудачным выбором начальных центров — из одного и того же кластера. Тем не менее, даже если начальные центры берутся из правильных кластеров, это необязательно гарантирует правильный результат. Начнем, например, с точек Ав, Бм и Ви (отметим, что данные компании производят разные продукты!). Попробуйте самостоятельно убедиться в том, что итоговый результат будет удручающим — *S*1 = {Ав, Ан, Ас}, S2 = {Бм, Бу}, S3 = {Бр, Ви, Ву} (см. также вопрос 5.13).

На рис. 5.13 продемонстрирован тот факт, что нестабильность результатов метода возникает не в каком-то специально подстроенном случае — такая вещь происходит достаточно часто. На рис. 5.13 показаны два чет-

ких кластера, которые можно представить как равномерно распределенные множества точек, и две различных инициализации: симметричная представлена на рис. 5.13, *а*, ассиметричная — на рис. 5.13, *б*. В случае $K = 2$ правило минимального расстояния приводит к проведению гиперплоскости, которая ортогонально разделяет два центра и проходит через середину отрезка, соединяющего эти два центра; гиперплоскость показана на рис. 5.13 как прямая, разделяющая центры. На рис. 5.13, *а* изображен случай, когда начальные центры более или менее симметрично расположены друг относительно друга. Поэтому линия, проходящая через середину, действительно отделит кластеры друг от друга. В случае рис. 5.13, *б* начальные центры сильно асимметричны. Поэтому разделяющая линия отсечет кусок одного из данных кластеров, изменяя тем самым местоположение будущих центров. Финальное деление все равно проходит через один из кластеров, что совершенно противоречит здравому смыслу.

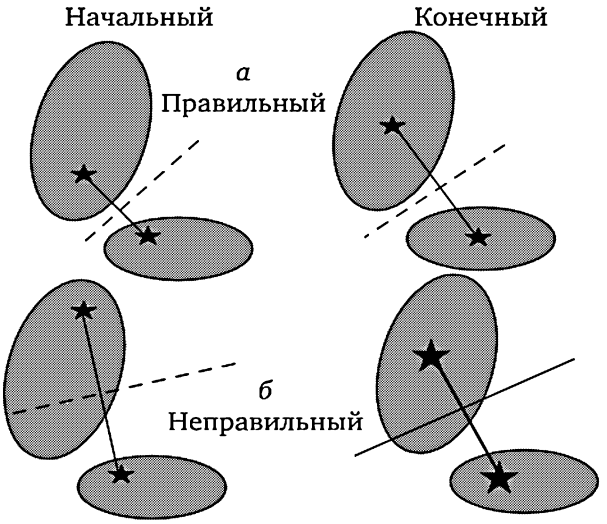


Рис. 5.13. **Ситуация двух четко разделенных кластеров при двух разных начальных инициализациях:**
а — правильное разделение на кластеры, *б* — нет

Еще один пример не оптимальности результатов метода *K*-средних, с использованием только четырех точек, представлен на рис. 5.14.

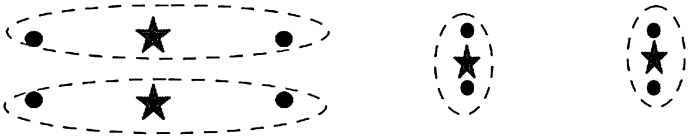


Рис. 5.14. **Пример неудачной работы метода *K*-средних при поиске двух кластеров на множестве четырех точек плоскости:**
на правой стороне рисунка разделение с помощью *K*-средних логично, а на левой — нет (здесь центры обозначены звездами)

Рабочий пример 5.6

Близкие кластеры могут оказаться дальними для *K*-средних

Ниже представлен еще один пример, в котором минимизация критерия квадратичной ошибки приводит к решению, противоречащему интуиции. Тем не менее, оптимальный результат не может быть получен с помощью метода *K*-средних из-за его локальной природы.

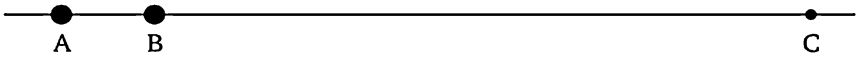


Рис. 5.15. **Вопрос: определить, какие два кластера объединятся по методу *K*-средних, A и B или B и C?**

Рассмотрим случай, представленный на рис. 5.15: три множества объектов, каждое попало в одну и ту же точку, A, B, C. Два из них состоят из 100 объектов в каждом (в точках A и B), а третье состоит из одного объекта в точке C. Примем, что расстояние между A и B равно 2, а между B и C — 10. В таком случае возможно только два случая двукластерного разбиения: (I) 200 точек из A и B в одном кластере и одна точка в другом кластере (C); (II) 100 точек из A в одном кластере и 101 точка — из B и C — в другом. Третье разбиение, состоящее из кластера B и кластера A + C, не может быть оптимальным, потому что кластер A + C более растянут, нежели схожий кластер B + C в случае (II).

Сравним значения критерия *K*-средних на этих двух вариантах при использовании квадрата Евклидова расстояния между точками и центрами.

В случае (I) центр кластера A + B будет расположен в середине интервала между A и B, на расстоянии 1 от каждого. Следовательно, сумма всех квадратичных Евклидовых расстояний равна $200 \cdot 1 = 200$. Так как кластер C состоит только из одного объекта, он ничего не вкладывает в значение критерия *K*-средних, который, таким образом, равен 200.

В случае (II) у кластера B + C есть центр, который является центром масс B и C, т. е. расположен на расстоянии $d = 10/101$ от B. Следовательно, значение критерия *K*-средних равно $100d^2 + (10 - d)^2$, что меньше, чем $100 \cdot (1/10)^2 + 10^2 = 101$, так как $d < 1/10$ и $10 - d < 10$. Вклад кластера A в критерий равен 0, потому что все 100 объектов расположены в A, который, следовательно, является центром данного кластера.

Таким образом, случай (II) более выгоден: значение критерия *K*-средних в данном случае не превышает 101, тогда как в случае (I) оно равно 200. При этом объединение B и C в (II) противоречит интуиции, так как A значительно ближе к B, чем C. Это означает, что критерий метода *K*-средних скорее способствует более равномерному распределению объектов между кластерами, чем объединению наиболее близких объектов.

Но нет худа без добра: параллельный метод *K*-средних ведет именно к интуитивному, хотя и не оптимальному, решению. Действительно, начав с наиболее удаленных точек A и C как исходных центров, мы всегда будем получать вариант (I) как результат!

Рабочий пример 5.7

Устойчивость критерия *K*-средних на нормализованных данных

Существует мнение, что критерий К-средних хорош только для разделения сферических кластеров. Но это не совсем так. Метод успешно работает и на кластерах, сильно отличающихся по форме. Чтобы убедиться в этом, сгенерируем два таких кластера на плоскости так, чтобы один образовывал кружок, а второй — сильно растянутую полосу.

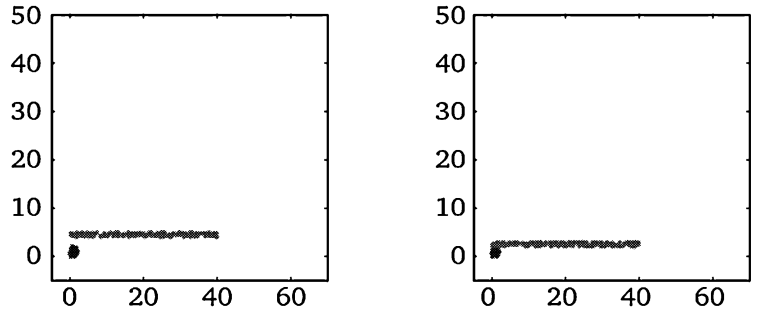


Рис. 5.16. **Двумерное множество, состоящее из двух кластеров разной формы, круга и полосы; координаты центров по оси ординат равны 1 (круг) и 5.5 (полоса) на левом рисунке, и 1 и 3.5, на правом**

Пусть кластер-круг состоит из 100, а кластер-полоса — из 200 объектов. Первый кластер — множество точек (x, y), где каждая компонента имеет Гауссово распределение со средним в 1 и стандартным отклонением 0.5. Второй кластер состоит из точек, равномерно распределенных в прямоугольнике со сторонами длиной 40 и 1, причем этот прямоугольник расположен параллельно оси x в районе значения $y = 5$ (рис. 5.16 слева) или в районе $y = 3$ (рис. 5.16 справа).

Если обрабатывать эти данные как есть, то метод *K*-средних не сможет отделить эти два кластера друг от друга. Как и на рис. 5.15, метод отделит круг и прилегающую часть полосы, порядка 50 объектов, от остальной ее части. Однако ситуация меняется, если перейти к стандартизованным данным. Если вычесть из каждого признака, x и y, его среднее, а потом разделить результат на размах или стандартное отклонение признака, метод *K*-средних сработает. Попробуйте самостоятельно убедиться, что на нормализованных данных левой стороны рис. 5.16 метод *K*-средних, при $K = 2$, начиная с наиболее удаленных друг от друга объектов, приводит в точности к двум сгенерированным кластерам. Конечно, результат меняется для не столь структурированных данных в правой части рис. 5.16. Если нормализовать данные правой части рис. 5.16 стандарт-

ными отклонениями признаков, то структура двух кластеров воспроизводится с очень небольшой ошибкой — только ближайшие 5 точек полосы добавляются к кругу, остальные 195 объектов образуют другой кластер. Несколько худший результат получается, если нормализовать признаки их размахами. В этом случае 32 объекта полосы присоединятся к кругу — ошибка 32/300, около 10 %. Докажите это самостоятельно.

5.2.4. Проблема инициализации *K*-средних

Для инициализации метода *K*-средних нужно задать:
(i) количество кластеров, K, и
(ii) начальные центры, $c = (c_1, c_2, \ldots, c_K)$.

В практических условиях определение каждого из этих параметров может стать проблемой, поскольку они зависят от уровня детализации и типологических особенностей содержательной проблемы, связанной с данными, которые остаются за гранью теории метода *K*-средних. Поэтому некоторые эксперты предлагают оставить пользователю решение вопроса о количестве и расположении центров как возможных прототипов в зависимости от его представлений о характере изучаемого явления. Вместе с тем существуют подходы к оценке количества и расположения начальных центров в зависимости от структуры рассматриваемого множества данных. Эти подходы можно систематизировать в зависимости от того, на каком этапе происходит оценка:

1) до применения метода кластеризации;
2) во время применения метода кластеризации;
3) после применения метода кластеризации.

Методы типа 2 связаны с применением методов иерархической кластеризации, которые последовательно объединяют «малые» кластеры в бóльшие (агломеративный подход) или же, напротив, разделяют «большие» кластеры на меньшие (дивизимный полход). В таких методах проблема выбора числа кластеров и их центров переводится в проблему остановки процесса объединений или разделения кластеров. Мы их не рассматриваем, поскольку они не вкладываются в метод *K*-средних.

Методы типа 3 основаны на многократном применении метода *K*-средних, начиная со случайных K центров, и обработке результатов. Обычно задаются каким-либо числом прогонов, например, $P = 100$, выбирают какой-либо диапазон изменения K, например, от 2 до 20, после чего делают P прогонов метода при каждом заданном K из этого диапазона. Результат каждого такого прогона оценивают критерием квадратичной ошибки $W(S, c)$ (5.15). Обозначим через W_K минимальное из этих P значений $W(S, c)$. Последовательность W_K при разных K из рассматриваемого диапазона значений используется для того, чтобы увидеть, какое K привело бы к лучше-

му значению W_K. К сожалению, лучшее W_K не обязательно является минимумом W_K, потому что минимальное значение критерия квадратичной ошибки может только убывать с ростом K. Имеется ряд предложений о том, как использовать флуктуации в поведении W_K, чтобы оценить «правильное» К. К сожалению, все они не очень надежны. Экспериментальный анализ показал, что одно из самых надежных — это «средне-потолочное» правило Хартигана. Правило Хартигана основано на том предположении, что если имеется K^* кластеров, четко отделенных друг от друга, то для $K < K^*$ кластеров $(K + 1)$-кластерное разбиение по методу K-средних должно походить на K-кластерное разбиение и получаться из него разделением одного из кластеров на две части. При этом W_{K+1} значительно меньше, чем W_K. С другой стороны, при $K > K^*$, и K-кластерное и $(K + 1)$-кластерное разбиения должны являться «правильными» K^*-кластерными разбиениями с несколькими «правильными» случайно разделенными кластерами, причем W_K и W_{K+1} не сильно различаются. На основе этих соображений Хартиган (Hartigan, 1975) предложил рассчитывать индекс

$$H_K = (W_K / W_{K+1} - 1)(N - K - 1), \tag{5.16}$$

начиная с $K = 2$ и последовательно увеличивая K. Здесь N — количество объектов. При увеличении K в качестве оценки K^* берется первое значение K, при котором H_K становится меньше, чем 10. В экспериментах Chiang и Mirkin (2010) правило Хартигана оказалось наилучшим из девяти различных критериев, причем порог 10 оказался не очень чувствительным к 10—20%-ным изменениям.

Рабочий пример 5.8

Индекс Хартигана для выбора количества кластеров

Рассмотрим значения H_K для данных об ирисах и городах побережья, полученные в результате 100 прогонов метода K-средних на данных, нормализованных вычитанием среднего и делением на размах, начиная с K случайных центров (табл. 5.14). Каждый прогон повторяется дважды (см. первый и второй наборы в табл. 5.14) для демонстрации типичных колебаний значений H_K с учетом того, что эмпирические значения W_K могут быть не оптимальными. В частности, в случае второго множества результатов, полученного методом K-средних на данных о городах видно нарушение правила: H_K положительно (поскольку нормально W_K должно уменьшаться с ростом K). Монотонность нарушается, потому что минимальные значения W_K на результатах 100 прогонов необязательно минимальны глобально.

«Естественное» количество кластеров в данных об ирисах по критерию Хартигана не 3, как утверждается ботаниками, а намного больше, 11! В данных о городах наш критерий определил бы 4 «естественных» кластера. Однако следует иметь в виду, что точное значение 10 в правиле Хартигана недостаточно для того, чтобы сделать определенный вывод — оно должно сопровождаться значительным изменением значения H_K.

Таблица 5.14

Значения индекса Хартигана H_K для K, меняющегося от 2 до 11, вычисленные по результатам двух различных 100-кратных прогонов кластеризации, стартующих со случайных K объектов в качестве начальных центров

Наборы данных		K=2	3	4	5	6	7	8	9	10	11
Ирисы	1-й набор	108.3	38.8	29.6	24.1	18.6	15.0	16.1	15.4	15.4	**9.4**
	2-й набор	108.3	38.8	29.6	24.1	18.7	15.4	15.6	15.7	16.0	**7.2**
Города	1-й набор	13.2	10.5	**9.3**	5.0	4.7	3.1	3.0	3.2	3.2	1.6
	2-й набор	13.2	10.5	**9.3**	5.8	4.1	2.5	3.0	7.2	—0.2	1.8

Сильное изменение происходит при $K = 5$. Вероятно, именно это число можно использовать в качестве «естественного» числа кластеров в данных о городах. Аналогично, значительное изменение значения H_K в данных об ирисах происходит при $K = 3$, именно это число является количеством естественных кластеров данного множества.

Задание 5.2. Рассчитайте значения индекса Хартигана по 100 прогонам метода K-средних при каждом $K = 3, 4, \ldots, 11$ как на данных об ирисах, так и на данных о городах английского побережья. Сравните полученные значения с величинами, содержащимися в табл. 5.14.

В целом, проведение нескольких прогонов метода K-средних кажется разумной стратегией, особенно в случае, когда количество объектов не слишком велико. Но с ростом количества объектов до нескольких тысяч или десятков тысяч количество прогонов метода, необходимых для достижения нужного значения W_K, может стать чрезмерно большим. Кроме того, критерий $W(S, c)$ сам по себе имеет ряд существенных изъянов; его следует использовать только в совокупности со стратегией, основанной на знании структуры данных.

Другое направление использования результатов P прогонов метода K-средних при различных K — это агрегирование всех P полученных разбиений при каждом K: чем ближе K к «правильному» значению K^*, тем точнее «усредненное» или «консенсусное» разбиение воспроизводит структуру данных. Это направление активно развивается в последнее время, но не будет здесь рассматриваться, так как выходит за рамки данного текста.

5.3. Пифагорово разложение и аномальные кластеры

5.3.1. Пифагорово разложение разброса данных и дополнительный критерий

Рассмотрим квадратичный критерий (5.14) и возведем «скобки» в квадрат:

$$D(S,c)=\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V}(y_{iv}-c_{kv})^2=\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V}(y_{iv}^2-2y_{iv}c_{kv}+{c_{kv}}^2).$$

Теперь рассмотрим три полученных слагаемых по отдельности. Первое слагаемое дает

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V}y_{iv}^2=\sum_{i=1}^{N}\sum_{v=1}^{V}y_{iv}^2,$$

потому что индекс i здесь «пробегает» по всем кластерам, а значит, и по всему множеству I. Второе слагаемое может быть преобразовано следующим образом:

$$-\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V}2y_{iv}c_{kv}=-2\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V}y_{iv}c_{kv}=-2\sum_{v=1}^{V}\sum_{k=1}^{K}c_{kv}\sum_{i\in S_k}y_{iv}=-2\sum_{v=1}^{V}\sum_{k=1}^{K}N_kc_{kv}^2.$$

Здесь N_k — количество объектов в кластере S_k. Последнее равенство вытекает из определения внутри-кластерного среднего $c_{kv}=\sum_{i\in S_k}y_{iv}/N_k,$, приводящего к тождеству $\sum_{i\in S_k}y_{iv}=c_{kv}N_k$.

