

# Введение в статистическое обучение с примерами на языке R

Джеймс Г.

Уиттон Д.

Хасты Т.

Тибширани Р.



# An Introduction to Statistical Learning with Applications in R

Gareth James

Daniela Witten

Trevor Hastie

Robert Tibshirani



# Введение в статистическое обучение с примерами на языке R

Джеймс Г.

Уиттон Д.

Хасты Т.

Тибширани Р.



Москва, 2016

**УДК 519.25/.6:004.434R**

**ББК 22.17с5**

**Д40**

**Джеймс Г., Уиттон Д., Хасты Т., Тибширани Р.**

**Д40** Введение в статистическое обучение с примерами на языке R. Пер. с англ. С. Э. Мاستицкого – М.: ДМК Пресс, 2016. – 436 с.: ил.

**ISBN 978-5-97060-293-5**

Книга представляет собой доступно изложенное введение в статистическое обучение – незаменимый набор инструментов, позволяющих извлечь полезную информацию из больших и сложных наборов данных, которые начали возникать в последние 20 лет в таких областях, как биология, экономика, маркетинг, физика и др. В этой книге описаны одни из наиболее важных методов моделирования и прогнозирования, а также примеры их практического применения. Рассмотренные темы включают линейную регрессию, классификацию, создание повторных выборок, регуляризацию, деревья решений, машины опорных векторов, кластеризацию и др. Описание этих методов сопровождается многочисленными иллюстрациями и практическими примерами. Поскольку цель этого учебника заключается в продвижении методов статистического обучения среди практикующих академических исследователей и промышленных аналитиков, каждая глава включает примеры практической реализации соответствующих методов с помощью R – чрезвычайно популярной среды статистических вычислений с открытым кодом.

Издание рассчитано на неспециалистов, которые хотели бы применять современные методы статистического обучения для анализа своих данных. Предполагается, что читатели ранее прослушали лишь курс по линейной регрессии и не обладают знаниями матричной алгебры.

**УДК 519.25/.6:004.434R**

**ББК 22.17с5**

Translation from the English language edition:

An Introduction to Statistical Learning

by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Copyright © Springer Science+Business Media New York 2013

Springer New York is a part of Springer Science+Business Media.

All Rights Reserved.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

**ISBN 978-1-4614-7137-0 (англ.)**

**ISBN 978-5-97060-293-5 (рус.)**

**Copyright © Springer Science+Business Media New York, 2013**

**© Издание, оформление, перевод, ДМК Пресс, 2016**

# Оглавление

От переводчика	10
Предисловие	11
<b>1 Введение</b>	<b>13</b>
<b>2 Статистическое обучение</b>	<b>27</b>
2.1 Что такое статистическое обучение?	27
2.1.1 Зачем оценивать $f$ ?	29
2.1.2 Как мы оцениваем $f$ ?	33
2.1.3 Компромисс между точностью предсказаний и интерпретируемостью модели	36
2.1.4 Обучение с учителем и без учителя	38
2.1.5 Различия между проблемами регрессии и классификации	40
2.2 Описание точности модели	41
2.2.1 Измерение качества модели	41
2.2.2 Компромисс между смещением и дисперсией	46
2.2.3 Задачи классификации	49
2.3 Лабораторная работа: введение в R	56
2.3.1 Основные команды	56
2.3.2 Графики	59
2.3.3 Индексирование данных	60
2.3.4 Загрузка данных	61
2.3.5 Дополнительные графические и количественные сводки	63
2.4 Упражнения	65
<b>3 Линейная регрессия</b>	<b>71</b>
3.1 Простая линейная регрессия	72
3.1.1 Оценивание коэффициентов	73
3.1.2 Точность оценок коэффициентов	75
3.1.3 Оценивание точности модели	80
3.2 Множественная линейная регрессия	83
3.2.1 Оценивание регрессионных коэффициентов	84
3.2.2 Некоторые важные вопросы	87
3.3 Другие аспекты регрессионной модели	95
3.3.1 Качественные предикторы	95
3.3.2 Расширения линейной модели	99
3.3.3 Потенциальные проблемы	105

