

Практическая статистика для специалистов Data Science

50 ВАЖНЕЙШИХ ПОНЯТИЙ



Питер Брюс и Эндрю Брюс

Practical Statistics for Data Scientists

50 Essential Concepts

Peter Bruce and Andrew Bruce

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Питер Брюс, Эндрю Брюс

Практическая статистика для специалистов Data Science

50 ВАЖНЕЙШИХ ПОНЯТИЙ

Санкт-Петербург
«БХВ-Петербург»

2018

УДК 004.6+519.2
ББК 32.81+22.172
Б89

Брюс, П.

Б89 Практическая статистика для специалистов Data Science: Пер. с англ. /
П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.: ил.
ISBN 978-5-9775-3974-6

Книга рассчитана на специалистов в области Data Science, обладающих некоторым опытом работы с языком программирования R и имеющих предварительное понятие о математической статистике. В ней в удобной и легкодоступной форме представлены ключевые понятия из статистики, которые относятся к науке о данных, а также объяснено, какие понятия важны и полезны с точки зрения науки о данных, какие менее важны и почему. Подробно раскрыты темы: разведочный анализ данных, распределения данных и выборок, статистические эксперименты и проверка значимости, регрессия и предсказание, классификация, статистическое машинное обучение и обучение без учителя.

Для аналитиков данных

УДК 004.6+519.2
ББК 32.81+22.172

Группа подготовки издания:

Руководитель проекта	<i>Евгений Рыбаков</i>
Зав. редакцией	<i>Екатерина Капалыгина</i>
Компьютерная верстка	<i>Ольги Сергиенко</i>
Оформление обложки	<i>Марины Дамбиевой</i>

© 2018 BHV

Authorized translation of the English edition of *Practical Statistics for Data Scientists* ISBN 9781491952962

© 2017 Andrew Bruce and Peter Bruce.

Авторизованный перевод английской редакции книги *Practical Statistics for Data Scientists* ISBN 9781491952962

© 2017 Andrew Bruce and Peter Bruce.

"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

ISBN 978-1-491-95296-2 (англ.)
ISBN 978-5-9775-3974-6 (рус.)

© 2017 Andrew Bruce and Peter Bruce
© Перевод на русский язык, оформление. ООО "БХВ-Петербург",
ООО "БХВ", 2018

Оглавление

Об авторах.....	13
Предисловие	15
Чего ожидать.....	15
Условные обозначения, принятые в книге	15
Использование примеров кода	16
Благодарности.....	16
Комментарий переводчика	17
Глава 1. Разведочный анализ данных	19
Элементы структурированных данных	20
Дополнительные материалы для чтения.....	22
Прямоугольные данные	23
Кадры данных и индексы.....	24
Непрямоугольные структуры данных.....	25
Дополнительные материалы для чтения.....	26
Оценки центрального положения	26
Среднее	27
Медиана и робастные оценки.....	28
Выбросы	29
Пример: оценки центрального положения численности населения и уровня убийств.....	30
Дополнительные материалы для чтения.....	31
Оценки вариабельности	31
Стандартное отклонение и связанные с ним оценки.....	33
Оценки на основе процентилей	35
Пример: оценки вариабельности населения штатов.....	36
Дополнительные материалы для чтения.....	37
Обследование распределения данных	37
Процентили и коробчатые диаграммы	38
Частотная таблица и гистограммы	39
Оценки плотности.....	41
Дополнительные материалы для чтения.....	43
Обследование двоичных и категориальных данных	43
Мода.....	45
Математическое ожидание	45
Дополнительные материалы для чтения.....	46
Корреляция.....	46
Диаграммы рассеяния	49
Дополнительные материалы для чтения.....	50

Исследование двух или более переменных	51
Шестиугольная сетка и контуры (отображение числовых данных против числовых)	51
Две категориальных переменных	54
Категориальные и числовые данные	55
Визуализация многочисленных переменных	56
Дополнительные материалы для чтения	58
Резюме	58
Глава 2. Распределения данных и выборки	59
Случайный отбор и смещенная выборка	60
Смещение	62
Произвольный выбор	63
Размер против качества: когда размер имеет значение?	64
Выборочное среднее против популяционного среднего	65
Дополнительные материалы для чтения	66
Систематическая ошибка отбора	66
Регрессия к среднему	67
Дополнительные материалы для чтения	69
Выборочное распределение статистики	69
Центральная предельная теорема	72
Стандартная ошибка	72
Дополнительные материалы для чтения	73
Бутстрап	74
Повторный отбор против бутстрапирования	77
Дополнительные материалы для чтения	77
Доверительные интервалы	77
Дополнительные материалы для чтения	80
Нормальное распределение	80
Стандартное нормальное распределение и квантиль-квантильные графики	82
Длиннохвостые распределения	84
Дополнительные материалы для чтения	85
t -Распределение Стьюдента	86
Дополнительные материалы для чтения	88
Биномиальное распределение	88
Дополнительные материалы для чтения	90
Распределение Пуассона и другие с ним связанные распределения	90
Распределения Пуассона	91
Экспоненциальное распределение	92
Оценка интенсивности отказов	92
Распределение Вейбулла	93
Дополнительные материалы для чтения	94
Резюме	94
Глава 3. Статистические эксперименты и проверка значимости	95
A/B -тестирование	95
Зачем нужна контрольная группа?	98
Почему только A/B ? Почему не C , D ?	99
Дополнительные материалы для чтения	100
Проверка статистических гипотез	100
Нулевая гипотеза	102