Третье слагаемое приводит к выражению

$$\sum_{v=1}^{V}\sum_{k=1}^{K}c_{kv}^2\sum_{i\in S_k}1=\sum_{v=1}^{V}\sum_{k=1}^{K}N_kc_{kv}^2.$$

Сводя полученные выражения вместе, получаем

$$D(S,c)=\sum_{i=1}^{N}\sum_{v=1}^{V}y_{iv}^2-\sum_{v=1}^{V}\sum_{k=1}^{K}N_kc_{kv}^2.$$

Это приводит к Пифагорейскому разложению разброса данных $T=\sum_{i=1}^{N}\sum_{v=1}^{V}y_{iv}^2$ (напомним, что понятие «разброс данных» вводится и обсуждается в разделе 5.1.1., см. рис. 5.3):

$$T=F(S,c)+D(S,c), \tag{5.17}$$

где

$$F(S,c)=\sum_{v=1}^{V}\sum_{k=1}^{K}N_kc_{kv}^2=\sum_{k=1}^{K}N_k\langle c_k,c_k\rangle. \tag{5.18}$$

В равенстве (5.17) величина $D(S, c)$ — это квадратичный критерий (5.14) качества разбиения алгоритма K-средних, а $F(S, c)$ — дополнительный критерий, характеризующий вклад получаемого разбиения в разброс данных. Чем $F(S, c)$ больше, тем разбиение лучше, потому что тем меньше $D(S, c)$. С учетом аддитивного характера критерия (5.18), можно утверждать, что он представляет собой сумму вкладов отдельных кластеров

$$B_k = N_k \langle c_k, c_k \rangle. \tag{5.19}$$

Величина вклада — это произведение численности кластера N_k и величины $\langle c_k, c_k \rangle$, равной скалярному произведению вектора на себя, т. е. квадрату расстояния между вектором c_k и точкой отсчета координат 0. Это придает простой геометрический смысл дополнительному критерию $F(S,c)$. Его максимизация требует, во-первых, больших и, во-вторых, аномальных кластеров.

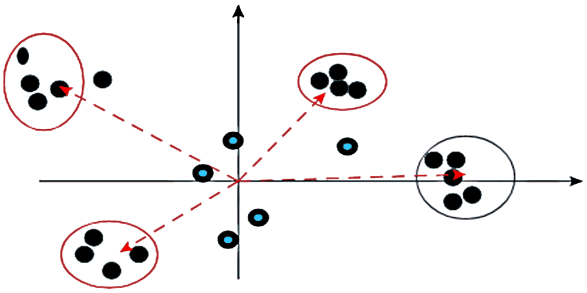


Рис. 5.17. **Геометрическая иллюстрация смысла дополнительного критерия (5.18): как можно дальше от нуля и как можно многочисленнее**

Первое определяется сомножителем N_k, а второе — сомножителем $\langle c_k, c_k \rangle$, требующим увеличения расстояния до начала отсчета. Обычное предварительное преобразование данных переносит точку общего среднего, т. е. вектора $c = (c_v)$, компонентами которого являются средние значения признаков на всем множестве данных, в точку начала координат 0. В этом случае расстояние $d(0, c_k) = \langle c_k, c_k \rangle$ характеризует уровень аномальности кластера, его отличие от общего среднего.

Возникает вопрос — в случае, когда ноль не является точкой общего среднего, то насколько применим термин «аномальность»? Ответ: вполне применим. Это доказывается в утверждении следующего вопроса.

Вопрос 5.15. Докажите, что оптимальное решение по критерию (5.18) не зависит от выбора точки отсчета координат.

Ответ. Рассмотрим произвольный V-мерный вектор $a = (a_v)$ и сдвинем точку отсчета в a, т. е. вычтем a из всех строк матрицы данных, $g_{iv} = y_{iv} - a_v$, При этом, для любого разбиения $S = \{S_1, S_2,$

..., $S_K\}$ и любого его класса S_k центр c_k в S_k заменится на $d_k = c_k - a$. (см. Задание 5.3). Это значит, что на сдвинутых данных величина критерия (5.18) будет равна

$$F(S,c) = \sum_{k=1}^{K} N_k \langle d_k, d_k \rangle = \sum_{k=1}^{K} N_k \langle c_k - a, c_k - a \rangle =$$
$$= \sum_{k=1}^{K} N_k [\langle c_k, c_k \rangle - 2\langle c_k, a \rangle + \langle a, a \rangle] =$$
$$= \sum_{k=1}^{K} N_k \langle c_k, c_k \rangle - 2\langle \sum_{k=1}^{K} N_k c_k, a \rangle + N\langle a, a \rangle =$$
$$= \sum_{k=1}^{K} N_k \langle c_k, c_k \rangle - 2N\langle c, a \rangle + N\langle a, a \rangle,$$

где c — точка общего среднего матрицы данных. Последнее равенство вытекает из тождества $Nc = \sum_{k=1}^{K} N_k c_k$, которое нетрудно доказать. Утверждение данного вопроса следует из того, что в полученном выражении для (5.18) полученные дополнительные слагаемые не зависят от разбиения S.

Таким образом, при использовании дополнительного критерия (5.18) всегда можно рассматривать общий центр данных $c = (c_v)$ в качестве точки отсчета, хотя бы виртуальной, и интерпретировать этот критерий как требование получения как можно более наполненных и как можно более аномальных кластеров.

Задание 5.3. Для данных о компаниях в табл. 5.6 убедитесь, что их сдвиг вычитанием вектора a = (1, 1, 1, 1, 1, 1, 1)/2 приводит к тому же сдвигу центров кластеров в табл. 5.12.

5.3.2. Общность моделей МГК и метода *K*-средних

Матричная факторизация — это представление заданной матрицы Y в виде произведения двух матриц, каждая из которых в некотором смысле проще, чем Y. Такое представление применяется во многих прикладных задачах — в рекомендательных системах, тематическом моделировании, анализе экспрессии генов и пр. Покажем, что критерий метода K-средних позволяет представить его как критерий матричной факторизации.

Напомним, что этот минимизируемый критерий имеет вид

$$D(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} (y_{iv} - c_{kv})^2, \qquad (5.20)$$

где $S = \{S_k\}$ — это совокупность искомых непересекающихся кластеров, включающих объекты $i \in I$, v — признаки, y_{iv} — v-я координата объекта i, представленного вектором $y_i = (y_{iv})$, а $c_k = (c_{kv})$ — центр кластера S_k (k = 1, 2, ..., K). Оказывается, этот критерий — не что иное как критерий наименьших квадратов для модели

$$y_{iv} = \sum_{k=1}^{K} z_{ik} c_{kv} + e_{iv}, \tag{5.21}$$

где $z_{ik} = 1$ для $i \in S_k$, и $z_{ik} = 0$ для $i \notin S_k$.

Действительно, любой данный объект $i \in I$ может принадлежать одному и только одному кластеру S_k, так что $z_{ik} = 1$ только для этого k, а для всех остальных кластеров l, $z_{il} = 0$. Это означает, что $\sum_{k=1}^{K} z_{ik} c_{kv}$ равно одному единственному c_{kv}, т. е. $e_{iv} = y_{iv} - \sum_{k=1}^{K} z_{ik} c_{kv} = y_{iv} - c_{kv}$, так что критерий (5.20) действительно выражает сумму квадратов невязок в модели (5.21).

Равенство (5.21) может быть выражено в матричной форме как:

$$Y = ZC^T + E, \tag{5.22}$$

где матрица $Y = (y_{iv}) - N \times V$ матрица данных, обычно стандартизованная; $Z = (z_{ik})$ — матрица $N \times K$, столбцами которой являются искомые бинарные векторы $z_k = (z_{ik})$ принадлежности объектов i кластерам S_k; $C = (c_{vk})$ — матрица $V \times K$, столбцами которой являются искомые центры $c_k = (c_{kv})$ кластеров S_k; и $E = (e_{iv})$ — матрица $N \times V$ минимизируемых невязок e_{iv}. А это — аппроксимационное уравнение матричной факторизации. Матрицы Z, C здесь более простые, чем Y, поскольку столбцы матрицы Z — 0/1 бинарные взаимно ортогональные векторы, а матрица C имеет размер $K \times V$, значительно меньший, чем размер Y.

Осталось напомнить, что уравнения (5.5)—(5.6) в методе главных компонент имеют ту же самую структуру факторизации матриц, как и (5.22), за исключением того, что матрицы Z, C в этом методе — это матрицы факторных баллов и нагрузок на признаки, соответственно, отнормированные так, что норма k-го столбца как в Z, так и в C, равна $\mu_k^{1/2}$ (k=1, 2, …, K). «Простота» этих матриц определяется тем, что их столбцы попарно взаимно ортогональны, причем «забирают» на себя максимально возможные вклады в разброс данных μ_k^2 (k=1, 2, …, K).

В случае, когда матрица C выбрана оптимально по критерию минимизации суммы квадратов невязок, матрицы ZC^T и E в (5.22) ортогональны друг другу. В этом случае имеет место Пифагорейское разложение квадратичного разброса данных на «объясненную» и «необъясненную» составляющие. В обозначениях матричной алгебры это разложение имеет вид:

$$Tr(Y^TY) = Tr(CZ^TZC^T) + Tr(E^TE) \tag{5.23}$$

где Tr(A) — это след квадратной матрицы A, определяемый как сумма всех ее диагональных элементов $Tr(A) = a_{11} + a_{22} + \ldots + a_{nn}$. В (5.23) крайние слагаемые выражают квадратичный разброс матрицы данных Y или невязок E, соответственно. Величина же в середине — суммарный «объясненный» вклад решения, главных компонент

или кластеров, соответственно. Для дальнейшего анализа удобно представить это слагаемое как $Tr(CZ^TZC^T) = Tr(C^TCZ^TZ)$. Выражение справа получено из левого выражения путем перестановки $A = CZ^TZ$ и $B = C^T$, поскольку, как известно, $Tr(AB) = Tr(BA)$ для любых A и B.

Рассмотрим выражение $Tr(C^TCZ^TZ)$ сначала для метода главных компонент. Поскольку и в Z, и C столбцы взаимно-ортогональны, обе $K \times K$ матрицы, Z^TZ и C^TC — диагональные с квадратами их норм, μ_k ($k = 1, 2, \ldots, K$) на диагонали. Следовательно, произведение этих матриц — диагональная матрица с величинами квадратов μ_k^2 ($k=1, 2, \ldots, K$) на диагонали. Эти значения — в точности вклады отдельных компонент, рассматривавшиеся ранее.

Для ситуации кластер-анализа выражение $Tr(C^TCZ^TZ)$ тоже должно иметь смысл суммы вкладов отдельных кластеров. В силу определения матрицы Z, очевидно, матрица Z^TZ — диагональная с (k, k)-м элементом, равным $\sum_i z_{ik}^2 = N_k$, $k = 1, 2, \ldots, K$. Это означает, что матрица C^TCZ^TZ тоже диагональная, с (k, k)-м элементом равным $N_k<c_k, c_k>$. Это в точности вклад k-го кластера согласно равенству (5.19).

5.3.3. Аномальные кластеры

Далее будет описан метод оценки числа и местоположения кластеров до применения метода K-средних. Можно считать, что метод в какой-то мере «разведает» структуру данного множества объектов. Он основан на последовательном выявлении и удалении так называемых «аномальных» групп в соответствии с критерием (5.18). «Аномальность» понимается как удаленность от некой «реперной» точки.

Реперная точка выбирается как проявление «нормы», «среднего» или «нормального» объекта, не обязательно среди множества наблюденных объектов. Например, при анализе оценок студентов по различным предметам, можно выбрать точку, представляющую «нормального студента», с оценками по контрольным работам и экзаменам, которые считаются нормальными в данной среде, а затем уже выделять группы, наиболее отклоняющиеся от выбранной реперной точки в ту или иную сторону. Или же менеджер банка может определить в качестве «нормы» клиентов с определенным уровнем образования и дохода, а затем выделять «аномальные», отклоняющиеся от нормы группы клиентов.

Аналогично, движущийся робот должен уметь сегментировать окружающую среду в соответствии со своим местоположением (реперная точка) для того, чтобы отделить удаленные объекты как представляющие наименьший интерес. Во многих случаях центр гравитации множества объектов, т. е. его среднее, может быть выбран как реперная точка отсчета.

Использование реперной точки позволяет сравнивать объекты не друг с другом, а именно с этой точкой, что существенно экономит объем вычислений: вместо перебора всех парных расстояний между объектами, есть возможность сконцентрировать усилия на вычислении только расстояний между объектами и реперной точкой, что даёт снижение порядка количества шагов с N^2 до N.

Аномальная группа конструируется как кластер, наиболее удаленный от реперной точки. Процесс начинается с того, что объект, наиболее удаленный от реперной точки, объявляется центром аномальной группы. Затем версия метода K-средних с $K = 2$ применяется к двум центрам: один — это реперная точка, которая не меняется все время процесса вычислений, а второй — это центр аномальной группы, который обновляется согласно стандартной процедуре. А именно, при заданном аномальном центре аномальная группа определяется как множество объектов, которые ближе к этому центру, чем к реперной точке. При заданной аномальной группе ее центр вычисляется как центр масс с помощью нахождения среднего всех входящих в группу объектов. Процедура повторяется до тех пор, пока не сойдется (см. рис. 5.18).

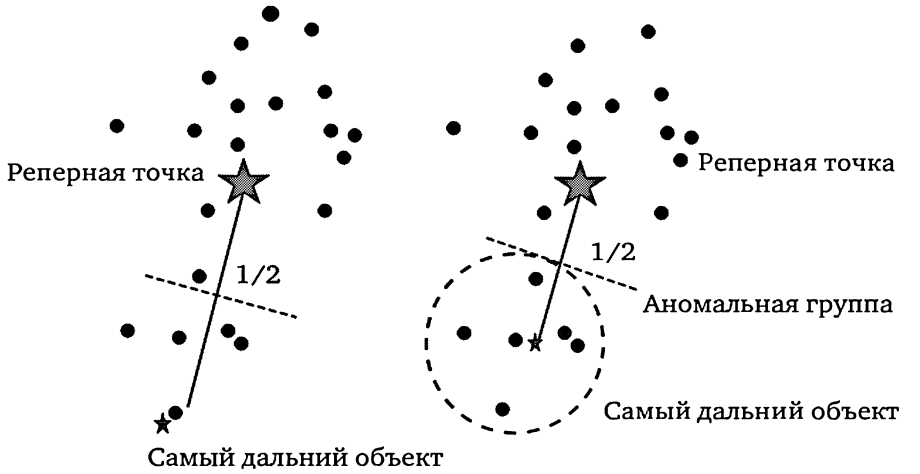


Рис. 5.18. **Извлечение аномального кластера при реперной точке, расположенной в центре масс (большая звезда); малая звезда представляет центр аномального кластера. Первая итерация показана на левой части рисунка, а финальная — на правой**

Таким образом, метод аномальной группы — это версия метода K-средних, в которой:

а) количество кластеров K равно 2;

б) центром одного из кластеров является 0, перенесенный в реперную точку и не меняющийся в процессе итераций;

в) начальный центр аномальной группы выбирается как максимально удаленный от точки 0.

Последнее свойство автоматизирует определение начального центра, исходя из того, что центр должен быть максимально удаленным от точки отсчета 0. Именно эта идея заложена в дополнительном критерии метода *K*-средних (5.18) (см. также рис. 5.19).

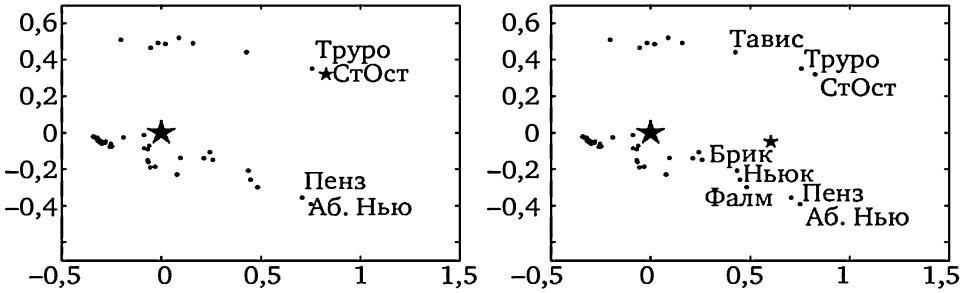


Рис. 5.19. **Первая и вторая итерации построения аномальной группы, визуализированные на плоскости первых главных компонент:** Разброс группы вдоль оси ординат отражает то, что эта ось слишком сильно соответствует признаку «наличие фермерского рынка»: «да» вверху, «нет» — внизу

Рабочий пример 5.9

Аномальная группа прибрежных городов Юго-Западной Англии

Применим метод аномальной группы к данным о прибрежных городах. В качестве реперной точки возьмем среднюю точку, куда и перенесём точку отсчета пространства, 0, а признаки нормируем делением на размах.

Таблица 5.15

Центр аномальной группы в данных о прибрежных городах в реальных и стандартизованных единицах

Центр	Нас	Нш	Тер	Бол	Ба	Ун
Реальный	18484	7.6	3.6	1.1	11.6	4.6
Стандартизованный	0.51	0.38	0.56	0.36	0.38	0.38
Центр	Ав	Ст	Бас	По	Юр	Фр
Реальный	4.1	1.0	1.4	6.4	1.2	0.4
Стандартизованный	0.30	0.26	0.44	0.47	0.30	0.18

Наиболее удаленным от 0 (напомним, 0 — точка средних значений!) является объект 44 (St. Austell); расстояние от него до нуля равно 4.33 (напомним: речь идет о квадрате евклидового расстояния). Объявляем этот объект центром аномальной группы, которую надо построить. Имеются только три объекта: 45, 42 и 41 (Newton Abbot, Penzance и Truro), которые находятся ближе к центру, чем к нулю. Это дает текущий кластер, состоящий из объектов 41, 42, 44 и 45. Вычислим среднее этих четырех городов и повторим операцию обновления кластера для этого нового

центра. Приходим к стабильной группе, состоящей из 8 объектов: 36, 39, 40, 41, 42, 43, 44, 45. Ее центр приведен в табл. 5.15.

Все компоненты стандартизированного центра в табл. 5.15 положительны, причем их бóльшая часть попадает в интервал 0.3—0.5, что близко к максимуму нормализованной шкалы. Это означает, что полученная аномальная группа состоит из городов с высоким уровнем значений показателей — все значения в центре больше, чем средние на всем множестве, на 30 %—50 % размаха значений признаков. Возможно, это связано с тем, что группа включает 8 из 11 городов с населением более 10 000 человек. Остальные три больших города не попали в группу из-за недостатка в них таких характеристик, как больницы и фермерские рынки. Тот факт, что размах шкалы численности населения в данных на порядки превышает остальные, не сильно влияет на вычисления, потому что они проводятся в нормализованных шкалах, при которых общий вклад признака «численность населения» в разброс данных невелик, около 8.5 %.