3.4	Маркетинговый план . . . . .	116
3.5	Сравнение линейной регрессии с методом $K$ ближайших соседей . . . . .	118
3.6	Лабораторная работа: линейная регрессия . . . . .	123
3.6.1	Библиотеки . . . . .	123
3.6.2	Простая линейная регрессия . . . . .	124
3.6.3	Множественная линейная регрессия . . . . .	127
3.6.4	Эффекты взаимодействия . . . . .	129
3.6.5	Нелинейные преобразования предикторов . . . . .	130
3.6.6	Качественные предикторы . . . . .	132
3.6.7	Написание функций . . . . .	134
3.7	Упражнения . . . . .	135
<b>4</b>	<b>Классификация</b>	<b>143</b>
4.1	Общее представление о классификации . . . . .	143
4.2	Почему не линейная регрессия? . . . . .	144
4.3	Логистическая регрессия . . . . .	146
4.3.1	Логистическая модель . . . . .	147
4.3.2	Оценивание регрессионных коэффициентов . . . . .	149
4.3.3	Предсказания . . . . .	150
4.3.4	Множественная логистическая модель . . . . .	151
4.3.5	Логистическая регрессия для зависимых переменных с числом классов $> 2$ . . . . .	154
4.4	Дискриминантный анализ . . . . .	154
4.4.1	Использование теоремы Байеса для классификации . . . . .	155
4.4.2	Линейный дискриминантный анализ для $p = 1$ . . . . .	155
4.4.3	Линейный дискриминантный анализ для $p > 1$ . . . . .	158
4.4.4	Квадратичный дискриминантный анализ . . . . .	166
4.5	Сравнение методов классификации . . . . .	168
4.6	Лабораторная работа: логистическая регрессия, LDA, QDA и KNN . . . . .	172
4.6.1	Данные по цене акций . . . . .	172
4.6.2	Логистическая регрессия . . . . .	173
4.6.3	Линейный дискриминантный анализ . . . . .	178
4.6.4	Квадратичный дискриминантный анализ . . . . .	180
4.6.5	Метод $K$ ближайших соседей . . . . .	181
4.6.6	Применение к данным по жилым прицепам . . . . .	182
4.7	Упражнения . . . . .	186
<b>5</b>	<b>Методы создания повторных выборок</b>	<b>192</b>
5.1	Перекрестная проверка . . . . .	193
5.1.1	Метод проверочной выборки . . . . .	193
5.1.2	Перекрестная проверка по отдельным наблюдениям . . . . .	196
5.1.3	$k$ -кратная перекрестная проверка . . . . .	198
5.1.4	Компромисс между смещением и дисперсией в контексте $k$ -кратной перекрестной проверки . . . . .	201
5.1.5	Перекрестная проверка при решении задач классификации . . . . .	202
5.2	Бутстреп . . . . .	205
5.3	Лабораторная работа: перекрестная проверка и бутстреп . . . . .	209

5.3.1	Метод проверочной выборки . . . . .	209
5.3.2	Перекрестная проверка по отдельным наблюдениям . . . . .	210
5.3.3	$k$ -кратная перекрестная проверка . . . . .	212
5.3.4	Бутстреп . . . . .	212
5.4	Упражнения . . . . .	215
<b>6</b>	<b>Отбор и регуляризация линейных моделей</b>	<b>221</b>
6.1	Отбор подмножества переменных . . . . .	223
6.1.1	Отбор оптимального подмножества . . . . .	223
6.1.2	Пошаговый отбор . . . . .	225
6.1.3	Выбор оптимальной модели . . . . .	228
6.2	Методы сжатия . . . . .	234
6.2.1	Гребневая регрессия . . . . .	234
6.2.2	Лассо . . . . .	239
6.2.3	Выбор гиперпараметра . . . . .	248
6.3	Методы снижения размерности . . . . .	250
6.3.1	Регрессия на главные компоненты . . . . .	251
6.3.2	Метод частных наименьших квадратов . . . . .	258
6.4	Особенности работы с данными большой размерности . . . . .	259
6.4.1	Данные большой размерности . . . . .	259
6.4.2	Что не так с большими размерностями? . . . . .	261
6.4.3	Регрессия для данных большой размерности . . . . .	263
6.4.4	Интерпретация результатов в задачах большой размерности . . . . .	264
6.5	Лабораторная работа 1: методы отбора подмножеств переменных . . . . .	265
6.5.1	Отбор оптимального подмножества . . . . .	265
6.5.2	Отбор путем пошагового включения и исключения переменных . . . . .	269
6.5.3	Нахождение оптимальной модели при помощи методов проверочной выборки и перекрестной проверки . . . . .	270
6.6	Лабораторная работа 2: гребневая регрессия и лассо . . . . .	273
6.6.1	Гребневая регрессия . . . . .	273
6.6.2	Лассо . . . . .	277
6.7	Лабораторная работа 3: регрессия при помощи методов PCR и PLS . . . . .	278
6.7.1	Регрессия на главные компоненты . . . . .	278
6.7.2	Регрессия по методу частных наименьших квадратов . . . . .	280
6.8	Упражнения . . . . .	282
<b>7</b>	<b>Выходя за пределы линейности</b>	<b>288</b>
7.1	Полиномиальная регрессия . . . . .	289
7.2	Ступенчатые функции . . . . .	291
7.3	Базисные функции . . . . .	292
7.4	Регрессионные сплайны . . . . .	294
7.4.1	Кусочно-полиномиальная регрессия . . . . .	294
7.4.2	Ограничения и сплайны . . . . .	295