Альтернативная гипотеза.....	102
Односторонняя и двухсторонняя проверки гипотез.....	103
Дополнительные материалы для чтения.....	104
Повторный отбор.....	104
Перестановочный тест.....	105
Пример: прилипчивость веб-страниц.....	105
Исчерпывающий и бутстраповский перестановочные тесты.....	108
Перестановочные тесты: сухой остаток для науки о данных.....	109
Дополнительные материалы для чтения.....	109
Статистическая значимость и p -значения.....	110
p -Значение.....	112
Альфа.....	112
Чему равно p -значение?.....	113
Ошибки 1-го и 2-го рода.....	114
Наука о данных и p -значения.....	114
Дополнительные материалы для чтения.....	115
Проверка на основе t -статистики.....	115
Дополнительные материалы для чтения.....	117
Множественное тестирование.....	117
Дополнительные материалы для чтения.....	121
Степени свободы.....	121
Дополнительные материалы для чтения.....	122
ANOVA.....	123
F -статистика.....	126
Двухсторонняя процедура ANOVA.....	127
Дополнительные материалы для чтения.....	127
Проверка на основе статистики хи-квадрат.....	128
Проверка χ^2 : подход на основе повторного отбора.....	128
Проверка χ^2 : статистическая теория.....	130
Точная проверка Фишера.....	131
Актуальность проверок для науки о данных.....	133
Дополнительные материалы для чтения.....	134
Алгоритм многорукого бандита.....	134
Дополнительные материалы для чтения.....	137
Мощность и размер выборки.....	138
Размер выборки.....	140
Дополнительные материалы для чтения.....	141
Резюме.....	142
Глава 4. Регрессия и предсказание.....	143
Простая линейная регрессия.....	143
Уравнение регрессии.....	144
Подогнанные значения и остатки.....	146
Наименьшие квадраты.....	148
Предсказание против объяснения (профилирование).....	149
Дополнительные материалы для чтения.....	150
Множественная линейная регрессия.....	150
Пример: данные о жилом фонде округа Кинг.....	151
Диагностика модели.....	152

Перекрестная проверка	154
Отбор модели и шаговая регрессия	155
Взвешенная регрессия	157
Предсказание на основе регрессии	158
Опасности экстраполяции	159
Доверительный и предсказательный интервалы	159
Факторные переменные в регрессии	161
Представление фиктивных переменных	162
Многоуровневые факторные переменные	164
Порядковые факторные переменные	165
Интерпретация уравнения регрессии	166
Коррелированные предикторы	167
Мультиколлинеарность	168
Искажающие переменные	169
Взаимодействия и главные эффекты	170
Проверка допущений: диагностика регрессии	172
Выбросы	173
Влиятельные значения	174
Гетероскедастичность, ненормальность и коррелированные ошибки	177
Графики частных остатков и нелинейность	179
Нелинейная регрессия	181
Параболическая регрессия	182
Сплайновая регрессия	183
Обобщенные аддитивные модели	185
Дополнительные материалы для чтения	187
Резюме	187
Глава 5. Классификация	189
Наивный байесовский алгоритм	190
Почему точная байесовская классификация непрактична?	191
Наивное решение	192
Числовые предикторные переменные	194
Дополнительные материалы для чтения	194
Дискриминантный анализ	195
Ковариационная матрица	196
Линейный дискриминант Фишера	196
Простой пример	197
Дополнительные материалы для чтения	199
Логистическая регрессия	199
Функция логистического отклика и логит-преобразование	200
Логистическая регрессия и обобщенная линейная модель	202
Обобщенные линейные модели	203
Предсказанные значения в логистической регрессии	203
Интерпретация коэффициентов и отношений шансов	204
Линейная и логистическая регрессии: сходства и различия	205
Подгонка модели	205
Диагностика модели	206
Дополнительные материалы для чтения	209
Оценивание моделей классификации	210
Матрица несоответствий	211

Проблема редкого класса	213
Прецизионность, полнота и специфичность	213
ROC-кривая	214
Метрический показатель AUC	216
Лифт	217
Дополнительные материалы для чтения.....	218
Стратегии в отношении несбалансированных данных	219
Понижающий отбор	220
Повышающий отбор и повышающая/понижающая перевесовка.....	220
Генерация данных.....	221
Стоимостно-ориентированная классификация	222
Обследование предсказаний	222
Дополнительные материалы для чтения.....	224
Резюме	224
Глава 6. Статистическое машинное обучение	225
<i>K</i> ближайших соседей	226
Небольшой пример: предсказание невозврата ссуды.....	227
Метрические показатели расстояния	229
Кодировщик с одним активным состоянием.....	230
Стандартизация (нормализация, <i>z</i> -оценки)	231
Выбор <i>K</i>	233
Метод KNN как конструктор признаков	234
Древовидные модели.....	235
Простой пример	237
Алгоритм рекурсивного сегментирования	238
Измерение однородности или разнородности	240
Остановка роста дерева	241
Предсказывание непрерывной величины	243
Каким образом деревья используются.....	243
Дополнительные материалы для чтения.....	244
Бэггинг и случайный лес.....	244
Бэггинг	246
Случайный лес	246
Важность переменных.....	249
Гиперпараметры	251
Бустинг	252
Алгоритм бустинга	253
XGBoost	254
Регуляризация: предотвращение перепогонки	256
Гиперпараметры и перекрестная проверка	259
Резюме	261
Глава 7. Обучение без учителя.....	263
Анализ главных компонент	264
Простой пример	265
Вычисление главных компонент.....	267
Интерпретация главных компонент.....	267
Дополнительные материалы для чтения.....	270