Данный процесс проиллюстрирован рис. 5.19. Звезды обозначают реперную точку и аномальный центр. Визуально аномальная группа на этом рисунке не очень-то «аномальна», в отличие от аномальной группы на рис. 5.18. А именно, аномальная группа распределена здесь поперек всей плоскости, что противоречит тому, что объекты группы должны быть ближе к ее центру, чем к реперной точке. Причина такой картины — не ошибка в данных, а недостаток нашего визуального отображения. Дело в том, что данная двумерная плоскость представляет все 12 признаков, но представляет их довольно избирательно. Хотя эта плоскость и вносит почтенные 76 % в разброс данных, дело в оси ординат — она слишком хорошо коррелирует с последним признаком, «Наличие фермерского рынка», что и отражается в разделении группы по этой оси на те города, где фермерский рынок есть и на те, где нет.

Самостоятельная работа

5.5. Удалите полученные 8 «аномальных» городов из выборки и на оставшихся данных (о 37 городах), не меняя стандартизации, опять примените метод аномального паттерна. Какой из оставшихся городов соответствует самой дальней от начала координат точке? Попробуйте отобразить полученный кластер на том же самом графике рис. 5.19.

Прежде чем описать математическую модель аномальной группы, дадим формулировку алгоритма выделения аномальной группы.

Предобработка. Определим реперную точку $a = (a_1, \ldots, a_V)$ (если нет явных предположений о «норме», то в качестве a возьмем общее среднее) и стандартизуем таблицу данных сдвигом начала координат в точку $a = (a_1, \ldots, a_V)$. Иными словами, из всех значений каждого признака v вычтем величину a_v ($v = 1, 2, \ldots, V$). Эта реперная точка не меняется при повторном применении алгоритма.

Алгоритм выделения аномальной группы ВАГ

1. **Инициализация аномального центра.** Найдем объект, максимально удаленный от начала координат, 0, и поместим туда начальный аномальный центр, c.

2. **Обновление аномальной группы.** Определим аномальную группу S вокруг c правилом: объект y_i относится к кластеру S, если $d(y_i, c) < d(y_i, 0)$.

3. **Обновление аномального центра.** Вычислим внутригрупповое среднее c' для множества объектов S и проверим, отличается ли оно от предыдущего центра c. Если c' и c не равны друг другу, обновляем центр присвоением $c \Leftarrow c'$ и переходим к шагу 2. В противном случае переходим к шагу 4.

4. **Выдача результатов.** Список группы S и центр c выдаются как результат работы алгоритма.

Нетрудно доказать, что метод аномальной группы, как и метод K-средних, альтернативно минимизирует некий критерий — своеобразную версию критерия метода K-средних $W(S, c)$ (5.15):

$$W(S,c) = \sum_{i \in S} d(y_i, c) + \sum_{i \notin S} d(y_i, 0), \tag{5.24}$$

где S — это искомое подмножество I (а не разбиение как в методе K-средних), а c — центр S в пространстве признаков. При этом метод аномальной группы отличается от метода K -средних при $K = 2$ в следующем: здесь имеется только один центр, c, который и обновляется в алгоритме; другой центр, 0, не меняется никогда и используется только для приписывания ему объектов, не входящих в аномальную группу. Поэтому метод K-средних при $K = 2$ на выходе дает два кластера, а метод аномальной группы — только один, наиболее удаленный от реперной точки, 0.

Можно показать, что критерий (5.24) — не что иное, как критерий наименьших квадратов для модели

$$y_{iv} = \begin{cases} c_v + e_v, i \in S, \\ 0 + e_v, i \notin S, \end{cases} \tag{5.25}$$

где i — номер объекта, v — номер признака, а S — множество объектов искомой аномальной группы, тогда как $c = (c_1, \ldots, c_v, \ldots, c_V)$ — ее центр. Эта модель имеет очень простую структуру: объект либо принадлежит аномальной группе, либо нет, причем в первом случае он должен как можно точнее совпадать с c, а во втором — с 0! (Не забудем, что 0 здесь — это реперная точка, например, точка средних значений всех признаков.) Подобная простая модель характерна и для метода K-средних: в ней каждый объект должен как можно точнее совпадать с центром кластера, которому он принадлежит. Подобные простые модели лежат в основе практически всех популярных методов анализа данных. В частности, путем раскрытия скобок в квадратичных расстояниях d, как это делалось в предыдущем разделе, критерий (5.24) может быть преобразован к следующему виду:

$$W(S,c) = \sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2 - |S| \sum_{v=1}^{V} c_v^2.$$

В этой формуле замечательны оба выражения справа от знака равенства. Первое из них представляет сумму квадратов всех элементов матрицы данных. Эта сумма — уже встречавшийся нам *разброс данных*, составленный из суммы квадратов расстояний от всех объектов до точки отсчета пространства, 0. Второе выражение — произведение численности группы S, $|S|$, на сумму квадратов компонент ее центра c. Эта последняя сумма — не что иное, как квадрат расстояния от c до 0, называемый в математике квадратом нормы c и обозначаемый через $\| c \|^2$. Таким образом, приходим к разложению разброса данных на сумму двух слагаемых

$$T(Y) = |S|\|c\|^2 + W(S, c), \tag{5.26}$$

где $T(Y)$ обозначает разброс данных. Первое слагаемое выражает ту часть разброса данных, которая объяснена аномальной группой, а второе — необъясненную часть. Поскольку метод минимизирует необъясненную часть, а разброс данных — постоянная величина, то одновременно максимизируется объясненная часть. Ее относительная величина, результат деления на разброс данных, выражает вклад аномальной группы в разброс данных. Чем больше вклад, тем лучше эта группа отделена от остальных данных. Выдача вклада может производиться на шаге 4 Алгоритма ВАГ.

5.3.4. Интеллектуальная версия метода *K*-средних

Метод аномальной группы ВАГ может быть использован для автоматического определения и количества кластеров, и начальных центров в методе *K*-средних. Для этого нужно последовательно применять его, сначала ко всему множеству, потом к множествам объектов, остающихся после удаления полученных аномальных групп. Главное — это не менять положения 0 после таких удалений. Мы называем метод *K*-средних, предваренный этим дополнением, «интеллектуальным» методом *K*-средних, или и*K*-средних, потому что он освобождает пользователя от необходимости участия в инициализации.

В методе и*K*-средних пользователю предлагается задать число t, которое является порогом разрешения и используется для того, чтобы отбросить все те аномальные группы, число элементов в которых равно или меньше t. Ничего не отбрасывается только при $t = 0$. При $t = 1$ все аномальные группы, состоящие только из одного объекта, *одиночки*, рассматриваются как не заслуживающие внимания и отправляются обратно в набор данных. Если $t = 10$, все группы, состоящие из 10 или менее объектов, отбрасываются, так как являются слишком маленькими и не заслуживающими внимания при данном уровне разрешения; на больших данных нужны более крупные детали.

Часто при анализе данных аномальные группы-одиночки возникают из-за ошибок в данных, как, скажем, когда человеку приписан возраст 5000 лет из-за пропущенной запятой в числе 50,00. Выделение аномальных групп при этом может служить полезным средством контроля данных.

Рабочий пример 5.10

Итерации метода аномальных групп по данным о прибрежных городах

Многократно примененный к нормализованным на размах признаков данным о прибрежных городах, алгоритм выделения аномальной группы отработал до тех пор, пока не осталось ни одного не сгруппированного объекта, получив в конце концов 12 групп, из которых 5 — одиночки. Эти одиночки — не артефакт, они действительно имеют довольно странные комбинации значений признаков. Например, объект 19 (Лискерд, 7044 жителей) имеет неожиданно большое количество гостиниц (6) и служб такси (2). Список семи неодиночных кластеров представлен в табл. 5.16, в порядке их отделения алгоритмом ВАГ.

Данная структура кластеров не сильно изменится, если, согласно алгоритму и*K*-средних, будет применен метод *K*-средних, инициализированный семью центрами нетривиальных аномальных групп (пять одиночек отправлены обратно в данные). Более того, похожие результаты были получены и при кластеризации набора всех 1300 «фермерских» английских городов, описанных 18ю характеристиками их развития: неодиночные кластеры имеют вполне похожие центры.

Вопрос 5.16. Почему в табл. 5.16 вклад аномальной группы 4, равный 18.6 %, больше, чем вклад предыдущей группы 3, 10.0 %?

Ответ. Из-за гораздо большего количества объектов, 18 в группе 4 против 6 в группе 3. Даже если центр группы 3 значительно дальше от 0, чем центр группы 4 (а именно это — причина того, что группа 3 получена раньше, чем группа 4), вклад рассчитывается с учетом количества объектов (см. формулу (5.18))!

Самостоятельная работа

5.6. Примените метод ВАГ к стандартизованным данным об ирисах (табл. 1.2) в итеративном режиме.

Несмотря на крайнюю упрощенность модели (5.25), во многих случаях аномальные группы достаточно хорошо отображают реальные данные. Их получают последовательно, одну за другой, удаляя объекты, попавшие в аномальную группу, вместе с их вкладами в разброс данных согласно формуле (5.18). Вклады могут быть использованы для остановки процесса вычислений, например, при сильном уменьшении вклада следующей группы.

Ниже дается точная формулировка алгоритма иK-средних(t), где t — порог разрешения, т. е. задаваемое пользователем минимальное количество объектов в аномальной группе, необходимое, чтобы она могла восприниматься как генератор отдельного кластера. В большинстве приложений, работающих с небольшими объемами данных (порядка десятков и сотен объектов), значение t может быть принято равным 2.

Таблица 5.16

Группы, полученные итеративным применением алгоритма выделения аномальной группы на данных о прибрежных городах

Номер группы	Размер	Содержимое	Вклад в разброс, %
1	8	36, 39, 40, 41, 42. 43, 44, 45	35.1
3	6	10, 15, 23, 25, 27, 32	10.0
4	18	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 19, 21	18.6
5	2	37, 38	2.4
6	2	20, 29	1.6
8	2	31, 35	1.7
11	2	22, 24	1.2

Алгоритм иK-средних(t)

0. Настройка. Предобработка и стандартизация данных. Полагаем $k = 1$ и $I_k = I$ — соответственно, номер кластера и множество объектов, на котором ищется аномальная группа. Выбирается реперная точка a и вычитается из всех строк таблицы данных (сдвиг точки отсчета в a).

1. Аномальная группа. Метод выделения аномальной группы применяется к I_k для нахождения k-той аномальной группы S_k и ее центра c_k.

2. Условие остановки. Если условие остановки (см. ниже) не выполняется, то удаляем S_k из I_k, производим замену $k \Leftarrow k + 1$ и $I_k \Leftarrow I_k - S_k$, после которой переходим на шаг 1. Если же условие остановки верно, то переходим к шагу 3.

3. Отбрасывание малых кластеров. Удалим все найденные группы, состоящие из $t - 1$ или менее объектов. (В зависимости от желания пользователя, объекты из удаленных групп либо возвращаются в данные для последующей кластеризации, либо удаляются из данных вовсе как «нехарактерные» объекты.) Обозначим количество оставшихся кластеров через K, а их центры через c_1, c_2, …, c_K.

4. Метод K-средних. Применяем метод K-средних, используя c_1, c_2, …, c_K в качестве начальных центров.

Условие остановки на шаге 2 может включать в себя любые или даже все из следующих условий:

(А) В множестве I не осталось объектов, не попавших в ту или иную аномальную группу, т. е. $S_k = I_k$.

(Б) Большой суммарный вклад: суммарный вклад первых аномальных групп в разброс данных достиг изначально заданного порога, например, 50 %.

(В) Малый вклад отдельной аномальной группы: вклад S_k в разброс данных слишком мал; например, сравним со средним вкладом одного объекта, $T(Y)/N$, где $T(Y)$ — разброс данных.

(Г) Количество кластеров достигло заранее заданного значения K.

Условие (А) выполняется часто, особенно если в данных есть «естественные» кластеры, сильно отличающиеся по вкладу в разброс данных. Условия (Б) и (В) могут быть рассмотрены в качестве уровней гранулярности анализа, задаваемых пользователем. В отличие от (Г), они основаны на структуре анализируемых данных, а не априорных представлениях.

Рабочий пример 5.11

Кластеризация выборки из одномерного нормального распределения методом и*K*-средних

Сгенерируем одномерную выборку X из 280 точек из Гауссового распределения с параметрами $N(0, 10)$ (см. рис. 5.20). Данное множество представлено в приложении (табл. А5.2). Многие бы сказали, что оно составляет единичный Гауссов кластер, так что нет никакого смысла искать в нем какие-либо кластеры. Мы применяем алгоритм и*K*-средних, чтобы проиллюстрировать принцип, заложенный в этом алгоритме.

Несмотря на симметричность Гауссова распределения, выборка несколько смещена в отрицательную сторону: среднее равно –0.89, а не 0, а медиана вообще равна –1.27. При этом максимальное расстояние до среднего приходится на максимальное положительное значение 32.02, а не на «максимально отрицательное» –30.27.

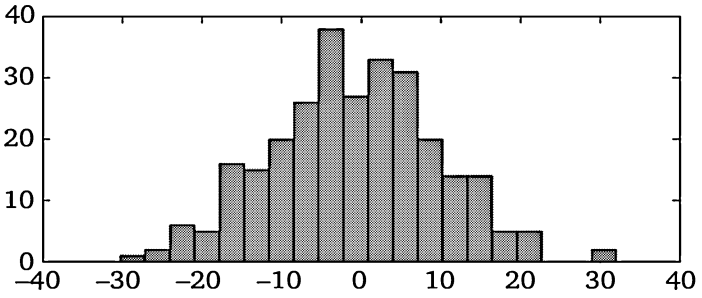


Рис. 5.20. **Гистограмма выборки из 280 элементов, сгенерированных командой 10*randn в среде программирования MATLAB из нормального распределения с параметрами N(0,10)**

Вычисление аномальной группы начинается с наиболее удаленного, т. е. максимального, значения 32.02 и в итоге составляет 83 объекта в интервале между 5.28 и максимумом. Такое разделение согласуется с тем, что происходит на практике. Например, для роста молодых мужчин, призываемых на военную службу, гистограмма тоже имеет колоколообразный вид, т. е. может рассматриваться как приближенно нормальная. Однако «объекты», соответствующие разным сторонам «колокола», отличаются в разных реальных ситуациях. Например, люди невысокого роста не смогут выполнить ряд специфических задач, тогда как слишком высоким будет трудновато в тесных помещениях.

Итеративное применение метода выделения аномальных групп отсекает крайние фрагменты распределения и его остатков. Аномальные группы, в порядке их формирования, включая внутри-групповые средние и вклады в разброс данных, представлены в табл. 5.17.

Таблица 5.17

Характеристики кластеров, полученных итеративным применением метода выделения аномальной группы к выборке из нормального распределения из табл. A5.2

Порядок получения	Размер группы	Лев.	Прав.	Среднее	Вклад
1	83	198	280	11,35	34,28
2	70	1	70	–14,32	46,03
3	47	71	117	–5,40	4,39
4	41	157	197	2,90	1,11
5	18	118	135	–2,54	0,38
6	10	147	156	0,27	0,002
7	6	136	141	–1,42	0,039
8	2	145	146	–0,49	0,002
9	3	142	144	–0,77	0,006

Кластеры представлены в порядке их получения, наряду с их размерами, левой и правой границами номеров объектов в порядке возрастания, а также средними и вкладами в разброс данных.

Последние из полученных групп расположены вокруг среднего значения, и они совсем невелики. Можно также видеть, что вклад следующей группы может быть больше, чем предыдущей, из-за локальности алгоритма выделения аномальной группы, несмотря на критерий, требующий, чтобы на каждой итерации отыскивался кластер с наибольшим вкладом. Суммарный вклад девяти кластеров составляет около 86 %, при этом последние пять кластеров, фактически, ничего не добавляют к этому.

Задание 5.3. Проведите подобное вычисление самостоятельно, сгенерировав 500 гауссовых значений. **Подсказка:** генерация выборки x, состоящей из 500 значений из Гауссова распределения с ну-

левым математическим ожиданием и стандартным отклонением 10, может быть осуществлена в MATLAB с помощью команды

```
x = 10*randn(500,1).
```

Метод и*K*-средних является достаточно гибким, позволяя отделять как выбросы, так и промежуточные объекты, а также и «трясину» объектов, находящихся рядом с общим средним. Например, можно удалить все малые аномальные группы перед тем, как окончательно применять метод *K*-средних. В некоторых задачах, например, в структурировании множества регионов для лучшего планирования или мониторинга или анализа климатических изменений не следует исключать из рассмотрения ни один объект. В других задачах, таких как формирование обзора корпуса документов, аномальные тексты, сильно отличающиеся от остальных, могут быть полностью удалены перед окончательным кластер-анализом.

В ряде экспериментов с перекрывающимися гауссовыми кластерами, описанными в Chiang and Mirkin (2010), метод и*K*-средних показал себя вполне конкурентоспособным и оказался лучше многих других способов выбора *K*, используемых в литературе.

Проект 5.1. Действительно ли метод главных компонент очищает структуру данных?

Существует мнение, что структура многомерных данных лучше выявляется, если данные сначала «очистить» при помощи метода главных компонент, используя всего несколько главных компонент вместо исходных признаков. Это мнение не является общепринятым. Одно из возражений относительно целесообразности применения метода главных компонент дано в учебнике А. Крыштановского (2008): в нем приводится пример структуры данных, которая становится менее выраженной при переходе к главным компонентам. Попробуем проверить, так ли это.

В примере Крыштановского рассматриваются данные, состоящие из двух Гауссовых кластеров, каждый из которых содержит пятьсот 15-мерных элементов. Первый кластер определяется с помощью следующих команд в MATLAB:

```
>>b(1:500,1) = 10*randn(500,1); % расчет первой координаты
>>b(1:500,2:15) = repmat(b(1:500,1),1,14)+20*randn(500, 14); %
расчет 14 координат
```

Первая координата кластера является Гауссовой со средним 0 и стандартным отклонением 10, тогда как остальные четырнадцать переменных добавляют к данной гауссовой величине еще одну со средним 0 и стандартным отклонением, равным 20. Поэтому это множество — выборка из 15-мерного Гауссова распределения с диагональной матрицей ковариации, центр которой находится в начале координат пространства, со стандартными отклонениями

признаков, равными 22.36, т. е. квадратному корню из $10^2 + 20^2$, за исключением первого признака, стандартное отклонение которого равно 10.