7.4.3	Представление сплайнов с помощью базисных функций . . . . .	296
7.4.4	Выбор числа и расположения узлов . . . . .	298
7.4.5	Сравнение с полиномиальной регрессией . . . . .	299
7.5	Сглаживающие сплайны . . . . .	300
7.5.1	Общее представление о сглаживающих сплайнах . . . . .	300
7.5.2	Нахождение параметра сглаживания $\lambda$ . . . . .	302
7.6	Локальная регрессия . . . . .	304
7.7	Обобщенные аддитивные модели . . . . .	306
7.7.1	GAM для регрессионных задач . . . . .	307
7.7.2	GAM для задач классификации . . . . .	310
7.8	Лабораторная работа: нелинейные модели . . . . .	312
7.8.1	Полиномиальная регрессия и ступенчатые функции . . . . .	312
7.8.2	Сплайны . . . . .	317
7.8.3	GAM . . . . .	319
7.9	Упражнения . . . . .	322
<b>8</b>	<b>Методы, основанные на деревьях решений</b>	<b>328</b>
8.1	Деревья решений: основные понятия . . . . .	328
8.1.1	Регрессионные деревья . . . . .	329
8.1.2	Деревья классификации . . . . .	337
8.1.3	Сравнение деревьев с линейными моделями . . . . .	339
8.1.4	Преимущества и недостатки деревьев решений . . . . .	341
8.2	Бэггинг, случайные леса, бустинг . . . . .	342
8.2.1	Бэггинг . . . . .	342
8.2.2	Случайные леса . . . . .	347
8.2.3	Бустинг . . . . .	349
8.3	Лабораторная работа: деревья решений . . . . .	351
8.3.1	Построение деревьев классификации . . . . .	351
8.3.2	Построение регрессионных деревьев . . . . .	355
8.3.3	Бэггинг и случайные леса . . . . .	356
8.3.4	Бустинг . . . . .	358
8.4	Упражнения . . . . .	359
<b>9</b>	<b>Машины опорных векторов</b>	<b>364</b>
9.1	Классификатор с максимальным зазором . . . . .	364
9.1.1	Что такое гиперплоскость? . . . . .	365
9.1.2	Классификация с использованием гиперплоскости . . . . .	365
9.1.3	Классификатор с максимальным зазором . . . . .	368
9.1.4	Построение классификатора с максимальным зазором . . . . .	370
9.1.5	Случай, когда разделяющая гиперплоскость не существует . . . . .	370
9.2	Классификаторы на опорных векторах . . . . .	371
9.2.1	Общие представления о классификаторах на опорных векторах . . . . .	371
9.2.2	Более подробное описание классификатора на опорных векторах . . . . .	374
9.3	Машины опорных векторов . . . . .	377



9.3.1	Классификация с использованием нелинейных решающих границ . . . . .	377
9.3.2	Машина опорных векторов . . . . .	378
9.3.3	Применение к данным по нарушению сердечной функции . . . . .	382
9.4	Машины опорных векторов для случаев с несколькими классами . . . . .	383
9.4.1	Классификация типа «один против одного» . . . . .	384
9.4.2	Классификация типа «один против всех» . . . . .	384
9.5	Связь с логистической регрессией . . . . .	384
9.6	Лабораторная работа: машины опорных векторов . . . . .	387
9.6.1	Классификатор на опорных векторах . . . . .	387
9.6.2	Машина опорных векторов . . . . .	391
9.6.3	ROC-кривые . . . . .	393
9.6.4	SVM с несколькими классами . . . . .	395
9.6.5	Применение к данным по экспрессии генов . . . . .	395
9.7	Упражнения . . . . .	397
<b>10</b>	<b>Обучение без учителя</b>	<b>402</b>
10.1	Трудность обучения без учителя . . . . .	402
10.2	Анализ главных компонент . . . . .	403
10.2.1	Что представляют собой главные компоненты? . . . .	404
10.2.2	Альтернативная интерпретация главных компонент .	408
10.2.3	Дополнительный материал по PCA . . . . .	409
10.2.4	Другие приложения PCA . . . . .	414
10.3	Методы кластеризации . . . . .	414
10.3.1	Кластеризация по методу $K$ средних . . . . .	415
10.3.2	Иерархическая кластеризация . . . . .	418
10.3.3	Практические аспекты применения кластеризации . .	429
10.4	Лабораторная работа 1: анализ главных компонент . . . .	432
10.5	Лабораторная работа 2: кластерный анализ . . . . .	434
10.5.1	Кластеризация по методу $K$ средних . . . . .	434
10.5.2	Иерархическая кластеризация . . . . .	436
10.6	Лабораторная работа 3: анализ данных NCI60 . . . . .	438
10.6.1	Применение PCA к данным NCI60 . . . . .	439
10.6.2	Кластеризация наблюдений из набора данных NCI60 . . . . .	441
10.7	Упражнения . . . . .	444

# От переводчика

В последние несколько лет наблюдается небывалый рост объема, скорости получения и сложности данных в самых разных областях жизнедеятельности человека. Неудивительно, что и спрос на специалистов, способных извлечь полезную информацию из этих потоков данных, сегодня высок, как никогда раньше. Важную роль в подготовке таких специалистов играет учебная литература по современным методам статистического анализа. Написать хороший учебник — это титанический труд, однако авторы книги, которую Вы сейчас держите в руках, справились с этой задачей блестяще. Простота изложения материала, многочисленные практические примеры и хорошо продуманные лабораторные работы и упражнения сделали книгу «An Introduction to Statistical Learning with Applications in R» чрезвычайно популярной в академических кругах и среди аналитиков коммерческих организаций во всем мире. Для меня было честью выполнить перевод этой работы, и я рад, что теперь она стала доступной и для русскоязычных читателей.