Кластеризация на основе K средних	270
Простой пример	271
Алгоритм K средних	272
Интерпретация кластеров	273
Выбор количества кластеров	275
Иерархическая кластеризация	277
Простой пример	277
Дендограмма	278
Агломеративный алгоритм	279
Меры различия	280
Модельно-ориентированная кластеризация	281
Многомерное нормальное распределение	282
Смеси нормальных распределений	283
Выбор количества кластеров	285
Дополнительные материалы для чтения	287
Шкалирование и категориальные переменные	287
Шкалирование переменных	288
Доминантные переменные	289
Категориальные данные и расстояние Говера	290
Проблемы кластеризации смешанных данных	293
Резюме	294
Библиография	295
Предметный указатель	297

*Мы хотим посвятить эту книгу памяти наших родителей,
Виктора Г. Брюса и Нэнси С. Брюс, которые воспитали в нас
страсть к математике и точным наукам,
а также нашим первым учителям, Джону У. Тьюки и Джулиану Саймону,
и нашему верному другу, Джеффу Уотсону, который вдохновил нас на то,
чтобы мы посвятили свою жизнь статистике*

Питер Брюс основал и расширил Институт статистического образования Statistics.com, который теперь предлагает порядка 100 курсов в области статистики, из которых примерно половина предназначена для аналитиков данных. Нанимая в качестве преподавателей ведущих авторов и шлифуя маркетинговую стратегию для привлечения внимания профессиональных аналитиков данных, Питер развил широкое представление о целевом рынке и свои собственные экспертные знания для его завоевания.

Эндрю Брюс имеет более чем 30-летний стаж работы в области статистики и науки о данных в академической сфере, правительстве и бизнесе. Он обладает степенью кандидата наук в области статистики Вашингтонского университета и опубликовал несколько работ в рецензируемых журналах. Он разработал статистико-ориентированные решения широкого спектра задач, с которыми сталкиваются разнообразные отрасли, начиная с солидных финансовых фирм до интернет-стартапов, и располагает глубоким пониманием практики науки о данных.

Предисловие

Книга рассчитана на аналитика данных, обладающего некоторым опытом работы с языком программирования R и имеющего предшествующий (возможно, обрывочный или сиюминутный) контакт с математической статистикой. Мы оба, авторы этой книги, пришли в мир науки о данных из мира статистики, и поэтому у нас есть определенное понимание того вклада, который статистика может привнести в науку о данных, как прикладную дисциплину. В то же время мы хорошо осведомлены об ограничениях традиционного статистического обучения: статистика как дисциплина насчитывает полтора столетия, и большинство учебников и курсов по статистике отягощены кинетикой и инерцией океанского лайнера.

В основе настоящей книги лежат две цели:

- ◆ представить в удобной, пригодной для навигации и легкодоступной форме ключевые понятия из статистики, которые относятся к науке о данных;
- ◆ объяснить, какие понятия важны и полезны с точки зрения науки о данных, какие менее важны и почему.

Чего ожидать

Ключевые термины

Наука о данных — это сплав многочисленных дисциплин, включая статистику, информатику, информационные технологии и конкретные предметные области. В результате при упоминании конкретной идеи могут использоваться несколько разных терминов. Ключевые термины и их синонимы в данной книге будут выделяться в специальной выноске, такой как эта.

Условные обозначения, принятые в книге

В книге используются следующие условные обозначения.

- ◆ *Курсив* указывает новые термины.
- ◆ **Полужирный шрифт** — URL-адреса, адреса электронной почты.
- ◆ Моноширинный шрифт используется для распечаток программ, а также внутри абзацев для ссылки на элементы программ, такие как переменные или имена функ-

ций, базы данных, типы данных, переменные окружения, операторы и ключевые слова.

- ◆ Моноширинный шрифт курсивом показывает текст, который должен быть заменен значениями пользователя либо значениями, определяемыми по контексту.



Данный элемент обозначает подсказку или совет.



Данный элемент обозначает общее замечание.



Данный элемент обозначает предупреждение или предостережение.

Использование примеров кода

Дополнительный материал (примеры кода, упражнения и пр.) доступен для скачивания по адресу <https://github.com/andrewgbuice/statistics-for-data-scientists>.