Элементы второго кластера генерируются подобным образом путем создания 500 следующих строк в той же матрице:

```
>>b(501:1000,1) = 20 + 10*randn(500,1);
>>b(501:1000,2:15) = repmat(b(501:1000,1),1,14)+20*ra
ndn(500,14)+10;
```

Первый признак имеет центр в точке 20, а остальные — в точке 30. Стандартные отклонения — такие же, как в первом кластере.

Так как стандартные отклонения превышают расстояния между центрами, кластеры нелегко различить: см. рис. 5.21, показывающий облако данных на плоскости двух первых главных компонент.

Метод и*K*-средних в применении к этим, предварительно центрированным и нормализованным по размаху данным, находит не два, а гораздо большее число кластеров, 13, при пороге разрешения, равном $t = 1$. Если же увеличить порог разрешения до $t = 200$, т. е. исключить все аномальные группы меньшего размера, то метод выдает два кластера, которые отличаются от сгенерированных на 96 элементов (см. первый столбец табл. 5.18, показывающей результаты вычислений), так что суммарная ошибка равна 9.6 %. Тот же метод, примененный к данным, центрированным и нормализованным с помощью стандартных отклонений, приводит к 99 ошибкам; относительно умеренное повышение доли ошибок, с учетом специфики сгенерированных данных.

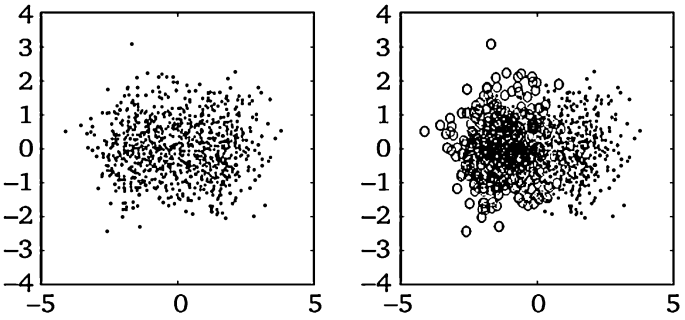


Рис. 5.21. **Данные, состоящие из двух кластеров, сгенерированных, как описано выше, после центрирования:**
данные представлены в виде точек на плоскости двух первых главных компонент (на левом рисунке); точки второго кластера представлены в виде кружков на правом рисунке

Так как Крыштановский (2008) работал с четырьмя главными компонентами, мы также вычислим первые четыре главные компоненты. Например, случай, когда данные предварительно центриро-

ваны по средним и нормированы по размаху, реализуется следующими командами MATLAB:

```
>>n = 1000; br = (b-repmat(mean(b),n,1))./ repmat(max(b)-min(b),n,1);
% стандартизация
>>[zr,mr,cr] = svd(br);
% сингулярное разложение матрицы данных
>> zr4  = zr(:,1:4);
% те же объекты, признаки – первые 4 главные компоненты
```

При нормировании по стандартным отклонениям используются те же команды, только в первой строке вместо max(b)-min(b) нужно поставить std(b). Четыре главные компоненты составляют порядка 66 % разброса данных как в первом, так и во втором случаях.

Данные, нормированные стандартными отклонениями, особенно важны в данном контексте, потому что Крыштановский [9] использовал стандартную версию метода главных компонент, которая основана на матрице корреляции между признаками. Выше использована версия, основанная на так называемом сингулярном разложении матрицы данных. Она более гибкая, чем стандартная версия, так как позволяет использовать различные предварительные преобразования данных. Две версии дают одинаковые результаты в случае, когда данные центрированы по средним и нормализованы по стандартным отклонениям, т. е. подвергнуты преобразованию z-скоринг. Метод иK-средних, примененный к каждому из трех множеств данных ((а) центрированные ненормализованные, (б) нормализованные по размаху, (в) нормализованные по стандартному отклонению) при пороге разрешения 200, показал достаточно согласованные результаты (см. табл. 5.18).

Таблица 5.18

Количество ошибок метода иK-средних при разных стандартизациях исходных данных и соответствующих четырех главных компонентах

Данные	Исходные данные		Четыре главные компоненты данных		
Нормализованные	Размахом	Стандартным отклонением	Ненормализованные	Размахом	Стандартным отклонением
Кластер 1	44	43	37	51	47
Кластер 2	52	56	47	47	45
Всего	96	99	84	98	92

Эти результаты показывают, что способ стандартизации значительно больше влияет на ошибку, чем факт очищения данных с помощью главных компонент. Они скорее подтверждают, нежели

опровергают, идею о том, что при использовании главных компонент «очищенные» данные лучше структурированы. Например, для стандартной версии главных компонент использование исходных данных приводит к 99 ошибкам, а «очищенных» — к 92 ошибкам. Неудовлетворительные результаты использования главных компонент Крыштановским (2008), возможно, связаны с тем, что применялась версия метода *K*-средних со случайными инициализациями для всех случаев нормализации без исключений, что помешало найти «правильную» пару начальных центров, в отличие от метода и*K*-средних, который такую пару находит автоматически, как противоположные друг другу «аномальности».

5.3.5. Правила интерпретации кластеров через их центры

Если кластеры получены не в чисто прикладных целях, а как способ обогащения теоретических представлений о явлении, к которому относятся данные, кластеры следует интерпретировать. К сожалению, эта проблематика мало интересует исследователей.

В отличие от случая главных компонент, проблема интерпретации кластеров — как и любых других групп объектов — допускает более или менее формализованную трактовку и, значит, доступна для автоматизации. Дело в том, что признаки в кластере представлены очень понятными средними значениями, в отличие от нагрузок на главные компоненты, имеющих значительно менее наглядный смысл весовых коэффициентов признаков.

Каждый кластер следует интерпретировать отдельно. Для этого используются признаки, выбранные пользователем. При этом признаки могут браться «со стороны», т. е. признаки, не участвовавшие в формировании кластеров, тоже вполне приемлемы. Важно: для интерпретации используются не стандартизованные, а исходные значения признаков. В соответствии с моделью метода к-средних, для интерпретации кластера используются средние значения признаков внутри этого кластера.

Таблица 5.19

Расчет относительных разностей средних как средства интерпретации для группы объектов первого таксона Т1 в данных Ирис

	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп
Среднее в Т1, cm	5.006	3.428	1.462	0.246
Общее среднее, cm	5.843	3.057	3.758	1.199
СреднТ1 – Общее, cm	–0.837	0.371	–2.296	–0.953
(СреднТ1 – Общее)/Общее, %	–14.33	12.12	**–61.10**	**–79.49**

Главный совет: взятые сами по себе, средние внутрикластерные значения не очень информативны. Информативными они становят-

ся при сравнении с общим средним значением признака. Рассмотрим, например, первый таксон, T1 — первые 50 объектов в данных Ирис (табл. 1.9, см. табл. 5.19).

В табл. 5.19 наибольшее впечатление производят нижние строки — абсолютная разность внутригрупповых и общих средних, и особенно — относительная разность в последней строке. Мы видим, что группа T1 существенно уступает общим средним по признакам длины и ширины лепестка — на 61 % и 79.5 %, соответственно. Чтобы помочь интуиции, попробуем перевести эти значения в наглядный контекст — например, средний вес мужчины в России можно принять за 75 кг. Вес на 30 % выше — порядка 100 кг (толстые), на 30 % ниже — порядка 50 кг (очень худые). То есть можно принять, что значения внутригрупповых средних, которые выше или ниже, чем общие средние на 30 %, являются характеристическими для кластера; их следует специально выделять при интерпретации. В табл. 5.19 они выделены жирным.

Уточним используемые понятия.

При заданном кластере k и количественном признаке v, относительная разность групповых средних и общего среднего, в процентах, определяется формулой

$$d_{kv} = 100[c_{kv}/c_v - 1], \tag{5.27}$$

где c_{kv} — среднее значение признака v в кластере k, а c_v — его среднее на всем множестве данных.

Это понятие может быть перенесено на бинарные признаки с учетом того, что общее среднее «фиктивного» 0/1 признака v равно p_v, т. е. доле кластера в соответствующей категории, а внутри-кластерное среднее $p(v/k) = p_{kv}/p_k$ — не что иное, как условная вероятность v при условии k, причем p_{kv} — это доля объектов, которые попали в кластер k и имеют категорию v.

При заданных кластере k и категории v относительная разность, в процентах, определяется формулой

$$q_{kv} = 100[p_{kv}/(p_k p_v) - 1]. \tag{5.28}$$

Обратим внимание, что относительная разность в (5.28) есть не что иное как коэффициент Кетле (3.20), характеризующий связь между категориями. В данном случае, это изменение вероятности категории v при условии, что объект принадлежит k-му кластеру. Таким образом, мы видим еще одно применение индексов Кетле — для интерпретации кластеров в терминах неколичественных категорий.

Теперь можно сформулировать правила интерпретации кластеров более четко.

Правила интерпретации кластера

1. **Отбор признаков «плюс» и «минус».**

Для каждого из рассматриваемых признаков рассчитайте относительную разность между внутри-кластерным и общим средним. Определите множество V^+ тех признаков, для которых относительная разность превышает 30 %, и множество V^- тех признаков, для которых относительная разность меньше, чем –30 %.

2. **Описание кластера.**

Опишите кластер утверждением, что значения признаков из V^+ значительно выше среднего, а значения признаков из V^- значительно ниже среднего. Если оба множества, V^+ и V^-, пусты, напишите, что кластер ничем особым от общего среднего не отличается.

3. **Концептуализация.**

Посмотрите, имеется ли что-то общее в описаниях кластеров, и опишите это общее — это и будет концептуализацией описаний кластеров. Если не получается, не огорчайтесь: возможно, в следующий раз вы окажетесь удачливее.

Рабочий пример 5.12

Интерпретация таксонов в данных Ирис

Применим сформулированные правила для интерпретации таксонов Т1, Т2 и Т3, т. е. списков 1—50, 51—100 и 101—150 в данных Ирис.

Таблица 5.20

Ступени интерпретации группы таксонов Т1, Т2 и Т3 в данных Ирис

№	Этап	Группа	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп
1	Средние	Т1	5.006	3.428	1.462	0.246
		Т2	5.936	2.770	4.260	1.326
		Т3	6.588	2.974	5.552	2.026
		Общее	5.843	3.057	3.758	1.199
2	Относительные разности, %	Т1	–14.330	12.124	**–61.100**	**–79.489**
		Т2	1.586	—9.398	13.358	10.561
		Т3	12.744	—2.726	**47.738**	**68.927**
3	Интерпретирующие признаки	Т1	$V^+ = \varnothing$, $V^- =$ {ДлЛеп, ШиЛеп}			
		Т2	$V^+ = \varnothing$, $V^- = \varnothing$			
		Т3	$V^+ =$ {ДлЛеп, ШиЛеп}, $V^- = \varnothing$			
4	Интерпретация	Т1	Лепестки значительно короче и уже среднего			
		Т2	Ничем особым не выделяется			
		Т3	Лепестки значительно длиннее и шире среднего			

Окончание табл. 5.20

№	Этап	Группа	ДлЧаш	ШиЧаш	ДлЛеп	ШиЛеп
5	Концептуализация	Т1	**Маленькие** лепестки			
		Т2	**Среднее**			
		Т3	**Большие** лепестки			
6	Дальнейшая концептуализация	—	В описаниях используются только измерения лепестка, но не чашелистика			

В строках табл. 5.20 представлены отдельные этапы этой работы:

— расчет внутри-кластерных средних (1);

— вычисление относительных разностей (5.19) (2);

— вычленение множеств интерпретирующих признаков V^+ и V^- (3);

— интерпретация, т. е. описание отклонений от среднего по признакам из V^+ и V^- (4);

— концептуализация, т. е. пересказ интерпретации с подчеркиванием того общего, что содержится в описаниях кластеров (5);

— супер-концептуализация: Дальнейшее обобщение (6).

В данном случае «супер-обобщающим» является утверждение — «В интерпретирующих описаниях используются только измерения лепестка, но не чашелистика.» Что оно означает? Не знаю. Требуется дополнительное знание. Понять, в чем дело — задача специалиста по ирисам. Со стороны анализа данных ничего дополнительного не предвидится.

Задание 5.4. Кластеры в данных о прибрежных городах и их интерпретация.

Рассмотрим данные об английских прибрежных городах из табл. 1.11 и результаты аномального кластеринга этих данных, стандартизованных вычитанием из каждого признака его среднего значения с последующим делением на размах, в табл. 5.16. Задача — продолжить кластер-анализ по методу иК-среднего с пороговым размером минимального кластера $t = 3$, при котором оставляются только аномальные кластеры размера 3 или более, а остальные отправляются назад в множество данных как пока что не кластеризованные. При этом остается всего три кластера. Окончательные результаты представлены в табл. 5.21. Как видим, средний кластер из 6 объектов не изменился; все «лишние» объекты распределились между первым и третьим кластерами.

Несмотря на то, что данные нормализованы, мы видим, что кластеры в основном следуют численности населения: первый кластер включает 10 самых больших городов, а третий — все остальные, за исключением 6 более или менее средних городов, образующих второй кластер.

Таблица 5.21

Кластеры прибрежных городов и относительные разности между внутри-кластерными и общими средними по отдельным признакам для их интерпретации. Значения относительной разности, выходящие за порог ±30 %, выделены жирным

Кластер		Нас	Ншк	Тер	Бол	Бан	Уни	Авт	Стр	Бас	Поч	ЮрУ	ФеР
30 36 37 38 39 40 41 42 43 44 45	Среднее # = 11	16510	6,55	3,18	1,18	9,91	4,00	4,00	0,91	1,27	5,27	1,27	0,27
	Отн. разн., %	**125.0**	**117**	**131.0**	**195.0**	**130.0**	**107.0**	**96.0**	**309.0**	**160.0**	**101.0**	**97.0**	**36.0**
10 15 23 25 27 32	Среднее # = 6	5268	2,17	0,83	0,50	4,67	1,83	1,67	0,00	0,50	1,67	0,67	1,00
	Отн. разн., %	–28.0	–28	**–40.0**	25.0	8.0	–5.0	–18.0	**–100.0**	2.0	**–36.0**	3.0	**400.0**
1 2 3 4 5 6 7 8 9 11 12 13 14 16 17 18 19 20 21 22 24 26 28 29 31 33 34 35	Среднее # = 28	4200	1,82	0,79	0,07	2,04	1,14	1,36	0,00	0,18	1,79	0,39	**–0.00**
	Отн. разн., %	**–43.0**	**–40**	**–43.0**	**–82.0**	**–53.0**	**–41.0**	**–34.0**	**–100.0**	**–63.0**	**–32.0**	**–39.0**	**–100.0**

Этот второй кластер отличается тем, что во всех его членах имеется Фермерский рынок, а остальные показатели — на среднем уровне, хотя и обнаруживается нехватка по трем из них: наличию терапевтов (возможно, компенсируемому присутствием больниц), нет строительных магазинов и маловато почтамтов. Кластеры 1 и 3 очень просто устроены. В первом наблюдается избыток всего, а в третьем — напротив, нехватка всего.

Кстати говоря

8. Классификация (разделение совокупности на группы)

8.1. — Я пью водку только в двух случаях. Первый — когда есть соленые огурцы. Второй — когда нет солёных огурцов.

8.2. Профессор математики студентам:

— Я делю людей на три типа: тех, кто умеет считать, и тех, кто не умеет...!!!

8.3.

Бежит слепой зайчик по тропинке и спотыкается об змею. Обращается к змее:

— Извините, я слепой и не видел вас, из-за слепоты, я даже не знаю кто я.

Змея отвечает:

— Я тебя понимаю. Я тоже слепая и не знаю кто я.

Зайчик предлагает:

— Давай ощупаем друг друга и определим кто мы...

Змея ощупывает зайчика и говорит:

— Ты мягкий, пушистый, с коротким хвостом и длинными ушами. Ты, наверное, зайчик.

Зайчик в свою очередь ощупал змею и говорит:

— Ты холодная, скользкая, у тебя маленькая голова и очень длинный язык. Ты, наверное, менеджер или руководитель проекта...

8.4. Психотерапевт: Давайте, определим причину вашего невроза. Скажите, что у вас за работа?

— Я сортирую апельсины.

— Так, так, расскажите поподробнее.

— Вниз по желобу скатываются апельсины, я стою внизу и должен их сортировать. В одну корзину большие, в среднюю поменьше и в маленькую — маленькие.

— Но в чем причина нервничать на этой спокойной работе?

— Спокойной? Да поймите же вы, наконец, что целый день я должен принимать решения, решения, решения!

8.5. Была с мужем на рыбалке. Узнала новые виды рыб. Рыбу-стерву и рыбу-мразь поймать практически невозможно. А вот рыба-красава ловится хорошо.

8.6. В метро меня бесят две категории людей: те, кто останавливается перед указателями, и приходится их отталкивать, чтобы пройти, и те, кто пихается, когда стоишь и спокойно читаешь указатель.

9. Обобщение

9.1. Американский мальчик, сидя со своим отцом за большим обеденным столом и прорабатывая уроки по изучению общественных отношений:

— Пап, а сколько людей работает в правительстве США?

Отец:

— Примерно половина из них.

9.2. Человек жалуется невропатологу на бессонницу. Врач внимательно слушает, потом достает целую кучу разных клякс, вынимает одну.

— Так, мистер Джонсон, что Вы видите на этой картинке?

Джонсон:

— Секс.

Доктор смотрит на кляксу, пожимает плечами, роется и достает новую:

— А что вы думаете об этой картинке?

Джонсон:

— Секс.

Врач делает пометки в блокноте, просматривает список обычных ответов, делает новые пометки:

— Ну хорошо. А вот эта клякса?

— Секс!!!

Доктор сурово смотрит на пациента:

— Мистер Джонсон, а вам не приходило в голову, что у вас ненормально сильная сексуальная озабоченность?

— У меня? Вот это да, доктор! Вы же сами пристаете ко мне с этими вашими грязными картинками...

9.3. Всероссийское радио начинает передачи и сообщает:

— В Москве — 15 часов, в Свердловске — 16, в Томске — 17, в Иркутске — 18, во Владивостоке — 23, в Петропавловске-Камчатском — полночь.

Слушатель:

— Ну и бардак в стране!

9.4. На вопрос «Пользуетесь ли вы интернетом?» утвердительно ответили 100 % россиян. Таков результат опроса, проведенного недавно в интернете.