К сожалению, в первом издании этой книги на русском языке, которое вышло в апреле 2016 г., был найден целый ряд опечаток и ошибок, возникших в ходе верстки<sup>1</sup>. Все обнаруженные с тех пор ошибки были учтены и исправлены в настоящем издании, за что я безмерно благодарен помогавшим с этой работой читателям. В случае обнаружения новых недостатков, сообщайте, пожалуйста, по адресу [rtutorialsbook@gmail.com](mailto:rtutorialsbook@gmail.com).

Я благодарен Дмитрию Мовчану и всей команде «ДМК Пресс» за помощь с подготовкой и изданием этой книги, а также Артему Груздеву, Дмитрию Дерябину и Александру Вишератину за оказанные ими консультации и советы по улучшению первых вариантов рукописи. Наконец, я хотел бы поблагодарить свою жену Светлану за ее поддержку во всех моих начинаниях, одним из которых стала работа над этим переводом.

*Сергей Мاستицкий*

*Лондон, декабрь 2016 г.*

---

<sup>1</sup> Полный список этих опечаток и ошибок можно найти на GitHub-странице книги: <https://github.com/ranalytics/islr-ru>.

# Предисловие

К статистическому обучению относят набор инструментов, предназначенных для моделирования и понимания сложно организованных данных. Это недавно разработанная область статистики, которая развивалась параллельно с достижениями в компьютерных науках и особенно машинном обучении. Данная область охватывает многие методы, включая лассо и разреженную регрессию, классификационные и регрессионные деревья, бустинг и метод опорных векторов.

Одновременно со взрывообразным ростом круга задач, связанных с «большими данными», статистическое обучение стало очень популярным во многих научных областях, а также в маркетинге, финансах и других бизнес-дисциплинах. Люди с навыками статистического обучения очень востребованны.

Одна из первых книг в этой области — «*Основы статистического обучения*» (ОСО)<sup>2</sup> — была опубликована в 2001 г., а в 2009 г. вышло ее второе издание. ОСО стала очень популярной книгой среди не только статистиков, но и специалистов из смежных областей. Одна из причин такой популярности заключается в относительно легкодоступном стиле изложения. Однако ОСО предназначена для людей с основательной математической подготовкой. Новая книга «*Введение в статистическое обучение*» возникла в связи с ощутимой необходимостью в более широком и не таком техническом изложении материала. В этой новой книге мы освещаем многие из тех же тем, которые присутствуют в ОСО, но уделяем основное внимание практическому применению соответствующих методов, а не их математическим деталям. Мы разработали лабораторные работы, иллюстрирующие реализацию каждого метода с использованием статистического пакета R. Эти лабораторные работы позволяют читателю получить ценный практический опыт.

Эта книга подойдет для студентов и магистрантов, углубленно изучающих статистику или родственные дисциплины, а также для представителей других наук, которые желают применять инструменты статистического обучения для анализа своих данных. Ее можно использовать в качестве учебника для курса, длящегося один или два семестра.

Мы благодарим за ценные комментарии следующих читателей черновых вариантов этой книги: Паллави Басу, Александру Чулдецову, Патрика Данахера, Уилла Фитиана, Луэллу Фу, Сэма Гросса, Макса Гразьера Г'Селла, Кортни Паулсон, Ксингао Кыяо, Элизу Шенг, Ноа Симон, Кена Минга Тана и Ксина Лу Тана.

---

<sup>2</sup> Hastie T., Tibshirani R., Friedman J. (2001) The Elements of Statistical Learning. Springer, 745 p.

*«Делать предсказания трудно, особенно в отношении будущего».*

*Йоги Берра*

Джеймс Гарет (Лос-Анджелес, США)

Даниела Уиттен (Сиэтл, США)

Тревор Хасты (Пало Альто, США)

Роберт Тибширани (Пало Альто, США)

# Глава 1

## Введение

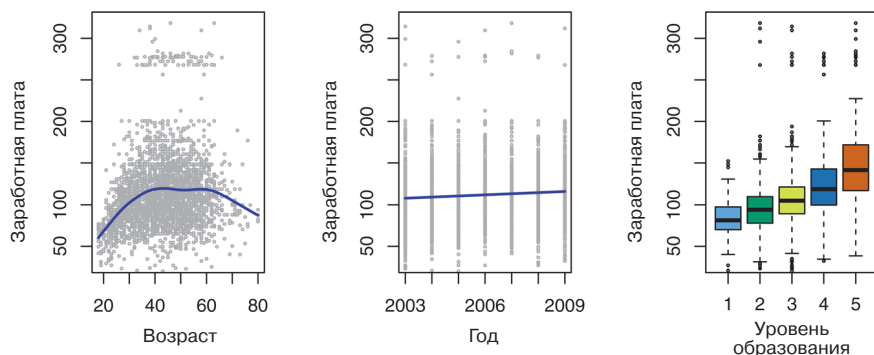
### Обзор задач статистического обучения