Эта книга предназначена для того, чтобы помочь вам решить ваши задачи. В целом, если код примеров предлагается вместе с книгой, то вы можете использовать его в своих программах и документации. Вам не нужно связываться с нами с просьбой о разрешении, если вы не воспроизводите значительную часть кода. Например, написание программы, которая использует несколько фрагментов кода из данной книги, официального разрешения не требует.

Адаптированный вариант примеров в виде электронного архива вы можете скачать по ссылке <ftp://ftp.bhv.ru/9785977539746.zip>, которая доступна также со страницы книги на сайте www.bhv.ru.

Благодарности

Авторы выражают признательность всем, кто помог воплотить эту книгу в реальность.

Герхард Пилхер (Gerhard Pilcher), генеральный директор консалтинговой фирмы Elder Research в области глубинного анализа данных, был свидетелем ранних черновиков книги и предоставил нам подробные и полезные поправки и комментарии. Так же как и Эн्या Макгверк (Anyu McGuirk) и Вей Сяо (Wei Xiao), специалисты

в области статистики в SAS, и Джэй Хилфигер (Jay Hilfiger), член авторского коллектива O'Reilly, которые предоставили полезные рекомендации на первоначальных стадиях работы над книгой.

В издательстве O'Reilly Шэннон Катт (Shannon Cutt) в дружеской атмосфере сопровождал нас в течение всего процесса публикации и в меру подстегивал нашу работу, в то время как Кристен Браун (Kristen Brown) гладко провела нашу книгу через производственную стадию. Рэйчел Монагэн (Rachel Monaghan) и Элайю Сасмэн (Eliahu Sussman) бережно и терпеливо проводили корректировку и улучшение нашего текста, тогда как Эллен Траутмэн-Зайг (Ellen Troutman-Zaig) подготовила предметный указатель. Мы также благодарим Мари Богуро (Marie Beaugureau), которая инициировала наш проект в O'Reilly, и Бена Бенгфорда (Ben Bengfort), автора O'Reilly и преподавателя в statistics.com, представившего нас издательству O'Reilly.

Мы извлекли большую пользу из многих бесед, которые Питер провел за последние годы с Галитом Шмуели (Galit Shmueli), соавтором других книжных проектов.

Наконец, мы хотели бы особо поблагодарить Элизабет Брюс (Elizabeth Bruce) и Дебору Доннелл (Deborah Donnell), чьи терпение и поддержка сделали это начинание возможным.

Комментарий переводчика

Прилагаемый к настоящей книге программный код протестирован в среде Windows 10 с использованием действующих версий программных библиотек (время перевода книги — октябрь-ноябрь 2017 г.). При тестировании исходного кода за основу взята среда R версии 3.4.2.

Адаптированный и скорректированный исходный код примеров лучше всего разместить в подпапке домашней папки пользователя. Например:

/home/r_projects/statistics-for-data-scientists-master

или

C:\Users\[ИМЯ_ПОЛЬЗОВАТЕЛЯ]\r_projects\statistics-for-data-scientists-master

Структура папки с прилагаемыми примерами такова:

data	Наборы данных, используемые в книге и в сценариях R
figures	Графики, полученные в результате выполнения сценариев R
src	Исходный код примеров в виде сценариев R

Разведочный анализ данных

Статистика как дисциплина получила свое развитие главным образом в прошлом столетии. Теория вероятностей — математический фундамент статистики — разрабатывалась с XVII по XIX в. на основе работ Томаса Байеса (Thomas Bayes), Пьера-Симона Лапласа (Pierre-Simon Laplace) и Карла Гаусса (Carl Gauss). В отличие от чисто теоретической природы вероятности, статистика является прикладной наукой, занимающейся анализом и моделированием данных. Современная статистика как строгая научная дисциплина восходит корнями к концу 1800-х гг. — Фрэнсису Гальтону (Francis Galton) и Карлу Пирсону (Karl Pearson). Р. А. Фишер (R. A. Fisher) в начале XX в. был ведущим новатором современной статистики, который ввел в употребление такие ключевые понятия, как *планирование эксперимента* и *оценка максимального правдоподобия*. Эти и многие другие статистические понятия в основном находятся в отдаленных уголках науки о данных. Главная цель настоящей книги состоит в том, чтобы помочь высветить эти понятия и разъяснить их важность — или ее отсутствие — в контексте науки о данных и больших данных.

В данной главе основное внимание уделяется первому шагу в любом проекте науки о данных: разведке данных. *Разведочный анализ данных* (exploratory data analysis, EDA) — это сравнительно новая область статистики. Классическая статистика фокусировалась почти исключительно на *статистическом выводе*, т. е. иногда сложном наборе процедур для получения выводов о популяциях (или генеральных совокупностях) на основе небольших выборок. В 1962 г. Джон У. Тьюки (John W. Tukey, рис. 1.1) в своей концептуальной статье "Будущее анализа данных" [Tukey-1962] призвал к преобразованию статистики. Он предложил новую научную дисциплину под названием *анализ данных*, которая включала статистический вывод в качестве всего лишь одного из компонентов. Тьюки наладил связи с инженерным и вычислительным сообществами (он ввел термины "*bit*" — от англ. *binary digit*, и "*программное обеспечение*"), а его исходные принципы оказались удивительно прочными и формируют часть фундамента науки о данных. Область разведочного анализа данных появилась благодаря книге Тьюри "*Анализ результатов наблюдений*" [Tukey-1977], теперь уже ставшей классической.