9.5. Статистика показывает, что на каждого мужчину свыше 85 лет приходится по 7 женщин. Но, увы, это уже слишком поздно!..

9.6. — Статистика установила, что каждый двухсотый мужчина имеет рост выше двух метров. Вы знаете об этом?

— Еще бы мне не знать! Он каждый раз сидит передо мной в кинотеатре.

9.7. Я окончил курсы по скоростному чтению. Получил диплом. На экзамене прочитал роман Сэлинджера «Над пропастью во ржи» за 20 минут. Там о Соединенных Штатах Америки.

9.8. Статистик утонул, пересекая реку, средняя глубина которой была один метр.

Заключение: место анализа данных

Необходимость объяснения паттернов для принятия решений

Анализ данных не обязательно связан с их серьезными преобразованиями. Часто простые действия, например, расчет и сравнение частот или средних значений, могут привести к немедленным выводам и соответствующим действиям. Особенно это относится к большим, да и малым, организациям, где основная трудность зачастую не столько этап анализа, сколько этап сбора данных. В этом плане трудно переоценить возможности, открываемые цифровыми системами вычислений и связи. Как-то я прочитал в газете репортаж о том, как одному из муниципалитетов Лондона после нескольких неудачных попыток удалось приобрести эффективную компьютерную систему. Эта система позволяет поддерживать связь со всеми служащими, занятыми выпиской штрафов за неправильную парковку автомобилей, и видеть плоды их деятельности в режиме реального времени — а этих служащих более двух сотен. Оказалось, например, что после обеда их производительность (количество выписанных штрафов) резко снижалась — были приняты дисциплинарные меры, и ситуация выправилась. Больше штрафов — больше доход муниципалитета. Выявлены служащие, чьи штрафные взыскания наиболее успешно оспаривались автомобилистами, что позволило провести дополнительное обучение именно этого персонала. Более того, были выявлены места, где к парковке невозможно или нет смысла придираться, что позволило сократить штат на 5—10 % и, значит, сэкономить на зарплате. Учитывая, что парковочные штрафы — один из важнейших источников «живых» денег в муниципалитете, можно понять, почему так доволен руководитель службы.

В более сложных случаях паттерны, получаемые в результате анализа данных, не столь очевидны, а иногда и просто противоречат существующим представлениям. В таких случаях лица, принимающие решения, не спешат действовать и, конечно же, их за это нельзя упрекнуть. Приведем три примера из литературы, чтобы сделать проблему нагляднее.

Доктор Сноу и вспышка холеры

Существует несколько версий популярной истории о борьбе доктора Джона Сноу с вспышкой холеры в Лондоне в 1854 году. Будем придерживаться версии, изложенной в статье Броди и др., опубликованной журналом Ланцет в 2000 г.[1] Согласно этой версии, доктор Сноу придерживался того взгляда, что холера передается через воду, в то время как господствовавшая теория утверждала, что холера передается миазмами по воздуху. Когда летом 1854 г. в Лондоне в очередной раз вспыхнула холера, сильно затронув, в частности, квартал лондонского района Сохо, Д-р Сноу обнаружил, что характер смертности в этом районе полностью соответствовал его взглядам.

На рис. 3.1 представлен фрагмент плана квартала, который он подготовил позже, чтобы объяснить и защитить свои взгляды. На плане помечены черными полосками случаи смерти в домах — они концентрируются вокруг водной колонки, представленной черным кружком, на Брод-стрит.

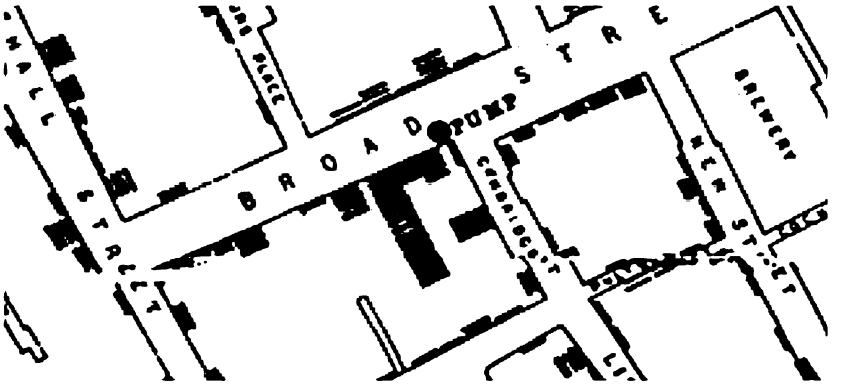


Рис. 3.1. **Фрагмент плана Д-ра Сноу, показывающего, что смертные случаи действительно концентрируются вокруг водоколонки на Брод-стрит**

Д-р Сноу убедил священника местной церкви, что эта колонка и есть источник смертельной заразы. Они сняли и унесли ручку колонки, чтобы из нее больше нельзя было брать воду. На этом обычно рассказ энтузиастов анализа и визуализации данных заканчивается. Увы, действительность оказалась менее благоприятной. Ручку колонки на следующий день пришлось вернуть на место, и в последующих дебатах доктору Сноу не удалось убедить руководство медицинской службы города в правильности своей теории. Холера и после этого случая возвращалась в Лондон еще пару раз, пока не окрепла и не победила идея, что причина болезней – бактерии. В частности, в 1883 г. Р. Кох обнаружил источник заразы —

[1] *Brody H., Rip M. R., Vinten-Johansen P., Paneth N., Rachman S.* Map-making and myth-making in Broad Street: the London cholera epidemic. 2000. P. 64—68.

холерный вибрион, что поставило точку на теории миазмов. (Любопытно, что первым холерный вибрион открыл итальянский ученый Ф. Пачини в том же 1854 г., но это открытие было проигнорировано научным сообществом как полная чепуха.)

Рак почек и малонаселенные штаты в США

Следующий материал основан на работе Х. Вайнера и Х. Цверлинга[1].

Около двадцати лет назад в США было обнаружено, что наибольший процент сильных учеников учится в малых, а не больших школах, что послужило для многих родителей сигналом к тому, что надо определять своих детей именно в малые школы. Вайнер и Цверлинг выступили с объяснением этого феномена, из которого вытекало, что малые школы также содержат и наибольший процент слабых учеников, так что нет никакой причины стремиться именно в эти школы. Поскольку данные о слабых учениках не собирались, они проиллюстрировали свою точку зрения на других, более доступных, материалах.

Эти материалы относятся к заболеваемости раком почек в США: на рис. 32 на карте США черным выделены графства с наименьшим уровнем заболеваний раком почек (слева) и графства с наивысшим уровнем заболеваний раком почек (справа). Конечно, эти графства разные, но сосредоточены они в основном в одних и тех же штатах. Эти штаты отличаются тем, что в них преобладает сельское население, они относительно мало населены, причем население в основном придерживается республиканских взглядов (христианский фундаментализм, опора на свои силы и пр.). В принципе, этих характеристик достаточно, чтобы объяснить каждый из наблюденных паттернов. Мало случаев рака почек? Конечно: сельский образ жизни, чистые вода и воздух, свежая незагрязненная пища. Много случаев рака почек? Тоже понятно: бедность, жирная пища, алкоголь, низкий уровень медицины. Единственный вопрос: можно ли совместить эти несовместимые причины?

Оказывается, да, можно совместить. Дело вовсе не в сельском образе жизни. Дело в малонаселенности. В малонаселенных графствах больше шансов для крайностей. Возьмем, например, наугад три шара из урны с равным количеством белых и черных шаров. При выборе каждого шара — две возможности, либо черный, либо белый. Значит, для трех шаров — 8 возможностей, каждая с вероятностью 1/8. При этом событие, состоящее в том, что один шар черный, а два белые, соответствует трем возможностям (ч-б-б, б-ч-б, б-б-ч), а событие, состоящее в том, что все шары белые — одной возмож-

[1] *H. Wainer, H. L. Zwerling*. Evidence That Smaller Schools Do Not Improve Student Achievement. URL: http://cog.state.pa.us)

ности (б-б-б), так же, как и событие, состоящее в том, что все шары черные. Если выбираются 4 шара, то вероятности «чистых», или экстремальных, событий «только белые» или «только черные» (б-б-б-б или ч-ч-ч-ч), уменьшаются вдвое, до 1/16. При выборе 7 шаров, вероятности экстремальных событий «только белые» или «только черные» уменьшаются еще в 8 раз, до 1/128. Чем больше шаров, тем меньше относительная вероятность экстремальных событий. Если уподобить графства совокупностям выбранных шаров, высокий уровень рака — с черным, а низкий уровень рака — с белым цветом, то получим, что вероятность экстремальных событий в малонаселенных графствах существенно превышает их вероятность в плотно населенных графствах. Вот и объяснение паттернов на рис. 32. Урновая модель хорошо разработана и понятна, поэтому указанный механизм, объясняющий данные паттерны, не встречает особых возражений.

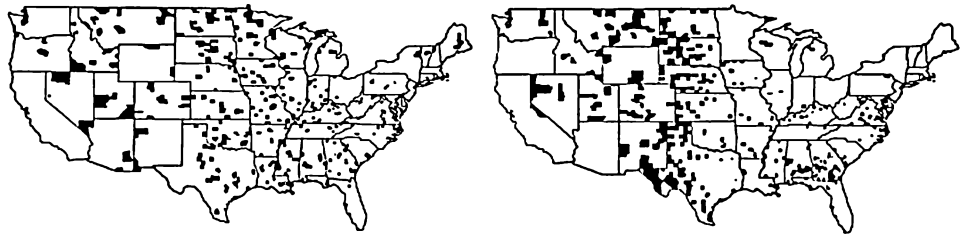


Рис. 3.2. **На карте США черным выделены графства с наименьшим уровнем заболеваний раком почек (слева) и графства с наивысшим уровнем заболеваний раком почек (справа)**

Факторы риска заболеваний органов дыхания

В 1981 г. в новосибирском Академгородке было проведено анкетирование около 50 000 человек — респондентами оказались практически все взрослые жители этого района, расположенного в двух десятках километрах от самого города. В Академгородке были очень часты случаи заболевания органов дыхания — тут были и бронхиты, и пневмонии, и гаймориты, и туберкулез – несколько десятков заболеваний в общей сложности.

Организатор обследования, В. Шанин, обратился к автору, руководившему тогда сектором математико-статистических методов анализа данных Института экономики и организации промышленного производства Сибирского отделения Академии Наук СССР, с просьбой помочь в обработке полученной анкетной информации. Он предполагал, что заболевания органов дыхания в Академгородке, как и везде, определяются двумя факторами риска: курением и пьянством. К счастью, в анкетах содержалась информация об индивидуальных привычках в потреблении и того, и другого, а также

масса других сведений, включая условия работы и быта респондентов.

Размер данных, десятки тысяч единиц, не позволял разместить их в памяти доступных нам в те годы компьютеров — данные располагались на отдельной магнитной ленте, доступ к которой занимал относительно большое время. Для проведения кластерного анализа этих данных мы разработали алгоритм, похожий на современные техники построения классификационных деревьев (см. материал П. С. Ростовцева [48] и монографию автора, в которой он размещен), и получили иерархическую группировку, в основном соответствующую делению на заболевания трех отдельных органов — легких, бронхов и носа (см. рис. 33). Она была получена последовательными делениями получаемых кластеров, максимизирующими суммарную меру сопряженности между конструируемым разбиением и 1/0 бинарными признаками, характеризующими наличие конкретных заболеваний. Затем мы стали смотреть таблицы сопряженности этого разбиения с различными характеристиками респондентов. К нашему удивлению, и, к досаде В. Шанина, мы не обнаружили особых связей между респираторными заболеваниями и курением или пьянством. Напротив, таблицы показывали практически полное отсутствие статистической связи между нашим разбиением и уровнями потребления алкоголя, равно как и курения, в том числе и с использованием предложенных нами тогда характеристик, оказавшихся индексами Кетле. Более того, мы обнаружили отрицательную связь между умеренным потреблением вина и респираторными заболеваниями в некоторых группах населения таких, как «женщины среднего возраста».

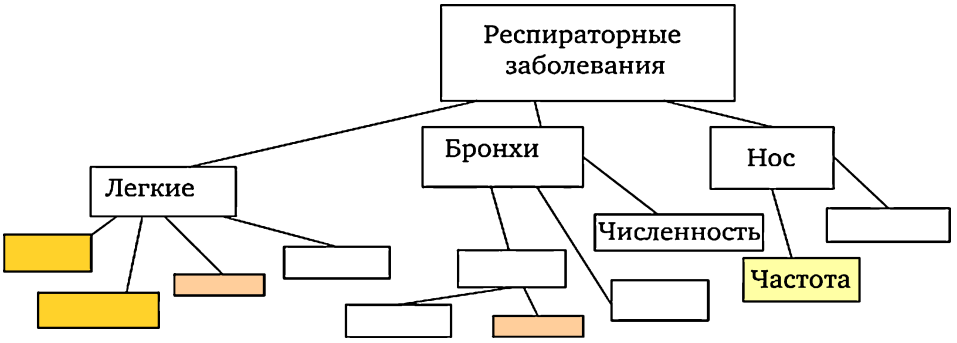


Рис. 3.3. **Иерархическая структура кластеров по признакам заболеваний органов дыхания, полученных нами в 1981 г. согласно методу, изложенному Ростовцевым [48]**

Серьезные связи нашего разбиения обнаружились для совсем других признаков. Оказалось, что респираторные заболевания в нашей выборке были связаны, прежде всего, с (а) наличием подобного заболевания в семье и (б) качеством жилищных условий. Получен-

ные результаты, вполне приемлемые и адекватные сегодня, абсолютно противоречили представлениям того времени и были отброшены как полная чепуха. С позиций сегодняшнего дня видится и то, что полностью ускользало тогда от нашего внимания. Объявленные факторы риска — алкоголь и курение — определялись индивидуальным поведением людей, как бы их нежеланием следовать здоровому образу жизни, тогда как обнаруженные нами связи указывали на социальные причины заболеваний — недостатки в жилищном и медицинском обеспечении со стороны государства. Этот аспект полностью противоречил объявленной цели социалистического государства СССР («все для блага человека») и уже только поэтому не имел права на существование.

Что показывают рассмотренные примеры? В общей ситуации, для того чтобы замеченный в данных паттерн мог являться основой принятия решений, необходимо предложить вразумительный механизм формирования паттерна. Желательно, чтобы этот механизм не противоречил имеющимся представлениям о явлении. Если же противоречит, то могут пройти годы и десятилетия, пока ситуация не прояснится. Ну и, конечно, само формулирование объясняющего механизма — это творческий акт, который может оказаться просто невозможным для неспециалиста по прикладной проблеме. Именно поэтому желательно привлечение к анализу данных специалистов в области, к которой относятся данные.

Анализ данных и смежные подходы

Анализ данных и искусственный интеллект

Под искусственным интеллектом понимают несколько разных вещей. Первоначально имелось в виду, что машина должна будет себя вести как человек, так что никто не сможет отличить, с кем ведет диалог — с человеком или машиной. Эту идею, возможно, в шутку, предложил один из основателей информатики Алан Тьюринг сразу после Второй мировой войны, еще до проникновения компьютеров в университеты. Идея яркая, но не совсем удачная по той причине, что обычно разработка новых концепций идет от простого к сложному, а, согласно Тьюрингу, вроде бы надо наоборот. То есть вместо того, чтобы сначала разобраться в отдельных блоках структуры интеллекта, таких как структурирование информации, поиск аналогий, обобщение, осмысление и прочее, проблема подменяется ее внешней оболочкой.

В 1960-е гг., когда компьютеры появились в университетах, тематику искусственного интеллекта подхватили математические логики. Искусственный интеллект, сказали они, — это мощнейшая машина логического вывода. Надо задать правильные аксиомы о том

или ином куске реальности, и машина автоматически выведет все его основные свойства путем использования и комбинирования различных логических умозаключений. Хотя за несколько десятилетий до этого австрийский математик Курт Гедель доказал, что само по себе это не выход: при достаточно богатой аксиоматике машина не может сама определить, что аксиоматика непротиворечива. А раз так, то все выводы становятся крайне ненадежны, ведь из противоречивых утверждений типа «А — это не А» можно логически вывести все что угодно. Древние софисты хорошо это понимали. И сами логики, и тем более их критики (если таковые были) должны были как-то уметь отговориться, апеллируя к критерию практики; науку, мол, нельзя отрывать от жизни. Так и происходило: логический подход развивался, были созданы концепции фрейма, неклассические логики, специальный машинный язык ПРОЛОГ и так далее. Это продолжалось примерно до начала нового столетия. Подобным же образом развивалась и математизация естественных языков, прежде всего в связи с проблемой перевода с языка на язык. Особенно далеко продвинулась математизация структуры предложений естественного языка.

К началу XXI столетия стало совершенно ясно, что все широковещательные обещания по развитию искусственного интеллекта, на которые не скупились его разработчики при получении многочисленных грантов и наград, не будут выполнены ни сейчас, ни в обозримом будущем — из их разработок ничего не выйдет. Выражение *artificial intelligence* в 1990-е гг. стало непрестижным, почти ругательным. Появились слова и дисциплины, такие как *computational intelligence* и *machine intelligence*, связанные с новыми подходами к искусственному интеллекту, опирающимися прежде всего на данные.

Например, понятие *computational intelligence* (приблизительный перевод — «вычислительный интеллект») объединяет такие относительно несвязанные направления, как искусственные нейронные сети, алгоритмизация решения оптимизационных задач через «эволюцию» сообществ взаимозаменяемых «частиц» и генетическое программирование. По-видимому, в настоящее время искусственный интеллект — это не более чем общее название разрозненных усилий по разработке тех или иных вычислительных моделей, реализующих те или иные аспекты интеллекта.

Все современные более или менее успешные реализации искусственного интеллекта (кроме роботов и генеративных уже-обученных трансформеров — текстоводов (GPT), конечно), т. е. «Яндекс. Пробки», Google Translate, «умный дом», «интернет вещей» — все это является результатом простых, но эффективных алгоритмов анализа данных, а не алгоритмов логического вывода. Просто наличие широкой сети взаимосвязанных компьютеров и возможность

невероятно быстрого — слово «мгновенный» здесь неуместно; время, затрачиваемое на глазной миг в тысячи раз медленнее — просмотра электронных файлов.