Под *статистическим обучением* понимают огромный набор инструментов, предназначенных для *понимания данных*. Эти инструменты можно разделить на две группы: *обучение с учителем* и *обучение без учителя*. В общих чертах статистическое обучение подразумевает построение статистической модели для предсказания, или оценивания, некоторой *выходной переменной* на основе одной или нескольких *входных переменных*. Подобные проблемы встречаются в настолько разнящихся областях, как бизнес, медицина, астрофизика и государственное управление. При обучении без учителя имеются входные переменные, но нет предсказываемой переменной; тем не менее мы можем выявить закономерности и структуру в таких данных. В качестве иллюстрации некоторых практических приложений статистического обучения ниже мы кратко обсудим три реальных набора данных, рассматриваемых в этой книге.

#### *Данные по заработной плате*

В этом примере мы исследуем связь нескольких факторов с уровнем заработной платы у группы мужчин из центрально-атлантического региона США (в этой книге мы будем ссылаться на соответствующие данные как «набор данных *Wage*»). В частности, мы хотим выяснить зависимость между заработной платой работника (переменная *wage*) и его возрастом (*age*), уровнем образования (*education*), а также календарным годом (*year*). Посмотрите, например, на график, представленный слева на рис. 1.1, где показана связь между заработной платой и возрастом работников из этого набора данных. Имеется свидетельство в пользу того, что *wage* увеличивается по мере возрастания *age*, а затем снова снижается примерно после 60 лет. Синяя линия, которая соответствует оценке среднего уровня *wage* для заданного значения *age*, позволяет увидеть этот тренд более четко.

Зная возраст работника, мы можем *предсказать* его заработную плату по этой кривой. Однако на рис. 1.1 также хорошо виден значительный разброс относительно этого среднего значения, из чего следует, что сама по себе переменная *age* вряд ли позволит с большой точностью предсказать *wage* для конкретного человека.



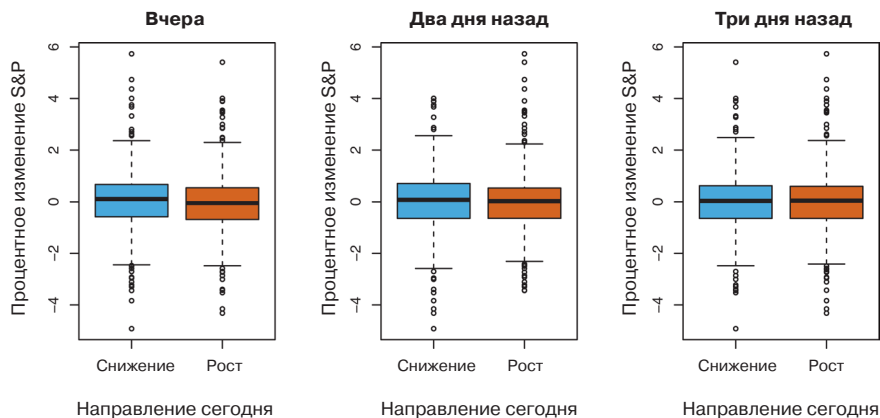
**РИСУНОК 1.1.** Таблица `Wage` с данными по заработной плате мужчин из центрально-атлантического региона США. Слева: `wage` как функция от `age`. В среднем `wage` увеличивается одновременно с `age` до возраста около 60 лет, после чего начинает снижаться. В центре: `wage` как функция от `year`. В период с 2003 по 2009 г. имеет место медленный, но устойчивый рост `wage` в среднем на 10 000\$ в год. Справа: диаграмма размахов `wage` как функции от `education`, где 1 соответствует самому низкому уровню образования (неоконченная средняя школа), а 5 – самому высокому уровню (ученая степень). В среднем `wage` возрастает с уровнем образования

У нас имеется также информация по уровню образования каждого работника и его заработной плате `wage` за каждый год `year`. Графики, представленные в центре и справа на рис. 1.1, показывают `wage` в зависимости от `year` и `education` и свидетельствуют о том, что каждый из этих факторов связан с `wage`. С 2003 по 2009 г. значения зарплаты с каждым годом линейно возрастают примерно на 10 000\$, хотя этот рост очень слабый, по сравнению с разбросом в данных. Зарплаты также выше у людей с более высоким уровнем образования: работники с наименьшим уровнем образования (1) в целом зарабатывают гораздо меньше, чем работники с самым высоким уровнем (5). Очевидно, что наиболее точное предсказание `wage` для конкретного человека будет получено при объединении информации по его возрасту `age`, уровню образования `education` и году `year`. В главе 3 мы обсудим линейную регрессию, которую можно применить для предсказания `wage` по этим данным. В идеале мы должны предсказывать `wage` с учетом нелинейного характера связи этой переменной с `age`. В главе 7 мы рассмотрим класс методов, предназначенных для решения данной проблемы.