Благодаря доступности вычислительных мощностей и выразительному программному обеспечению для анализа данных разведочный анализ данных эволюционировал далеко за пределы своей исходной области. Ключевыми факторами этой дисциплины явились быстрая разработка новой технологии, доступ к более разнообразным и большим по объему данным и более широкое применение количественного анализа во множестве дисциплин. Дэвид Донохо (David Donoho), препода-

ватель статистики в Стэнфордском университете и прежний выпускник Тьюки, написал превосходную статью "50 лет науки о данных" на основе своей презентации на семинаре в честь 100-летия Тьюки, проходившем в Принстоне, шт. Нью-Джерси [Donoho-2015]. В ней Донохо прослеживает возникновение науки о данных к новаторскому вкладу Тьюки в анализ данных.



Рис. 1.1. Джон Тьюки, выдающийся статистик, чьи идеи, разработанные более чем 50 лет назад, формируют фундамент науки о данных

Элементы структурированных данных

Данные поступают из многих источников: показаний датчиков, событий, текста, изображений и видео. *Интернет вещей* (Internet of Things, IoT) извергает потоки информации. Значительная часть этих данных не структурирована: изображения представляют собой набор пикселей, при этом каждый пиксел содержит информацию о цвете в формате RGB (красный, зеленый, синий). Тексты состоят из последовательностей словарных и несловарных символов, часто разбитых на разделы, подразделы и т. д. Потоки нажатий клавиш представляют собой последовательности действий пользователя, взаимодействующего с приложением или веб-страницей. По сути дела, основная задача науки о данных состоит в том, чтобы переработать этот поток сырых данных в информацию, полезную в практической деятельности. Для применения рассмотренных в этой книге статистических понятий неструктурированные сырые данные должны быть обработаны и помещены в структурированную форму — подобно той, которая может появляться из реляционной базы данных — либо быть собранными для статистического исследования.

Ключевые термины

Непрерывные данные (continuous)

Данные, которые могут принимать любое значение в интервале.

Синонимы: интервал, число с плавающей точкой, числовое значение.

Дискретные данные (discrete)

Данные, которые могут принимать только целочисленные значения, такие как количественные значения.

Синонимы: целое число, количество.

Категориальные данные (categorical)

Данные, которые могут принимать только определенный набор значений, в частности набор возможных категорий.

Синонимы: перечисления, перечислимые данные, факторы, именованные данные, полихотомические данные.

Двоичные данные (binary)

Особый случай категориальных данных всего с двумя категориями значений (0/1, истина/ложь).

Синонимы: дихотомический, логический, флаг, индикатор, булево значение.

Порядковые данные (ordinal)

Категориальные данные с явно выраженной упорядоченностью.

Синонимы: порядковый фактор.

Существует два основных типа структурированных данных: числовой и категориальный. Числовые данные поступают в двух формах: *непрерывной*, как например скорость ветра или продолжительность времени, и *дискретной*, как например количество возникновений события. *Категориальные* данные принимают только фиксированный набор значений, например тип экрана телевизора (плазма, LCD, LED и т. д.) или название штата (Алабама, Аляска и т. д.). *Двоичные* данные представляют собой важный особый случай категориальных данных. Эти данные принимают только одно из двух значений, таких как 0/1, да/нет или истина/ложь. Еще один полезный тип категориальных данных представлен *порядковыми* данными, в которых категории упорядочены; их примером является числовая характеристика (1, 2, 3, 4 или 5).

Зачем заморачиваться таксономией типов данных? Оказывается, что в целях анализа данных и предсказательного моделирования тип данных играет важную роль для определения типа визуального отображения, анализа данных либо статистической модели. По сути дела, в программных системах для науки о данных, таких как R и Python, эти типы данных используются для улучшения вычислительной производительности. А более важно, что тип данных переменной определяет то, каким образом программная система будет обращаться с вычислениями для этой переменной.

У разработчиков программного обеспечения (ПО) и программистов баз данных (БД) может возникнуть вопрос: зачем в аналитике нужны понятия "*категориальные*" и "*порядковые*" данные? В конце концов, категории являются просто набором текстовых (либо числовых) значений, и основная БД автоматически работает с их внутренним представлением. Однако четкая идентификация данных как категориальных, в отличие от текстовых, действительно предлагает некоторые преимущества.

- ◆ Знание, что данные категориальные, может служить сигналом для программной системы о том, каким образом должны вести себя статистические процедуры, такие как создание графика или подгонка модели. В частности, в R и Python порядковые данные могут быть представлены как порядковый фактор `ordered.factor`, сохраняя определенную пользователем упорядоченность в графиках, таблицах и моделях.
- ◆ Могут быть оптимизированы хранение и индексация данных (как в реляционной базе данных).
- ◆ Возможные значения, которые принимает конкретная категориальная переменная, реализуются в ПО (как, например, перечисление `enum`).