То же относится к российскому поисковику Яндекс. Возьмем какой-нибудь его сервис, например, Яндекс.Пробки, охватывающий в настоящее время около сотни городов в России и соседних государствах. Этот сервис «замеряет» скорость движения транспорта на участках основных транспортных улиц и магистралей города. Основной сигнал, учитываемый Яндексом — местоположение приборов-навигаторов, используемых водителями для ориентировки при движении к намеченной цели.

Весь город разбит на участки, так что, сопоставляя моменты времени при входе и выходе из участка, сервис определяет скорость автомобиля, на котором установлен навигатор. Уровень загруженности магистрали на данном участке определяется усреднением скоростей отдельных навигаторов. Случайные помехи типа данных от навигаторов, используемых пешеходами, легко отсекаются с помощью других признаков; например, того, что пешеходы включают навигатор лишь на короткое время. Или, еще проще, при усреднении можно отсекать по 1 % крайних, очень низких и очень высоких, скоростей, не вдаваясь в подробности и причины. Сопоставляя карту скоростей в данный момент с подобными картами в тот же день недели, то же время суток, при той же погоде, нетрудно прогнозировать распределение скоростей в ближайшее время и даже давать рекомендации об оптимальных маршрутах. Опять — ничего особо интеллектуального. Огромные скорости вычислений, огромная память, беспроводная связь электронных устройств, сенсоров, процессоров и серверов.

Обратимся к одному из наиболее обещающих начинаний последнего времени — Интернету вещей. Считается, что в ближайшие годы количество связанных беспроводной связью устройств достигнет десятков миллиардов. Они легко смогут осуществлять простейшие операции. Вы попали в пробку — автомобиль может послать сообщение на смартфон вашего друга: «Я опаздываю, наиболее вероятное опоздание — 40 минут». Электронный будильник может вас разбудить и послать сигнал кофейной машине, чтобы та начала кипятить воду и заваривать кофе. Принтер может послать заказ на установку нового сердечника, поскольку старый на исходе. Датчик, вшитый в пиджак, сообщит и вам, и вашему компьютеру, что вы были наиболее активны в такой-то промежуток времени.

Оставляю воображению читателя концепцию умного дома для пожилых, продвигаемую в последнее время в передовых странах. Перспективы беспредельны. При этом дедуктивные способности, которыми так гордятся Шерлоки Холмсы и Я-роботы (по А. Азимову),

могут спать спокойно. Они практически не понадобятся. Вы наврали в своем резюме, что окончили Гарвард? Нет необходимости собирать косвенные факты и делать сложные умозаключения. Можно просто поискать вашу фамилию в электронных архивах Гарварда.

Чуть сложнее работа таких современных чудодейственных систем искусственного интеллекта, как нейронные сети GPT-3 и GPT-4 — недавние разработки компании OpenAI, вполне удовлетворяющие тесту Тьюринга на интеллектуальность (см. [41]). Первая из них имеет 96 нейронных слоев, вторая, говорят, аж 120. В первой имеется порядка 175 миллиардов соединений между нейронами и, значит, столько же числовых параметров для оценки, вторая — говорят, в 10 раз больше. Обе системы в первом приближении нацелены на оценку условной вероятности следующего слова в последовательностях слов на основе обучения по огромным массивам текстов, вне всякой связи с грамматическими правилами. Скажем, для начальной последовательности «Сегодня я пойду на», наибольшими вероятностями будут оценены такие существительные как «улицу» и «работу», а наименьшими такие как «луна» или «коромысло» — просто потому, что в триллионах текстов, используемых для обучения, первые заканчивают такую фразу значительно чаще, чем вторые.

Современный искусственный интеллект работает на неинтеллектуальных принципах. Он обязан своими успехами:

а) огромному количеству практически мгновенных беспроводных связей;

б) огромной скорости машинных процессов;

в) практически безграничной памяти;

г) огромному количеству всевозможных данных в электронных хранилищах;

д) относительно простым методам их анализа.

Анализ данных и машинное обучение

Термины «анализ данных» и «машинное обучение» часто рассматривают как взаимозаменяемые, т. е., синонимичные. В какой-то степени это можно объяснить тем, что зачастую эти дисциплины используют одни и те же методы многомерного анализа. Однако способы использования этих методов могут отличаться, и довольно сильно.

Автор придерживается той точки зрения, что дисциплина анализа данных отличается тем, что обращается к данным как средству обогащения знаний о предметной области, прежде всего, в теоретическом аспекте, т. е. с точки зрения уточнения используемых понятий и отношений между ними. Машинное обучение же концентрируется на таких методах обработки данных, которые могут

адаптироваться к данным, т. е. «учиться» в процессе решения задачи. Такое понимание близко к определениям, даваемым международными источниками[1].

Различия в определениях носят не совсем схоластический характер. Можно указать по крайней мере три отличительных характеристики, вытекающие из этих определений:

1) Разное понимание «истины». Как уже отмечалось в подпараграфе 3.2.3, в анализе данных «истинность» относится, прежде всего, к наблюдениям, тогда как в машинном обучении «истинность» — это атрибут, прежде всего, модельных решений.

2) Разные критерии оценки качества модели. В машинном обучении качество модели определяется ее способностью адекватно работать на новых материалах, не использованных в процессе ее «тренировки». Поэтому в типичном случае данные разделяются на «тренировочные», используемые для «обучения» модели, и «тестирующие», используемые для проверки ее работоспособности. Напротив, в анализе данных качество модели определяется уровнем аппроксимации данных и способностью приводить к новым знаниям — выделение тестирующей части данных здесь бесполезно.

3) Отношение к интерпретируемости результатов. В анализе данных интерпретация — основа основ. Какое отношение к знаниям имеют результаты, которые не удается проинтерпретировать, т. е. понять? Напротив, в машинном обучении компьютер не нуждается в интерпретации. Естественным образом выделяются области, где одно полезно, а второе — нет. Например, не получится убедить судью или врача принять выводы компьютера, если к ним не прилагается разумное объяснение. Напротив, при серьезной аварии или пожаре, при необходимости предотвращения взрыва, и т. п. — придется довериться «интуиции» компьютерного алгоритма, если человеческая интуиция не срабатывает. Надо отметить, правда, что в самое последнее время в машинном обучении наметилось движение, ориентированное на разработку интерпретируемых (говорят, «объясняемых (*explainable*)») решений.

Практика различает термины «анализ данных» и «машинное обучение» самым существенным образом. Ниже приводятся количественные характеристики откликов на соответствующие запросы — русскоязычные (табл. 3.1) и англоязычные (табл. 3.2).

Несмотря на то, что практика действительно часто смешивает понятия «анализ данных» и «машинное обучение», так что среди объявлений в одной категории попадаются вакансии из другой, общая картина, очевидная из данных таблиц, такова. «Анализ данных» существенно опережает «машинное обучение» как по числу страниц, так и по числу вакансий.

---

[1] См. например, сайт https://azure.microsoft.com/ (дата посещения 06.03.2023).

Таблица 3.1

Реакция релевантных веб-сайтов на русскоязычные запросы (по результатам посещения 2 сентября 2023 г.)

Запрос	Google, страниц	HH.ru, вакансий	Superjob.ru, вакансий
Машинное обучение	2 660 000	2 474	859
Анализ данных	70 100 000	13 605	1 190

Таблица 3.2

Реакция релевантных веб-сайтов на англоязычные запросы (по результатам посещения 2 сентября 2023 г.)

Запрос	Google, pages	LinkedIn, jobs	Jooble.org, jobs
Machine learning	1 970 000 000	58 842	6 313
Data analysis	3 120 000 000	674 932	297 472

Анализ данных и математическая статистика

Имеются два основных толкования термина «анализ данных».

Согласно первому, наиболее популярному толкованию, анализ данных — это как бы математическая статистика «в широком смысле слова». При этом к анализу данных относятся все вычислительные методы обработки данных, в отличие от классической математической статистики, которая покрывает только методы, основанные на вероятностных моделях данных. Классическая математическая статистика следует так называемому научному подходу, согласно которому любые данные рассматриваются только в связи с некоторой заранее принятой моделью того, как устроено наблюдаемое явление или процесс. В типичном случае, согласно этому подходу, предполагается, что объективно существует некоторый механизм, порождающий вероятностное распределение на множестве всех исходов, а рассматриваемые данные получены в результате случайного выбора из этого распределения. Задача исследования данных в таком случае — пролить свет на распределение, из которого они получены, и на этой основе решать задачи исследования — объяснение, прогнозирование или принятие решений. С такой точкой зрения связана и терминология математической статистики. Наблюдаемые признаки в этом подходе — вовсе не признаки, а «переменные» — реализации так называемых «случайных величин», анализируемых в рамках математической теории вероятностей.

Как все понимают, при первоначальном изучении процесса или явления модельный подход не всегда возможен, так как недостаточно знаний о структуре, механизме и пр. изучаемого. Вот тут-то математическая статистика и использует эвристические, инженерные

методы для предварительного анализа имеющихся данных, с тем, чтобы хоть как-то разобраться в природе явления. При этом никто не ощущает необходимости в том, чтобы разобраться в структуре системы инженерных методов: зачем, ведь это не более чем подпорки для формирования правильной модели; их делают из любого подходящего материала. Подобным образом геометр, развивающий теорию «правильных» четырехугольников — квадратов, ромбов и пр. — отбрасывает все другие, неправильные, фигуры, как не вписывающиеся в теорию.

Рис. 3.3. **Данные, как и теории, могут оказаться не фальсифицируемыми (т. е. верными для всех случаев)**

Анализ данных «в узком смысле» пытается вычленить из моря всевозможных «достатистических» или «нестатистических» методов какую-то часть, поддающуюся систематизации. При этом возможны самые разнообразные критерии систематизации. Например, в дисциплине «майнинг данных» (*data mining*) значительная часть построений концентрируется вокруг различных уточнений понятия «интересности» наблюдений или закономерностей. Напротив, некоторые технически ориентированные исследователи кладут во главу угла используемые методы решения возникающих трудных задач оптимизации, например, «методы, инспирированные природой: генетические алгоритмы, метод роя частиц, метод муравьиной колонии и пр.», или же формы представления данных и результатов (машинные рассуждения, нечеткая логика и пр.). В монографии [12] автор

предложил базироваться на двух основных структурных элементах знания: понятиях и утверждениях о связях между понятиями. Имеется в виду, что цель анализа данных — уточнение или обогащение существующего знания об исследуемом явлении или процессе. При этом два самых прямых способа такого обогащения — это (а) порождение новых понятий из понятий, представляемых признаками, входящими в таблицу данных, и (б) порождение новых закономерностей, т. е. утверждений о связи признаков, подтвержденных данными. Эти две задачи: (а) агрегация или суммаризация (термин, заимствованный из английского, см. *summarization* [32], и широко применяемый в России в некоторых приложениях инженерной информатики) и (б) коррелирование (*correlation*, [32]). Эти соображения в значительной мере определяют содержание данного пособия.

Математический аппарат математической статистики ориентирован на решение задач, связанных с тем, какого рода информацию о модели порождения данных и ее параметрах можно получить из наблюдений. Возникает вопрос, какие темы может затрагивать теория анализа данных в отсутствие модели порождения данных. Прежде чем на него ответить, заметим, что говорить об отсутствии модели порождения данных в анализе данных не всегда уместно. Для некоторых задач модель порождения данных имеется — это, например, модель факторизации матриц, присущая как методу главных компонент, так и кластер-анализу по методу *K*-средних. Конечно, эта модель менее определенная, чем, скажем, Гауссово распределение вероятностей. Однако не менее генеративная, в том смысле, что предполагает, что данные порождаются в соответствии с моделью. В таких случаях задача именно такова, как она сформулирована в соответствующих разделах данного текста: определить характеристики модели с наименьшими ошибками. Другой класс задач теории анализа данных — формулирование таких моделей и методов, которые остаются адекватными при различных форматах представления данных — в виде таблиц объект — признак, или таблиц сходства между объектами, или таблиц сопряженности между признаками. В состав теоретического каркаса анализа данных должны войти изучение свойств конкретных постановок задач и методов их решения, установление теоретических связей между различными подходами, формирование новых классов задач и пр.

Имеется довольно много групп методов анализа данных, связанных общностью цели или организованным сообществом исследователей, имеющих свои издания, проводящих свои семинары и конференции, подчас сильно пересекающиеся. Рассмотрим наиболее популярные.

Классификация [*Classification*] — построение классификации, структурирующее рассматриваемое множество явлений в совокуп-

ность отдельных классов, отражающих важные свойства этих явлений. В настоящее время этот термин также применяется к задачам отнесения отдельных объектов к заранее заданным классам.

Кластер-анализ [*Cluster analysis*] — совокупность методов, разделяющих объекты таблицы данных в множества — кластеры — таким образом, чтобы сходные объекты попадали в один и тот же кластер, а несходные — в разные кластеры.

Вычислительный интеллект [*Computational intelligence*] — дисциплина, использующая нечеткие (*fuzzy*) множества; алгоритмы, инспирированные природой (*nature-inspired algorithms*); нейронные сети (*neural nets*) и другие подобные средства, чтобы имитировать человеческий интеллект в его способности адаптироваться к природе данных в процессе вычислений.

Майнинг данных [*Data mining*] — совокупность методов для отыскания интересных закономерностей по данным, организованным в виде компьютерной базы или хранилища данных. Эти «интересные закономерности» образуют как бы вновь обнаруженное знание. Поэтому майнинг данных обычно рассматривается как часть общего процесса накопления или порождения знаний (*knowledge discovery*).

Извлечение документов, извлечение информации [*Document retrieval, information retrieval*], часто также переводится несколько неточно как поиск документов или информации — совокупность критериев и методов для поиска и извлечения документов из баз и хранилищ данных по запросу. Эта область особенно популярна в связи с развитием поисковых систем интернета, таких как Яндекс или Google.

Факторный анализ [*Factor analysis*] — совокупность методов для измерения ненаблюдаемых, скрытых характеристик, таких как уровень интеллекта ученика или уровень социально-экономического развития территории, по косвенным, измеримым характеристика объектов.

Генетические алгоритмы [*Genetic algorithms*] — подход к глобальному поиску решений сложных задач оптимизации путем имитации процесса наследования генов в популяции. Для этого организуется процесс эволюции некоторого множества возможных решений, каждое из которых представлено в виде линейной «хромосомы». При переходе от поколения к поколению используются вероятностные механизмы генерации «брачных пар», «кроссовера», «мутаций», «сохранения элиты».

Обнаружение знаний [*Knowledge discovery*] — совокупность методов для отыскания количественных формул и концептуальных утверждений, связывающих различные аспекты данных между собой.

Математическая статистика [*Mathematical statistics*] — подход, предполагающий, что данные порождены в соответствии с некоторой вероятностной моделью и являются средством оценки тех или иных параметров модели или проверки статистических гипотез о них. С одной стороны, такая модель может быть наиболее точной формой знания о рассматриваемом явлении. С другой стороны, какую-то модель можно предположить даже и при слабом уровне знаний, а потом целенаправленно проводить эксперименты и собирать данные, чтобы подтвердить или улучшить модель.

Нейронные сети [*Neural networks*] — подход к моделированию связи между входными и выходными признаками, используя структуру взаимосвязанных искусственных нейронов (устройств, испускающих выходной сигнал при накоплении достаточного количества входных сигналов); параметры сети обычно подбираются в процессе машинного обучения.

Инспирированные природой алгоритмы [*Nature-inspired algorithms*] — современные методы оптимизации сложных функций, основанные не на изучении свойств задачи, как в классической математике, а с помощью процессов последовательного изменения популяции решений таким образом, чтобы имитировать какой-либо биологический или социальный процесс (движение роя пчел или колонии муравьев, репетиция оркестра и пр.).

Оптимизация [*Optimization*] — область вычислительной математики, в которой разрабатываются методы анализа и решения проблем отыскания минимума или максимума так называемой «целевой функции». В анализе данных это обычно: (а) минимизация суммы квадратов невязок между данными наблюдений и данными модели, порождаемыми «решающим правилом», получаемым как результат анализа данных (метод наименьших квадратов), или (б) максимизация так называемой функции максимального правдоподобия.

Распознавание образов [*Pattern recognition*] — несколько устаревшее название для дисциплины, занимающейся построением классификационных решающих правил (распознавание с учителем, *supervised learning*) или кластеров (распознавание без учителя, *unsupervised learning*) по данным наблюдений.

Социальная статистика [*Social statistics*] — дисциплина, связанная с методами измерения социальных и экономических индексов по выборочным данным или данным государственной статистики.

Анализ текстов [*Text analysis*] — совокупность подходов и методов для автоматизации анализа текстовых документов, как например, установления степени сходства текстов, категоризации документов, формирования аннотаций и пр.

Список литературы

Книги

1. *Айвазян, С. А.* Прикладная статистика: основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — Москва : Финансы и статистика, 1983.
2. *Айвазян, С. А.* Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. — Москва : Финансы и статистика, 1989.
3. *Аркадьев, А. Г.* Обучение машины классификации объектов / А. Г. Аркадьев, Э. М. Браверман. — Москва : Наука, 1971.
4. *Бурков, А.* Машинное обучение без лишних слов / А. Бурков. — Санкт-Петербург : Издательский дом «Питер», 2020.
5. *Грас, Дж.* Data Science: Наука о данных с нуля ; перевод с английского / Дж. Грас. — Санкт-Петербург : БХВ-Петербург, 2021.
6. *Загоруйко, Н. Г. Когнитивный анализ данных* / Н. Г. Загоруйко. — ИМ : Новосибирск, 2013.
7. *Загоруйко, Н. Г.* Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. — ИМ : Новосибирск, 1999.
8. *Калинина, В. Н.* Компьютерный практикум по прикладной статистике и эконометрике / В. Н. Калинина, В. И. Соловьев. — Москва : Вега-Инфо, 2010.
9. *Крыштановский, А. О.* Анализ социологических данных с помощью пакета SPSS / А. О. Крыштановский. — Москва : ИД Высшей школы экономики, 2008.
10. *Кулаичев, А. П.* Методы и средства комплексного анализа данных / А. П. Кулаичев. — 4-е изд. — Москва : ФОРУМ ; ИНФРА-М, 2006.
11. *Лагутин, М. Б.* Наглядная математическая статистика / М. Б. Лагутин. — Москва : Бином, 2009.
12. *Миркин, Б. Г. Анализ качественных признаков и структур* / Б. Г. Миркин. — Москва : Финансы и статистика, 1980.
13. *Миркин, Б. Г.* Группировки в социально-экономических исследованиях / Б. Г. Миркин. — Москва : Финансы и статистика, 1985.
14. *Миркин, Б. Г.* Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2017.
15. *Мхитарян, В. С.* Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2024.