### Данные по рынку акций

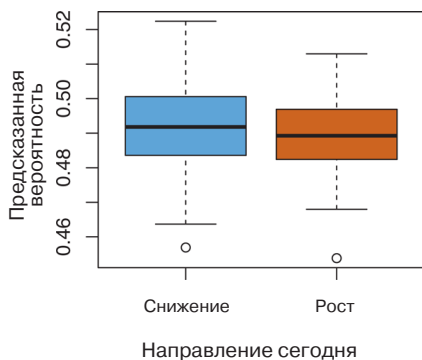
В случае с набором данных `Wage` предсказывается *непрерывное*, или *количественное*, выходное значение. Часто такую ситуацию называют проблемой *восстановления регрессии*. Однако в некоторых случаях мы можем столкнуться с необходимостью предсказать нечисловое значение, т. е. *категориальную*, или *качественную*, выходную переменную. Так, в главе 4 мы рассмотрим набор данных по рынку акций, который описывает днев-

ные изменения индекса Standard & Poor's 500 (S&P) в течение 5-летнего периода (с 2001 по 2005 г.). Мы будем ссылаться на него как на «набор данных Smarket». Задача заключается в предсказании *возрастания* или *снижения* индекса на основе его удельного изменения за последние 5 дней. Здесь проблема статистического обучения не подразумевает предсказания числового значения. Вместо этого предсказывается рост (Up) или снижение (Down) рынка акций для того или иного дня. Это известно как проблема *классификации*. Модель, способная с высокой точностью предсказывать направление движения рынка, была бы очень полезной!



**РИСУНОК 1.2.** Слева: диаграмма размахов, отражающая процентное изменение индекса S&P по сравнению со вчерашним значением для дней, когда происходили рост или снижение рынка (по данным Smarket). В центре и справа: то же, но показаны процентные изменения по сравнению с двумя и тремя предыдущими днями соответственно

На рис. 1.2 слева представлена диаграмма размахов, отражающая процентные изменения индекса акций по сравнению с предыдущим днем: для 648 дней, когда в следующие за ними дни рынок вырос, и для 602 дней, когда рынок ушел вниз. Эти две диаграммы почти идентичны, что указывает на невозможность простой стратегии по использованию вчерашнего состояния индекса S&P для предсказания его сегодняшнего положения. Остальные графики, на которых приведены диаграммы размахов для процентных изменений в сравнении с двумя и тремя предыдущими днями, также указывают на отсутствие выраженной связи между прошлым и текущим состояниями индекса. Безусловно, отсутствие связи здесь ожидаемо, иначе при наличии тесных корреляций между следующими друг за другом днями мы могли бы использовать простую торговую стратегию для получения прибыли. Тем не менее в главе 4 мы подробно исследуем эти данные при помощи нескольких методов статистического обучения. Интересно, что есть некоторые указания на наличие слабых закономерностей в этих данных, предполагающие возможность правильного предсказания направления движения рынка примерно в 60% случаев (по крайней мере, для этого 5-летнего периода; рис. 1.3).



**РИСУНОК 1.3.** Мы подошли к квадратичную дискриминантную модель для части данных *Smarket*, соответствующей периоду с 2001 по 2004 г., и предсказали вероятность снижения рынка акций для данных за 2005 г. В среднем предсказанная вероятность снижения рынка выше для дней, когда снижение в действительности имело место. На основе этих результатов мы можем правильно предсказать направление движения рынка в 60% случаев

### Данные по экспрессии генов

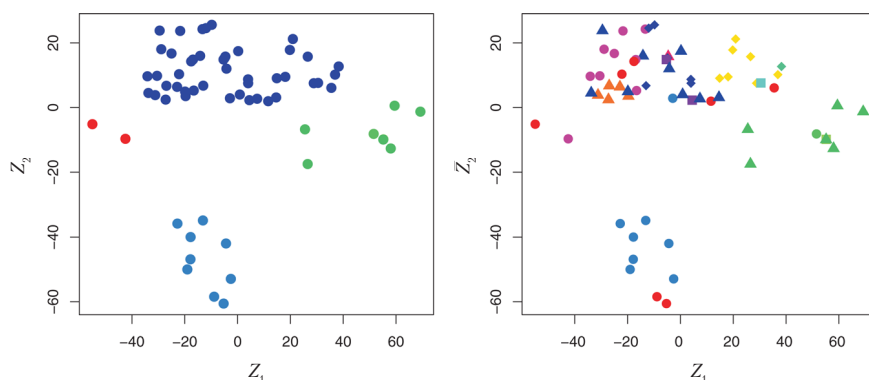
Два предыдущих примера иллюстрируют данные, в которых есть как входные, так и выходные переменные. Однако еще один важный класс проблем охватывает ситуации, в которых мы наблюдаем только входные переменные, без соответствующей зависимой переменной. Например, выполняя маркетинговые исследования, мы могли бы располагать демографической информацией для ряда уже имеющихся или потенциальных клиентов. У нас может возникнуть желание понять, какие клиенты похожи друг на друга, для чего мы объединили бы отдельных людей в группы в соответствии с их наблюдаемыми характеристиками. Такая ситуация известна как *проблема кластеризации*. В отличие от предыдущих примеров, здесь мы не пытаемся предсказать какую-либо выходную переменную.