Третье "преимущество" может привести к непреднамеренному или неожиданному поведению: по умолчанию функции импорта данных в R (например, `read.csv`) ведут себя таким образом, что автоматически преобразуют столбец текста в `factor`. Последующие операции на этом столбце будут исходить из предположения, что единственно допустимыми значениями для этого столбца являются значения, которые были импортированы первоначально, и присвоение нового текстового значения выдаст предупреждение и сообщение об отсутствии значения NA (not available).

Ключевые идеи для структурированных данных

- В программной системе данные обычно классифицируются по типу.
- Тип данных может быть непрерывным, дискретным, категориальным (который включает двоичный тип) и порядковым.
- Типизация данных в программной системе сигнализирует программной системе, каким образом обрабатывать данные.

Дополнительные материалы для чтения

- ◆ Типы данных могут вызывать путаницу, поскольку одни и те же данные можно отнести к разным типам, и таксономия в одной программной системе может отличаться от таксономии в другой. Веб-сайт R-tutorial знакомит с таксономией языка R (<http://www.r-tutor.com/r-introduction/basic-data-types>).
- ◆ В БД применяется более подробная классификация типов данных, включая учет уровней прецизионности, фиксированную либо переменную длину полей и мно-

Прямоугольные данные

В науке о данных типичной опорной конструкцией для анализа является объект с *прямоугольными данными* наподобие электронной таблицы или таблицы базы данных.

Ключевые термины

Кадр данных (data frame)

Прямоугольные данные (как в электронной таблице) — это типичная структура данных для статистических и машинно-обучаемых моделей.

Признак (feature)

Столбец в таблице принято называть *признаком*.

Синонимы: атрибут, вход, предиктор, переменная.

Исход (outcome)

Многие проекты науки о данных сопряжены с предсказанием *исхода* — нередко в формате да/нет (например, в табл. 1.1 это ответ на вопрос "были ли торги состязательными или нет?"). Для предсказания *исхода* в эксперименте или статистическом исследовании иногда используются *признаки*.

Синонимы: зависимая переменная, отклик, цель, выход.

Записи (records)

Строку в таблице принято называть *записью*.

Синонимы: случай, образец, прецедент, экземпляр, наблюдение, шаблон, паттерн, выборка.

Прямоугольные данные по существу представляют собой двумерную матрицу, в которой строки обозначают записи (случаи), а столбцы — признаки (переменные). Исходно данные поступают в такой форме не всегда: неструктурированные данные (например, текст) необходимо обработать и привести к такому виду, чтобы их можно было представить как набор признаков в прямоугольных данных (см. разд. "Элементы структурированных данных" ранее в этой главе). Данные в реляционных БД должны быть извлечены и помещены в единственную таблицу для большинства аналитических и модельных задач.

В табл. 1.1 показана смесь измерительных и количественных данных (например, длительность и цена) и категориальных данных (например, категория и валюта). Как уже упоминалось ранее, специальной формой категориальной переменной является двоичная переменная (да/нет или 0/1), которую можно увидеть в самом правом столбце табл. 1.1 — это индикаторная переменная, показывающая, были ли торги состязательными.

Таблица 1.1. Типичный формат данных

Категория	Валюта	Рейтинг продавца	Длительность	День закрытия	Цена закрытия	Цена открытия	Конкурентно-способность?
Music/Movie/Game	US	3249	5	Mon	0,01	0,01	0
Music/Movie/Game	US	3249	5	Mon	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	1
Automotive	US	3115	7	Tue	0,01	0,01	1

Кадры данных и индексы

Традиционные таблицы БД имеют один или несколько столбцов, называемых *индексом*. Он может значительно повысить эффективность определенных SQL-запросов. В Python при использовании библиотеки `pandas` основной прямоугольной структурой данных является объект `DataFrame`, содержащий таблицу данных. По умолчанию для `DataFrame` создается автоматический целочисленный индекс, который основывается на порядке следования строк. В программной библиотеке `pandas` для повышения эффективности определенных операций также можно задавать многоуровневые/иерархические индексы.

В R основной прямоугольной структурой данных является объект `data.frame`, кадр данных. Объект `data.frame` тоже имеет неявный целочисленный индекс на основе порядка следования строк. Хотя посредством атрибута `row.names`¹ можно создать пользовательский ключ, нативный (родной) для R, объект `data.frame` не поддерживает задаваемые пользователем или многоуровневые индексы. Для преодоления этого недостатка широкое распространение получили два новых программных пакета: `data.table` и `dplyr`. Оба пакета поддерживают многоуровневые индексы и обеспечивают значительное ускорение в работе с объектом `data.frame`.



Различия в терминологии

Терминология для прямоугольных данных может вызывать путаницу. В статистике и науке о данных используются разные термины, которые говорят об одном и том же. Для статистиков в модели существуют предикторные переменные, которые используются для предсказания отклика либо зависимой переменной. В отличие от них для аналитика данных существуют признаки,

¹ Атрибут кадра данных `row.names` — это символьный вектор длиной, которая соответствует числу строк в кадре данных, без дубликатов и пропущенных значений. — *Прим. пер.*

которые применяются для предсказания целевой переменной. В особенности сбивает с толку один синоним: специалисты по информатике используют термин "выборка" для обозначения одиночной строки, тогда как для статистика выборка означает набор строк.