16. *Митчелл, М.* Идиот или гений: как работает и на что способен искусственный интеллект / М. Митчелл. — Москва : АСТ Корпус, 2022.

17. *Мюллер, Д. П.* Питон и наука о данных для «чайников» ; перевод с английского / Д. П. Мюллер, Л. Массарон. — Москва ; Санкт-Петербург : Диалектика, 2020.

18. *Петрунин, Ю. Ю.* Информационные технологии анализа данных (Data Analysis) / Ю. Ю. Петрунин. — 3-е изд. — Москва : КДУ, 2021.

19. *Плошко, Б. Г.* История статистики / Б. Г. Плошко, И. И. Елисеева. — Москва : Финансы и статистика, 1990.

20. *Толстова, Ю. Н.* Математическая статистика для социологов : учебник и практикум для вузов / Ю. Н. Толстова. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2024.

21. *Тюрин, Ю. Н.* Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. — Москва : ИД ФОРУМ, 2010.

22. *Berthold, M.* Intelligent Data Analysis / M. Berthold, D. Hand. — Springer-Verlag, 2007.

23. *Chambers, J. M.* Graphical methods for data analysis / J. M. Chambers (el al.). — Chapman and Hall/CRC, 2018

24. *Davison, A. C.* Bootstrap Methods and Their Application / A. C. Davison, D. V. Hinkley. — Cambridge University Press, 2005.

25. *Denis, D. J.* SPSS data analysis for univariate, bivariate, and multivariate statistics / D. J. Denis. — John Wiley & Sons, 2018.

26. *Duda, R. O.* Pattern Classification / R. O. Duda, P. E. Hart, D. G. Stork. — Wiley-Interscience, 2001.

27. *Grus, J.* Data science from scratch: first principles with python / J. Grus. — O'Reilly Media, 2019.

28. *Han, J.* Data Mining: Concepts and Techniques / J. Han, M. Kamber. — 2nd ed. — Morgan Kaufmann Publishers, 2006.

29. *Kendall, M. G.* Advanced Statistics: Inference and Relationship / M. G. Kendall, A. Stewart. — 3rd ed. — Griffin: London, 1973.

30. *Lohninger, H.* Teach Me Data Analysis / H. Lohninger. — Berlin ; New York ; Tokyo : Springer-Verlag, 1999.

31. *Mirkin, B.* Mathematical Classification and Clustering / B. Mirkin. — Kluwer Academic Press, 1996.

32. *Mirkin, B.* Core Data Analysis: Summarization, Correlation, Visualization / B. Mirkin. — 2nd ed.— Springer-London, 2019.

33. *Mirkin, B.* Clustering: A Data Recovery Approach / B. Mirkin. — 2nd ed. — Chapman & Hall/CRC, 2012.

34. *Mitchell, T. M.* Machine Learning / T. M. Mitchell. — McGraw Hill, 2010.

35. *Silverman, B. W.* Density Estimation for Statistics and Data Analysis / B. W. Silverman. — Routledge, 2018.

36. *Soukup, T.* Visual Data Mining / T. Soukup, I. Davidson. — Wiley and Son, 2002.

37. *Webb, A.* Statistical Pattern Recognition / A. Webb. — Wiley and Son, 2002.

Периодические издания

38. *Carpenter, J.* Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians / J. Carpenter, J. Bithell // Statistics in Medicine. 2000. —№ 19. — P. 1141—1164.

39. *Dale, A. I.* Bayes or Laplace? An examination of the origin and early applications of Bayes' theorem / A. I. Dale // Archive for History of Exact Sciences. — P. 23—47.

40. *Fawcett, T.* An introduction to ROC analysis / T. Fawcett // Pattern Recognition Letters. — 2006. — № 27. — P. 861—874.

41. *Hacker, P.* Regulating ChatGPT and other large generative AI models / P. Hacker, A. Engel, M. Mauer // Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. — 2023. — P. 1112—1123.

42. *Jahoda, G.* Quetelet and the emergence of the behavioral sciences / G. Jahoda // SpringerPlus. — 2015. — 4(1). — P. 1—10.

43. *Mirkin, B.* Eleven ways to look at the chi-squared coefficient for contingency tables / B. Mirkin // The American Statistician. — 2001. — 55. — № 2.— P. 111—120.

44. *Mirkin, B.* A straightforward approach to chi-squared analysis of associations in contingency tables / B. Mirkin. // In E. Beh (Ed.): Analysis of Categorical Data from Historical Perspectives — Essays in Honour of Shizuhiko Nishisato. — Singapore : Springer Nature, 2023. — P. 59—72.

45. *Миркин, Б. Г.* Анализ данных и искусственный интеллект / Б. Г. Миркин // Постнаука [сайт]. — URL: https://postnauka.org/talks/80147 (дата обращения 06.03.2024).

46. *Мухия, С.* Решение задачи Титаник на Kaggle для начинающих (A beginner's guide to Kaggle's Titanic problem) / С. Мухия ; перевод с английского Н. Арзамасова // Neurohive [сайт]. — 2019. — URL: https://neurohive.io/ru/osnovy-data-science/razbor-resheniya-zadachi-titanik-na-kaggle-dlya-nachinajushhih/ (дата обращения 06.03.2024).

47. *Pearson, K.* On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling / K. Pearson // Philosophical Magazine. — 1900. — 50. — P. 157—175.

48. *Ростовцев, П. С.* Методы иерархического группирования / П. С. Ростовцев // *Миркин, Б. Г.* Группировки в социально-экономических исследованиях. — Москва : Финансы и статистика, 1985. — С. 126—133.

Приложение 1

Основы вычислительной среды MATLAB и ее аналогов

MATLAB — это программная среда, которая позволяет непрофессиональному математику или программисту относительно легко кодировать численные алгоритмы и обрабатывать данные относительно небольшой размерности (до или около миллиона чисел) на ноутбуке или домашнем ПК. Имеется несколько бесплатных аналогов этой программы, прежде всего, Октав (GNU Octave, https://octave.org/) и СкайЛаб (SciLab, https://www.scilab.org/). Бóльшая часть ниже излагаемого материала относится и к этим пакетам. Кроме того, записи программ на MATLAB могут рассматриваться как реализации псевдокода для соответствующих алгоритмов.

Данный материал описывает основы работы на MATLAB для непрофессионального пользователя. На образовательной платформе «Юрайт» в приложении 2 к данному курсу вы можете найти коды программ, упоминавшихся в отдельных главах:

— cm.m — отыскание центра Минковского с помощью эволюционного метода.;

— plan.m — совокупность модулей для построения регрессии по степенному закону с помощью двух разных подходов: (а) применение эволюционного метода минимизации расстояния Минковского; (б) линеаризация. Включает модуль для запоминания результатов в ASCII-файл (может быть модифицирован на другие задачи), а также множества чисел, порожденных компьютером с помощью генератора случайных чисел: (а) три выборки по 50 чисел из трех распределений — файл short.dat (см. Проекты 2.2, 2.3) и (б) 280 чисел из $N(0, 10)$.

Как начать работу

Пользователь сам должен решить, как организовать память в компьютере. Можно рекомендовать для всех вычислений на MATLAB организовать отдельную папку, скажем, Code, с двумя подпапками, Data и Result, в которых хранятся данные и результаты вычислений, соответственно.

Затем надо щелкнуть по пиктограмме MATLAB, и на экране возникнет многооконное приложение. Главным является командное окно [Command Window]. Можно указать MATLAB рабочую

папку обычным способом, выбрав папку в соответствующем окне, или же исполнив в командном окне команду типа

```
cd <Path_To_Working_Directory>
```

MATLAB этот путь запомнит, и в следующий раз туда можно перейти с помощью маленького окна, расположенного вверху.

MATLAB организован как совокупность пакетов, посвященных различной математической проблематике, каждый в своей собственной папке, где находятся файлы отдельных программ, каждый с расширением .m. Помощь можно запросить обычным образом, кликнув на символ ?, а можно непосредственно в рабочей памяти, написав команду help. Если ввести это слово, будут показаны все имеющиеся пакеты, среди них, например, matlab\datafun. В ответ на команду help datafun на экране появятся все команды, имеющиеся в пакете, включая, например, max -largest component. Команда help max вызовет на экран объяснение операции max.

Загрузка и запоминание файлов

Числовые данные следует организовывать как таблицу «объект — признак»: строки соответствуют объектам (наблюдениям), причем элементы строки разделены либо запятыми, либо пробелами; столбцы — признакам (так организованы файлы studn.dat и studn.var). Такую структуру данных, все компоненты которой — количественные, называют двумерным массивом [2D array], что соответствует математическому понятию матрицы. Одномерные массивы соответствуют отдельным объектам или признакам. Массив — основной формат количественных данных в MATLAB. Он работает по принципу шахматной доски. Например, arr(i,k) обозначает элемент массива с именем arr, который находится на пересечении *i*-й строки и *k*-го столбца. Файлы Excel имеют подобную структуру, плюс к тому в Экселе разрешается размещать буквенные выражения. В MATLAB же предпочтительнее иметь дело с файлами, содержащими однородную информацию: либо числа, либо буквы, но не то и другое. Определяющая особенность числовых массивов состоит в том, что каждая строка должна иметь одно и то же количество чисел.

Чтобы загрузить такой файл из памяти в рабочую область MATLAB, можно использовать команды из пакета iofun. Простейшая — операция load для загрузки числовых файлов, организованных как описано выше, в рабочую область:

```
>> arr = load('Data\iris.dat'); % файл iris.dat из папки Data
                                % загружается в переменную arr
% Символ " %" используется для обозначения комментариев,
% читаемых людьми, но пропускаемых компьютером.
% точка с запятой ";" отделяет одну инструкцию
% компьютеру от другой;
```

```
% если точка с запятой пропущена в конце строки,
% то результат распечатывается на экране, что позволяет
% легко организовать проверку вычислений;
% iris.dat это файл 150x4 таблицы данных о 150 цветках ириса,
% охарактеризованных 4 признаками;
% Все названия признаков сохранены
% в отдельном файле iris.var директории Data.
```

Одномерный массив можно создать, например, командой

```
>> a = [3 4 7 0];
```

которая помещает в a 1×4 массив, который можно перевести в 4×1 массив с помощью операции транспозиции

```
>> b = a'
```

Поскольку здесь нет точки с запятой в конце, результат появится на экране:

```
3
4
7
0
```

Чтобы извлечь его вторую компоненту, 4, используется команда

```
>> c = b(2)
```

Аналогично, команда

```
>> d = arr(7,8)
```

помещает значение из 7-й строки и 8-го столбца массива arr в переменную d рабочей области.

Для сохранения числового массива в рабочей области можно использовать команду save, которая допускает несколько разных форматов запоминания, включая внутренний формат .mat MATLAB (см. выдачу команды help save). Чтобы сохранить массив X в файл Result\good.res в формате ASCII (это стандартный формат, покрывающий символику стандартной компьютерной клавиатуры), можно использовать команду

```
>> save Result\good.res X -ascii
```

Другой популярный формат данных в MATLAB — это «строки» [*strings*], выделяемые одиночными кавычками «'». Этот формат используется для представления и хранения имен и названий. Поскольку названия могут имет разную длину, их не удается сохранять в массивах. Поэтому используется другой формат данных, поддерживаемый в MATLAB, «ячейка» (*cell*). Этот формат представляется фигурными скобками (для массивов используются круглые скобки): arr(i) — *i*-й элемент массива arr, тогда как brr{i} — *i*-й элемент ячейки brr, причем этот элемент может быть и числом, и строкой,

и матрицей, и даже другой ячейкой. MATLAB поддерживает и другие структуры данных, включая изображения и звук, но это выходит за рамки данного текста.

Таблица А1

Таблица с данными о четырех признаках пяти студентов в MS Excel

Студент	Возраст	Число детей	Профессия	Оценка
Джон	35	0	ИТ	94
Пегги	28	2	ДА	67
Фред	27	1	ДА	85
Крис	28	0	ПР	48
Лиз	25	0	ИТ	87

Примечание: в столбце Профессия: ИТ — Информационные технологии; ДА — Деловая администрация; ПР — Прочее, в столбце Оценка — Оценка экзаменационной работы, в шкале 0—100 %.

MATLAB работает с внешними данными в популярных форматах без перевода их в форму числовой таблицы. Например, для ввода и вывода файлов MS Excel (с расширением .xls) в MATLAB есть команды xlsread и xlswrite. Хотя это и кажется просто, но пользователю не следует ожидать очень уж удобного сопряжения MS Excel с MATLAB. Например, файл Excel, воспроизведенный в табл. А1, с помощью команды xlsread будет переписан в три структуры данных — одна для числовой части, вторая — для текста, а третья — для всего файла. Точнее, команда

```
>> [nn,tt,rr] = xlsread('Data\student.xls');
% nn массив числовых значений, tt – ячейка текста,
% rr – ячейка, покрывающая все данные из файла
```

произведет числовой массив nn:

$$nn = \begin{matrix} 35 & 0 & NaN & 94 \\ 28 & 2 & NaN & 67 \\ 27 & 1 & NaN & 85, \\ 28 & 0 & NaN & 48 \\ 25 & 0 & NaN & 87 \end{matrix}$$

а также ячейку текста tt размера 8×5:

```
tt =
    'Признак'    'Возраст'    'Число детей'    'Профессия'    'Оценка'
    ''           ''           ''               ''             ''
    'Джон'       ''           ''               'ИТ'           ''
    'Джон'       ''           ''               'ДА'           ''
    'Фред'       ''           ''               'ДА'           ''
    'Крис'       ''           ''               'ПР'           ''
    'Лиз'        ''           ''               'ИТ'           ''
```

А также ячейку 8 × 5, rr:

	'Признак'	'Возраст'	'Число детей'	'Профессия'	'Оценка'
	[NaN]	[NaN]	[NaN]	[NaN]	[NaN]
	'Джон'	[35]	[0]	'ИТ'	[94]
rr =	'Джон'	[28]	[2]	'ДА'	[67]
	'Фред'	[27]	[1]	'ДА'	[85]
	'Крис'	[28]	[0]	'ПР'	[48]
	'Лиз'	[25]	[0]	'ИТ'	[87]

Символ NaN используется в MATLAB для обозначения неопределенного числа, возникающего при делении чисел на нуль.

Как видно, эти таблицы не такие уж и удобные для обработки из-за многочисленных вхождений символа NaN и кавычек.

Работа с подмножествами объектов и признаков

Имея массив arr размера 100 × 8, можно легко создать массив, включающий, скажем, только три признака из 8, например, «Возраст», «Дети» и «Оценка по программированию». Для этого надо создать массив индексов этих переменных, скажем, 4, 5 и 7, соответственно:

```
>> ii = [4 5 7]
% двоеточие в конце отсутствует, чтобы массив ii
% появился на экране;
```

Если имена 8 признаков хранятся в «ячеечном» файле b, то команды для формирования файла всех объектов с редуцированным множеством признаков могут иметь вид:

```
>> newa = arr(:,ii); % newa – новый массив числовых данных
>> newb = b(ii);
% newb – новое множество названий признаков: оно определяется
% через круглые, не фигурные скобки, несмотря на то, что и b,
% и newb имеют структуру ячейки, а не массива
```

Аналогично определяются подмножества объектов. Если, например, нужен файл данных только о тех студентах, которые получили 60 % или выше на экзамене по Программированию, следует сначала найти множество индексов этих студентов с помощью команды find:

```
>> jj = find(arr(:,7)> = 60);
% jj – множество объектов, определенных предикатом операции
find()
% arr(:,7) – это 7-й столбец массива arr;
% он соответствует признаку «оценка по программированию».
```

Теперь можно применить arr к jj:

```
>> al = arr(jj,:);
% данные о всех признаках на объектах, попавших в jj
```

Размеры массива данных можно определить с помощью команды

```
>>size(al)
```

Если в ответ появится

```
55 8,
```

то, значит, массив al состоит из 55 строк и 8 столбцов. При необходимости иметь переменные для этих значений команда слегка меняется:

```
>>[n,m] = size(al)
```

Это помещает 55 в n, а 8 в m.

Теперь можно показать, как преобразовать в MATLAB «грязные» данные, полученные выше с помощью операции xlsread из файла Excel. В частности, получен файл nn с бессмысленным третьим столбцом. Чтобы его удалить, мы сначала определяем число столбцов

```
>> [rnn,cnn] = size(nn);
% т. е. число столбцов – cnn
```

Потом удаляем столбец 3 из множества всех столбцов с помощью операции теоретико-множественного вычитания множеств:

```
>> vv = setdiff([1:cnn],3);
% операция setdiff(x,y) удаляет из x все те элементы
% массива y, которые входят в x
% [1:cnn] – это массив всех натуральных чисел от 1 до cnn
% включительно, например, [1:4] = [1 2 3 4]
```

Операция

```
>> nnr = nn(:,vv);
```

сохраняет весь nn, за исключением столбца 3, в массив nnr:

$$nnr = \begin{matrix} 35 & 0 & 94 \\ 28 & 2 & 67 \\ 27 & 1 & 85 \\ 28 & 0 & 48 \\ 25 & 0 & 87 \end{matrix}$$

Теперь можно обратиться к проблеме формирования файла названий признаков. Сначала надо получить файл структуры ячейка для всех признаков. Они составляют конец первой строки ячейки tt, получаемый удалением первого названия «Признак», как видно из распечатки tt, приведенной выше. Таким образом, команда

```
>> fe = tt(1,2:5);
```

порождает ячейку fe размера 1×4, состоящей из названий 4 признаков.

Теперь можно удалить признак 3 — применением fe к полученному выше массиву vv:

```
>>fer = fe(vv);
```

Ячейка fer включает строки ‘Возраст’, ‘Число детей’, ‘Оценка’, индексированные числами 1, 2 и 3, и соответствующие столбцам массива nnr.

Многие другие операции MATLAB вводятся в основном тексте, особенно в проектах, рабочих примерах и заданиях.