Мы посвящаем главу 10 обсуждению методов статистического обучения, предназначенных для решения проблем, в которых нет естественной выходной переменной. Мы рассматриваем набор данных *NCI60*, состоящий из 6830 значений уровня экспрессии генов в 64 линиях раковых клеток. Вместо предсказания какой-то конкретной выходной переменной нам интересно выяснить наличие групп, или кластеров, среди этих клеточных линий на основе измерений генной экспрессии. Решить этот вопрос нелегко, отчасти из-за наличия тысяч значений уровня экспрессии для каждой линии, которое затрудняет визуализацию данных.

График, показанный на рис. 1.4 слева, решает эту проблему путем представления каждой из 64 клеточных линий при помощи всего лишь двух чисел —  $Z_1$  и  $Z_2$ . Это первые две *главные компоненты* данных, которые сводят 6830 значений уровня экспрессии по каждой линии до двух чисел, или *измерений*. Несмотря на вероятность потери некоторой части информации в результате такого снижения размерности, теперь появля-



ется возможность визуально исследовать данные на наличие кластеров. Выбор числа кластеров часто бывает трудной проблемой. Однако график, приведенный на рис. 1.4 слева, указывает на наличие не менее четырех групп клеточных линий, которые мы пометили разными цветами. Теперь мы можем подробнее изучить клеточные линии из каждого кластера на предмет их сходства по типу рака и тем самым лучше понять взаимосвязь между уровнями генной экспрессии и раком.



**РИСУНОК 1.4.** Слева: представление данных по уровню экспрессии генов *NCI60* в двумерном пространстве, образованном переменными  $Z_1$  и  $Z_2$ . Каждая точка соответствует одной из 64 клеточных линий. Клеточные линии образуют примерно четыре группы, которые мы представили разными цветами. Справа: то же, что и слева, за тем исключением, что мы выделили каждый из 14 типов рака при помощи символов разной формы и цвета. Клеточные линии, соответствующие одному типу рака, стремятся располагаться в этом двумерном пространстве рядом

Оказывается, что в случае с этим конкретным набором данных клеточные линии соответствуют 14 различным типам рака. (Эта информация, однако, не была использована при создании диаграммы на рис. 1.4 слева.) Справа на рис. 1.4 показаны те же данные, однако 14 типов рака отмечены символами разной формы и цвета. Хорошо видно, что клеточные линии с одинаковым типом рака стремятся располагаться близко друг к другу в этом двумерном представлении. Кроме того, несмотря на то что информация по раку не была использована для создания первого графика, полученная кластеризация в значительной мере соответствует действительным типам рака, наблюдаемым на втором графике. Это в определенной мере является независимым подтверждением верности нашего кластерного анализа.

## Краткая история развития статистического обучения

Несмотря на то что термин «статистическое обучение» достаточно новый, многие из основополагающих концепций этой дисциплины были раз-

работаны очень давно. В начале XIX века Лежандр и Гаусс опубликовали статьи по *методу наименьших квадратов*. Этот подход был впервые успешно применен для решения проблем астрономии. Линейная регрессия используется для предсказания значений количественных переменных, таких, например, как заработная плата. Для предсказания значений качественных переменных (например, выживет пациент или нет, пойдет рынок акций вверх или вниз) Фишер в 1936 г. предложил *линейный дискриминантный анализ*. В 1940–х г. разные авторы предложили альтернативный подход — *логистическую регрессию*. В начале 1970–х г. Нельдер и Вейдербурн ввели термин «*обобщенные линейные модели*» для целого класса методов статистического обучения, которые включают как линейную, так и логистическую регрессию в качестве частных случаев.

К концу 1970–х г. стали доступными многие другие методы обучения на основе данных. Однако почти всегда это были линейные методы, поскольку с вычислительной точки зрения подгонка *нелинейных* зависимостей в то время была неосуществимой. К началу 1980–х гг. вычислительные технологии, наконец, были усовершенствованы до уровня, который больше не ограничивал работу с нелинейными методами. В середине 1980–х г. Брейман, Фридман, Ольшен и Стоун ввели *деревья регрессии и классификации* и стали одними из первых, кто детально продемонстрировал большой потенциал для практической реализации этого метода, включая перекрестную проверку для выбора модели. В 1986 г. Хасте и Тибширани ввели термин «*обобщенные аддитивные модели*» для класса нелинейных дополнений обобщенных линейных моделей, а также разработали соответствующее программное обеспечение.