Непрямоугольные структуры данных

Помимо прямоугольных данных существуют и другие структуры данных.

Временной ряд содержит последовательные данные измерений одной и той же переменной. Эти данные представляют собой сырой материал для статистических методов предсказания, и они также являются ключевым компонентом данных, производимых устройствами — Интернет вещей.

Пространственные структуры данных, которые используются в картографической и геопространственной аналитике, более сложны и вариативны, чем прямоугольные структуры данных. В их *объектном* представлении центральной частью данных являются объект (например, дом) и его пространственные координаты. В *полевой* проекции, в отличие от него, основное внимание уделяется небольшим единицам пространства и значению соответствующего метрического показателя (яркости пиксела, например).

Графовые (или сетевые) структуры данных используются для представления физических, социальных и абстрактных связей. Например, граф социальной сети, такой как Facebook или LinkedIn, может представлять связи между людьми в сети. Соединенные дорогами центры распределения являются примером физической сети. Графовые структуры широко применяются в определенных типах задач, таких как оптимизация сети и рекомендательные системы.

В науке о данных каждый из этих типов данных имеет свою специализированную методологию. В центре внимания этой книги находятся прямоугольные данные — основополагающий структурный элемент в предсказательном моделировании.



Графы и графики в статистике

В информатике и информационных технологиях термин "граф" (graph) обычно обозначает описание связей среди объектов и основную структуру данных. В статистике термин "график" (graph) используется для обозначения самых разных графиков и визуализаций, а не только связей между объектами, и этот термин применяется исключительно для обозначения визуализаций, а не структуры данных.

Ключевые идеи для прямоугольных данных

- В науке о данных базовой структурой данных является прямоугольная матрица, в которой строки — это записи, а столбцы — переменные (признаки).
- Терминология может вызывать путаницу, поскольку существует множество синонимов, вытекающих из разных дисциплин, которые вносят свой вклад в науку о данных (статистика, информатика и информационные технологии).

Дополнительные материалы для чтения

- ◆ Документация по кадрам данных в R (<https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html>).
- ◆ Документация по кадрам (таблицам) данных в Python (<http://pandas.pydata.org/pandas-docs/stable/dsintro.html#dataframe>).

Оценки центрального положения

Переменные с измерительными или количественными данными могут иметь тысячи различных значений. Основной этап исследования данных состоит в получении "типичного значения" для каждого признака (переменной): оценки того, где расположено большинство данных (т. е. их центральной тенденции).

Ключевые термины

Среднее (mean)

Сумма всех значений, деленная на количество значений.

Синоним: среднее арифметическое.

Среднее взвешенное (weighted mean)

Сумма произведений всех значений на их веса, деленная на сумму весов.

Синоним: среднее арифметическое взвешенное.

Медиана (median)

Такое значение, при котором половина сортированных данных находится выше и ниже данного значения.

Синоним: 50-й процентиль.

Медиана взвешенная (weighted median)

Такое значение, при котором половина суммы весов находится выше и ниже сортированных данных.

Среднее усеченное (trimmed mean)

Среднее число всех значений после отбрасывания фиксированного числа предельных значений.

Синоним: обрезанное среднее.

Робастный (robust)

Не чувствительный к предельным значениям.

Синоним: устойчивый.

Выброс (outlier)

Значение данных, которое сильно отличается от большинства данных.

Синоним: предельное значение.

На первый взгляд обобщить данные довольно тривиально: просто взять *среднее арифметическое* данных (см. разд. "Среднее" далее в этой главе). На самом деле, несмотря на то, что среднее вычисляется довольно просто и его выгодно использовать, оно не всегда бывает лучшей мерой центрального значения. По этой причине в статистике были разработаны и популяризированы несколько альтернативных оценок среднего значения.



Метрические и оценочные показатели

В статистике термин "*оценки*" часто используется для значений, вычисляемых из данных, которые находятся под рукой, чтобы отличить то, что мы видим, исходя из этих данных, от теоретически истинного или точного положения дел. Аналитики данных и бизнес-аналитики с большей вероятностью будут называть такие значения *метрическими показателями*, или *метриками*. Эта разница отражает подходы, принятые в статистике, в отличие от науки о данных: учитывается неопределенность, которая лежит в основе статистики, тогда как центром внимания науки о данных являются конкретные деловые или организационные цели. Следовательно, статистики оценивают, а аналитики измеряют.

Среднее

Самой элементарной оценкой центрального положения является среднее значение, или *среднее арифметическое*. Среднее — это сумма всех значений, деленная на число значений. Рассмотрим следующий ряд чисел: {3, 5, 1, 2}. Среднее составит $(3 + 5 + 1 + 2) / 4 = 11 / 4 = 2,75$. Вы часто будете встречать символ $\bar{x}$ (произносится "х с чертой"), который обозначает среднее значение выборки из популяции, или генеральной совокупности. Формула среднего значения для ряда из n значений $x_1, x_2, \dots, x_n$ следующая:

$$\text{Среднее} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$



N (или n) обозначает общее число записей или наблюдений. В статистике это обозначение используется с заглавной буквы, если оно относится к популяции, и строчной, если оно относится к выборке из популяции. В науке о данных это различие не является принципиальным, и поэтому можно увидеть и то и другое.