Приложение 2

Две программы на Матлабе

Эволюционный процесс для отыскания центра Минковского

```
% cm.m, computing Minkowski p-distance central point c
% of a series x along with the average distance d
% and its proportion pe in the sum

function [c,d,pe]=cm(x,p)

n=length(x);
lb=min(x);
rb=max(x); %-----------------lb, rb are boundaries of the area
(i)---
de=0;
for ik=1:n
    de=de+(abs(x(ik)))^p;
end
de=de/n;%---------------------------average p-th power of the
data

%--population setting (ii) and setting the limit, iter,
to iterations
pp=15; %population size
feas=(rb-lb)*rand(pp,1)+lb; % generated population
of p c values
                            % within the range
flag=1;
count=0;
iter=5000;
%---------evaluation of the initially generated population
(iii)----
funp=0;
for ii=1:pp
    vv(ii)=mink(p,x,feas(ii));
end
[funi, ini]=min(vv);
soli=feas(ini) %initial best c value
funi %initial error
si=1;%0.5; %step of change
%-------------evolution of the population (iv) --------------
--
while flag==1
    count=count+1;
```

```
    feas=feas+si*randn(pp,1);
    % Gaussian mutation added with step si
    for ii=1:pp
        feas(ii)=max(lb, feas(ii));
        feas(ii,:)=min(rb,feas(ii));% keeping the population
                                    % within the range
        vec(ii)=mink(p,x,feas(ii)); % evaluation
    end
%-------------- elite maintenance (v) ---------------
    [fun, in]=min(vec); %best distance value
    sol=feas(in,:);%corresponding c value
    [wf,wi]=max(vec);
    wun=feas(wi); %worst c
    if wf>funi
        feas(wi)=soli;
        vec(wi)=funi; % changing the worst for the elite
    end
    if fun < funi
        soli=sol;
        funi=fun;
    end
    if (count>=iter)
        flag=0;
    end
   pe=funi/de;
%------------ screen the results of every 1000th iteration
    if rem(count,1000)==0
        %funp=funi;
        disp([soli pe]);
    end
end
c=soli;
d=funi;
pe=d/de;

return

%---computing the quality of ce, the average deviation
in p-th power
function dis=mink(p,x,ce)

nn=length(x);
dis=0;
for ik=1:nn
    dis=dis+(abs(x(ik)-ce))^p;
end
dis=dis/nn;

return
```

При заданных показателе степени p и массиве x программа cm(p,x) старается найти как можно меньшее значение общему отклонению точек x от центра, вычисляемому подпрограммой mink.

Оценка степенного закона: эволюционный метод и линеаризация

```
% plan.m, анализ степенного закона в предположении, ч
% то предиктор x и цель y уже находятся в рабочей памяти Мат-
Лаба
% степенной закон:                    y=ax^b                (1)
% линеаризованная версия:             log(y)=log(a)+b*log(x)(2)


function plan(x,y)

% --linear analysis of log(x) and log(y)/
% линейный анализ логарифмов

for ii=1:length(x);xc(ii)=max(.05,x(ii));
yc(ii)=max(0.05,y(ii));end;
% 0.05 instead of 0 to make logarithms possible
xll=log(xc);
yll=log(yc);
[all,bll,cll, rvll]=lr(xll,yll);

% all - the slope, bll - the intercept of the linear regression
% cll - the correlation coefficient, rvll - the residual variance
% of the linear regression
yle = all*xll+bll; % yll, оцененная по линейной регрессии
cd=cll^2; % determinacy coefficient, it should be cd=1-rvll
cd
rvll
% figure(1);plot(xll,yll,'k.',xll,yle,'rp');

% --- линеаризация: оценка уравнения (1)  через оценку уравне-
ния (2)

[al,bl, rl]=llr(x,y);
% al the estimate of a, bl the estimate of b and
% rl the proportion of the residual variance in the variance
of y

% ylr - the linearized rule estimate for the power law
for ii=1:length(x);ylr(ii)=al*x(ii)^bl;end;

%---as is: fitting equation (1) by straightforwardly minimizing
%---the residual variance with an evolutionary algorithm /
%---непосредственная оценка (1) (эволюционно)

[an,bn,f, rn]=nlr(x,y);
% an the estimate of a, bn the estimate of b and
% rn the proportion of the residual variance in the variance
of y
for ii=1:length(x);yn(ii)=an*x(ii)^bn;end; %estimated power law

% output: two-plot figure, real on the left, log on the right /
```

```
% двухоконная фигура, слева в реальных шкалах,
% справа – в логарифмах
% figure(2);
subplot(1,2,1);
plot(x,y,'k.',x,ylr,'b.',x,yn,'r.');%data scatter with two
estimated power laws,
% две оценки: синий – линеаризованная, красный – как есть
subplot(1,2,2);plot(xll,yll,'k.',xll,yle,'rp');

% output: text file of the results / выдача в виде
% текстового файла (см. ниже)
saveplan('rep', cll, al, bl, rl, an, bn, rn,cd);

return

% llr.m,  fitting a nonlinear regression function y=ax^b
% using linearization/ оценка через линеаризованную версию
% x is predictor, y is target,
% a,b -regression parameters to be fitted

function [a,b, residvar]=llr(xt,yt);

% regression is power law y=a*x^b as reflected
% in the procedure
% residvar is the average square error's proportion
% to the variance of y;
% xt, yt are predictor and target

%-----an elementary check of length compatibility--------
ll=length(xt);
if ll~=length(yt)
    disp('Something wrong is with the data');
    pause;
end

%--------- calculating a and b using the linearization
for ii=1:ll;xc(ii)=max(.05,xt(ii));yc(ii)=max(0.05,yt(ii));end;
% putting 0.05 instead of zero
% to make possible logarithms of the data
xl=log(xc); % taking log of x and y
yl=log(yc);

[al,bl,dl]=lr(xl,yl);
b=al;
a=exp(bl);
ab=[a b];
residvar=delta(ab,xt,yt)/var(yt,1);
return

%--computing the quality of the approximation y=a*(x^b)
% which is the residual variance
```

```
function esq=delta(tt,x,y)%tt=[a, b]; x predictor, y target
a=tt(1);
b=tt(2);
esq=0;
for ii=1:length(x)
    yp(ii)=a*(x(ii)^b); % можно взять и другую функцию
    esq=esq+(y(ii)-yp(ii))^2;
end
esq=esq/length(x);
return;

% nlr.m, evolutionary fitting of a nonlinear regression
% function y=f(x,a,b) / эволюционный метод
% x is predictor, y is target,
% a,b -regression prameters to be fitted

function [a,b, funi,residvar]=nlr(xt,yt);

% in this version the regression equation is power law
y=a*x^b which
% is reflected only in the subroutine 'delta' in the bottom
% for computing the value of the average error squared;
% вид минимизируемой функции «зашит» в 'delta' внизу
% funi is the average square error's best value;
% residvar is its proportion to the variance of y;
% xt, yt are predictor and target

%-----an elementary check of length compatibility--------
ll=length(xt);
if ll~=length(yt)
    disp('Something is wrong with the data');
    pause;
end
% --- determine rectangle at which (a,b)-populations fluctuate
% оценка области изменения a и b
[ab,bb]=ddr(xt,yt);

lb=[ab(1) bb(1)];
rb=[ab(2) bb(2)];
lb
rb
% интервалы изменения оцениваемых величин
disp('Hit ENTER if you wish to proceed. ');
pause;
% -organisation of iterations, iter the limit to their number
% число итераций равно iter
p=15; % population size / размер популяции
for ii=1:p;feas(ii,:)=(rb-lb).*rand(1,2)+lb;end;
% generated population of p pairs coefficients within the range
flag=1;
count=0;
iter=10000;%5000;
% evaluation of the initially generated population
```

```
% оценка начальной популяции
funp=0;
for ii=1:p
    vv(ii)=delta(feas(ii,:),xt,yt);
end
[funi, ini]=min(vv);
soli=feas(ini,:) %initial coeffts
funi %initial error
si=1;%0.5; %step of change
% ---------evolution of the population/ эволюция популяции----
while flag==1
    count=count+1;
    feas=feas+si*randn(p,2); %mutation added with step si/
мутация
    for ii=1:p
        feas(ii,:)=max([lb;feas(ii,:)]);
        feas(ii,:)=min([rb;feas(ii,:)]);% keeping the
population
                                          % within the range
        vec(ii)=delta(feas(ii,:),xt,yt);% evaluation / оценка
    end

    [fun, in]=min(vec); % best approximation value
                        % наилучший член популяции
    sol=feas(in,:);% corresponding  parameters
    [wf,wi]=max(vec);
    wun=feas(wi,:); % worst case / наихудший член популяции
    if wf>funi
        feas(wi,:)=soli;
        vec(wi)=funi;
% changing the worst for the best of the previous generation
    end
    if fun < funi
        soli=sol;
        funi=fun;
    end
    if (count>=iter)
        flag=0;
    end
 residvar=funi/var(yt,1);
% ---screen the results of every 500th iteration ---
% ---вывод результатов на экран каждые 500 итераций---
    if rem(count,500)==0
        %funp=funi;
        disp([soli residvar]);
    end
end
a=soli(1);
b=soli(2);
return

% --computing the quality of the approximation y==a*(x^b) --
% оценка качества оценки
```

```
function esq=delta(tt,x,y)%tt=[a, b]; x predictor, y target
a=tt(1);
b=tt(2);
esq=0;
for ii=1:length(x)
    yp(ii)=a*(x(ii)^b); % this is a power law function
    esq=esq+(y(ii)-yp(ii))^2;
end
esq=esq/length(x); % the average difference squared
return;


% ddr.m, determination of the domain for power law y=a*x^b with
b
% определение допустимой области
% restricted
function [ab,bb]=ddr(x,y)
n=length(x);
bm=(log(y(1))-log(y(2)))/(log(x(1))-log(x(2)));
am=y(1)/(x(1)^bm);
ab=[am am];
bb=[bm bm];
%------finding extreme values for a and b using pairwise
equations
bs=0;as=0; bsq=0;asq=0;
count=0;
for ii=1:(n-1);
    if min(x(ii),y(ii))>.25
    for jj=(ii+1):n
        if min(x(jj),y(jj))>.25
        if (x(ii)/x(jj)<0.75)|(x(ii)/x(jj)>1.25)
            count=count+1;
            bt=(log(y(ii))-log(y(jj)))/(log(x(ii))-log(x(jj)));
            aij=y(ii)/(x(ii)^bt);
            aij=min(aij,100);%restriction
            %if (aij>100)
             %   disp([ii jj]); aij
            %end;
            bs=bs+bt;
            bsq=bsq+bt*bt;
            as=as+aij;
            asq=asq+aij*aij;
            if bt>bb(2)
                bb(2)=bt;
            end;
            if bt<bb(1)
                bb(1)=bt;
            end;
            if aij>ab(2)
                ab(2)=aij;
            end;
            if aij<ab(1)
                ab(1)=aij;
```

```
            end;
        end;
        end;
    end;
    end;
end;
as=as/count
asq=asq/count;
sas=sqrt(asq-as^2)
bs=bs/count
bsq=bsq/count;
sbs=sqrt(bsq-bs^2)
ab(1)=as-4*sas;ab(2)=as+4*sas;
bb(1)=bs-4*sbs;bb(2)=bs+4*sbs;
count
return

% saveplan.m, saving  results of the power-law analysis
in plan.m
% сохранение отчета о результатах

function saveplan(file, cc, al, bl, rl, an, bn, rn,cd);

ct =num2str(cc);
first=['Results of the power-law analysis y=ax^b' ];
alla=['On the level of logarithms, the correlation is  '
num2str(cc)];
alex=['Explained proportion of log(y)-variance is '
num2str(100*cd) '%'];
nt=[ ];
lt1=['Linearized estimate parameter values are a= ' num2str(al)
', b= ' num2str(bl)];
lt2=['Explained proportion of y-variance is r= '
num2str(100*(1-rl)) '%' ];
nt1=['"As is" estimate parameter values are a= ' num2str(an) ',
b= ' num2str(bn)];
nt2=['Explained proportion of y-variance is r= '
num2str(100*(1-rn)) '%'];
alltext=strvcat(alla, lt1,lt2,nt1,nt2);

oul=[' These are visualized on the Figure produced:']
our=[' The power-law estimates on the left, the logarithms,
on the right'];
    alltt=strvcat(alltext, oul, our);
    alltt
Filename=[ file '.out'];
fid= fopen(Filename, 'at');
if fid~=-1
    fprintf(fid, '%s\n', first);
    fprintf(fid, '%s\n', '     ');
    fprintf(fid, '%s\n', alla);
    fprintf(fid, '%s\n', alex);
    fprintf(fid, '%s\n', '     ');
```

```
    fprintf(fid, '%s\n', lt1);
    fprintf(fid, '%s\n', lt2);
     fprintf(fid, '%s\n', '     ');
     fprintf(fid, '%s\n', nt1);
    fprintf(fid, '%s\n', nt2);
     fprintf(fid, '%s\n', '     ');
     fprintf(fid, '%s\n', oul);
    fprintf(fid, '%s\n', our);
    fprintf(fid, '%s\n', '     ');
    fprintf(fid, '%s\n', '     ');
    fclose(fid);
end;
return
```

Приложение 3

Две случайные выборки для экспериментов

Таблица 1

Три случайные выборки из трех разных распределений

8		20	1512
12	21	50	
11	23	48	
10	21	206	
9		9	12
7		20	199
10	22	51	
12	18	50	
9		20	198
13	21	843	
9		5	12
13	13	8	
10	10	7	
11	14	9	
9		18	39
9		13	12
7		21	51
11	20	46	
11	21	50	
9		18	54

8		20	1391
10	19	49	
10	19	41	
13	24	35	
12	23	45	
10	13	11	
12	9	9	
10	21	49	
7		10	10
8		17	52
12	8	8	
11	20	48	
12	17	199	
8		11	9
8		11	13
9		20	978
12	17	51	
9		20	6233
13	19	23	
10	21	47	
11	11	8	
11	20	973	
11	7	43	
13	20	201	
9		18	200
10	19	49	
9		10	7
14	20	36	
9		10	8
11	21	203	

Таблица 2

Выборка 280 значений из $N(0, 10)$, отсортированных в порядке возрастания

−30.29	−12.48	−7.01	−2.99	1.76	5.58	10.35
−25.57	−12.29	−6.94	−2.91	1.98	5.59	10.50
−25.34	−12.27	−6.83	−2.83	1.98	5.63	10.94
−23.79	−11.89	−6.79	−2.78	2.07	5.65	10.98
−23.34	−11.61	−6.65	−2.75	2.08	5.65	11.08
−22.38	−11.50	−6.64	−2.66	2.14	5.74	11.13
−22.37	−11.33	−6.11	−2.66	2.14	5.74	11.64
−21.78	−11.10	−6.02	−2.58	2.18	5.81	12.28
−21.05	−10.78	−5.98	−2.52	2.21	5.82	12.33
−20.89	−10.57	−5.87	−2.23	2.27	5.89	12.59
−20.65	−10.52	−5.53	−2.07	2.28	6.13	12.79
−19.10	−10.44	−5.35	−2.06	2.29	6.26	12.93
−18.16	−10.13	−5.33	−1.91	2.36	6.29	13.15
−17.95	−10.09	−5.22	−1.90	2.37	6.51	13.24
−17.79	−10.08	−5.17	−1.74	2.56	6.55	13.42
−17.58	−10.06	−4.91	−1.60	2.71	6.59	13.44
−16.47	−9.79	−4.82	−1.51	2.79	6.59	13.48
−16.43	−9.11	−4.62	−1.44	2.85	6.65	13.56
−16.31	−9.08	−4.58	−1.42	2.91	7.00	13.99
−16.19	−9.01	−4.53	−1.28	2.94	7.09	14.27
−16.15	−8.95	−4.43	−1.26	2.98	7.16	14.69
−16.14	−8.93	−4.26	−0.80	3.16	7.30	14.95
−15.90	−8.71	−4.18	−0.79	3.21	7.58	15.35
−15.89	−8.53	−4.17	−0.73	3.27	7.99	15.74
−15.67	−8.49	−4.08	−0.50	3.27	8.34	15.82
−15.56	−8.01	−4.01	−0.49	3.46	8.57	15.84
−15.50	−7.98	−3.98	−0.23	3.66	8.58	15.99
−15.04	−7.97	−3.95	−0.21	3.74	8.70	16.03
−15.00	−7.75	−3.84	−0.08	3.80	8.85	16.84
−14.91	−7.67	−3.78	−0.02	4.29	8.87	16.87
−14.16	−7.48	−3.74	0.03	4.39	8.97	17.29
−14.14	−7.46	−3.65	0.33	4.41	9.02	17.62

Окончание табл. 2

–14.04	–7.44	–3.61	0.65	4.42	9.08	18.43
–13.88	–7.37	–3.59	0.70	4.48	9.12	19.57
–13.84	–7.37	–3.47	0.78	4.60	9.39	19.58
–13.72	–7.35	–3.46	0.80	4.78	9.57	20.80
–13.58	–7.27	–3.39	1.10	4.94	9.83	22.38
–13.33	–7.24	–3.14	1.20	5.28	10.02	22.66
–12.98	–7.20	–3.02	1.38	5.41	10.08	29.50
–12.68	–7.03	–3.01	1.58	5.54	10.09	32.03

Наши книги можно приобрести:

Учебным заведениям и библиотекам:
в отделе по работе с вузами
тел.: (495) 744-00-12, e-mail: vuz@urait.ru

Частным лицам:
список магазинов смотрите на сайте urait.ru
в разделе «Частным лицам»

Магазинам и корпоративным клиентам:
в отделе продаж
тел.: (495) 744-00-12, e-mail: sales@urait.ru

Отзывы об издании присылайте в редакцию
e-mail: gred@urait.ru

Новые издания и дополнительные материалы доступны
на образовательной платформе «Юрайт» urait.ru,
а также в мобильном приложении «Юрайт.Библиотека»

Учебное издание

Миркин Борис Григорьевич

БАЗОВЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебник и практикум для вузов

Формат $70\times100\,^{1}/_{16}$.
Гарнитура «Charter». Печать цифровая.
Усл. печ. л. 23,04.

ООО «Издательство Юрайт»
111123, г. Москва, ул. Плеханова, д. 4а.
Тел.: (495) 744-00-12. E-mail: izdat@urait.ru, www.urait.ru