Разновидностью среднего является *среднее усеченное*, которое вычисляется путем отбрасывания фиксированного числа сортированных значений с каждого конца последовательности и затем взятия среднего арифметического оставшихся значений. Если представить сортированные значения как $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, где $x_{(1)}$ — самое маленькое значение, а $x_{(n)}$ — самое большое, то формула для вычисления усеченного среднего с пропуском p самых малых и самых больших значений будет следующей:

$$\text{Среднее усеченное} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}.$$

Среднее усеченное устраняет влияние предельных значений. Например, в международных состязаниях по прыжкам в воду верхние и нижние баллы пяти судей отбрасываются, и итоговым баллом является среднеарифметический балл трех оставшихся судей [Wikipedia-2016]. Такой подход не дает одному судье манипулировать баллом, возможно, чтобы оказать содействие спортсмену из своей страны. Усеченные средние получили широкое распространение и во многих случаях предпочтительны вместо обычного среднего (см. разд. "Медиана и робастные оценки" далее в этой главе).

Еще один вид среднего значения — это *среднее взвешенное*, которое вычисляется путем умножения каждого значения данных x_i на свой вес w_i и деления их суммы на сумму весов. Формула среднего взвешенного выглядит так:

$$\text{Среднее взвешенное} = \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

Существует два главных побудительных мотива для использования среднего взвешенного.

- ◆ Некоторые значения внутренне более переменчивы, чем другие, и сильно переменчивым наблюдениям придается более низкий вес. Например, если мы берем среднее данных, поступающих от многочисленных датчиков, и один из датчиков менее точен, тогда вес данных от этого датчика можно понизить.
- ◆ Собранные данные не одинаково представляют разные группы, которые мы заинтересованы измерить. Например, в зависимости от того, каким образом проводится онлайн-эксперимент, у нас может не быть набора данных, который точно отражает все группы в базе пользователей. Для того чтобы это исправить, можно придать более высокий вес значениям из тех групп, которые были представлены недостаточно.

Медиана и робастные оценки

Медиана — это число, расположенное в сортированном списке данных ровно посередине. Если число данных четное, срединным значением является то, которое не находится в наборе данных фактически, а является средним арифметическим двух значений, которые делят сортированные данные на верхнюю и нижнюю половины. По сравнению со средним, в котором используются абсолютно все наблюдения, медиана зависит только от значений в центре сортированных данных. Хотя это может выглядеть как недостаток, поскольку среднее значение намного более чувствительно к данным, существует много примеров, в которых медиана является

лучшим метрическим показателем центрального положения. Скажем, мы хотим взглянуть на типичные доходы домохозяйств в округах вокруг озера Вашингтон в Сिएтле. При сравнении округа Медина с округом Уиндермир использование среднего значения дало бы совершенно разные результаты, потому что в Медине живет Билл Гейтс. Если же мы станем использовать медиану, то уже не будет иметь значения, насколько богатым является Билл Гейтс — позиция срединного наблюдения останется той же.

По тем же самым причинам, по которым используется среднее взвешенное, можно вычислить и *медиану взвешенную*. Как и с медианой, мы сначала выполняем сортировку данных, несмотря на то, что с каждым значением данных связан вес. В отличие от срединного числа медиана взвешенная — это такое значение, в котором сумма весов равна для нижней и верхней половин сортированного списка. Как и медиана, взвешенная медиана устойчива к выбросам.

Выбросы

Медиана называется *робастной* оценкой центрального положения, поскольку она не находится под влиянием *выбросов* (предельных случаев), которые могут исказить результаты. Выброс — это любое значение, которое сильно удалено от других значений в наборе данных. Точное определение выброса несколько субъективно, несмотря на то, что в различных сводных данных и графиках используются определенные правила (см. разд. "*Процентили и коробчатые диаграммы*" далее в этой главе). Выброс как таковой не делает значение данных недопустимым или ошибочным (как в предыдущем примере с Биллом Гейтсом). Однако выбросы часто являются результатом ошибок данных, таких как смешивание данных с разными единицами измерения (километры против метров) или плохие показания датчика. Когда выбросы являются результатом неправильных данных, среднее значение приводит к плохой оценке центрального положения, в то время как медиана будет по-прежнему допустимой. В любом случае выбросы должны быть идентифицированы и обычно заслуживают дальнейшего обследования.



Обнаружение аномалий

В отличие от типичного анализа данных, где выбросы иногда информативны, а иногда — досадная помеха, в *обнаружении аномалий* целевыми объектами являются именно выбросы, и значительный массив данных преимущественно служит для определения "нормы", с которой соразмеряются аномалии.

Медиана не единственная робастная оценка центрального положения. На самом деле, для того чтобы предотвратить влияние выбросов, широко используется и среднее усеченное. Например, усечение нижних и верхних 10% данных (общепринятый выбор) обеспечит защиту от выбросов во всех, кроме самых малых, наборах данных. Среднее усеченное может считаться компромиссом между медианой и средним: оно устойчиво к предельным значениям в данных, но использует больше данных для вычисления оценки центрального положения.