С тех пор благодаря появлению *машинного обучения* и других дисциплин, статистическое обучение развилось в новую ветвь статистики, уделяющую основное внимание обучению с учителем и без учителя, а также прогнозированию. В последние годы прогресс в статистическом обучении был связан с ростом доступности мощного и относительно удобного программного обеспечения, каковым является популярная и бесплатная система R. Потенциально это может привести к дальнейшей трансформации дисциплины из набора методов, используемых и разрабатываемых статистиками и специалистами в области компьютерных наук, в неотъемлемый набор инструментов для гораздо более широкого сообщества.

## Об этой книге

Книга «Основы статистического обучения» (ОСО), написанная Хасте, Тибширани и Фридманом, была впервые опубликована в 2001 г. С тех пор она превратилась в важную справочную работу по фундаментальным основам статистического обучения. Ее успех обусловлен широким и детальным рассмотрением многих тем статистического обучения, а также тем фактом, что (в сравнении со многими специализированными учебниками по статистике) она доступна для широкой аудитории. Однако больше всего успех ОСО связан с тематикой этой книги. На момент публикации интерес к области статистического обучения начинал свой взрывообразный рост. ОСО стала одной из первых доступных и всеобъемлющих вводных работ по этой теме.

С момента публикации ОСО статистическое обучение продолжило свой расцвет. Развитие этой дисциплины приняло две формы. Наиболее заметный рост был связан с разработкой новых и усовершенствованных подходов статистического обучения, предназначенных для получения ответов на широкий круг вопросов в ряде научных областей. Однако статистическое обучение расширило также и свою аудиторию. В 1990-х г. рост доступности вычислительных ресурсов вызвал волну интереса к этой области со стороны неспециалистов по статистике, которым не терпелось начать использовать современные статистические инструменты для анализа своих данных. К сожалению, высокотехническая природа этих методов означала, что сообщество их пользователей оставалось ограниченным преимущественно экспертами по статистике, компьютерным и смежным областям, имеющими необходимую подготовку (и время) для освоения и реализации соответствующих методов.

В последние годы новое и усовершенствованное программное обеспечение значительно облегчило практическое применение многих методов статистического обучения. В то же время во многих областях, таких как бизнес, здравоохранение, генетика, социальные науки и т. д., произошло осознание того, что статистическое обучение является мощным инструментом для решения важных практических задач. Как следствие оно перестало быть чем-то, что представляет преимущественно академический интерес, и превратилось в популярную дисциплину с огромной потенциальной аудиторией. Несомненно, этот тренд продолжится по мере роста доступности огромных объемов данных и программного обеспечения, предназначенного для их анализа.

Цель книги *«Введение в статистическое обучение»* (ВСО) состоит в том, чтобы содействовать превращению статистического обучения из академической в популярную практическую дисциплину. ВСО не предназначена для замены ОСО, которая является гораздо более обстоятельной работой как по числу рассматриваемых в ней методов, так и по глубине их описания. Мы рассматриваем ОСО в качестве справочника для профессионалов (имеющих ученые степени по статистике, машинному обучению или сходным направлениям), которым необходимо понимать технические детали, лежащие в основе подходов статистического обучения. Однако сообщество пользователей методов машинного обучения расширилось и включает людей с более широким кругом интересов и с разным образованием. Поэтому мы убеждены, что сейчас появилось место для менее технической и более доступной версии ОСО.

В ходе преподавания этих тем на протяжении многих лет мы обнаружили, что они представляют интерес для магистрантов и аспирантов из настолько далеких друг от друга дисциплин, как бизнес-администрирование, биология и компьютерные науки, а также для ориентированных на количественные дисциплины студентов старших курсов. Для этой разнородной аудитории важно иметь возможность понимать модели, их предпосылки, а также сильные и слабые стороны различных методов. Однако многие технические детали методов статистического обучения, такие как алгоритмы оптимизации и теоретические свойства методов, для этой аудитории не представляют большого интереса. Мы убеждены, что таким студентам не нужно иметь глубокого понимания этих аспектов, для того чтобы начать осознанно применять различные методы и сделать

вклад в соответствующие научные дисциплины с помощью инструментария статистического обучения.

Книга ВСО основана на следующих четырех предпосылках.

1. *Многие методы статистического обучения применимы и полезны для широкого круга академических и практических дисциплин, выходящих далеко за рамки статистической науки.* Мы убеждены, что многие современные процедуры статистического обучения должны стать (и станут) настолько же широко доступными и используемыми, как классические методы наподобие линейной регрессии. В связи с этим вместо попытки охватить все возможные подходы (а это невыполнимая задача) мы сконцентрировались на представлении методов, которые считаем наиболее широко применимыми.
2. *Статистическое обучение не следует рассматривать как набор «черных ящиков».* Не существует метода, который одинаково хорошо сработает во всех возможных ситуациях. Без понимания всех «винтиков» внутри «ящика» и без взаимодействия с этими «винтиками» невозможно выбрать наилучший «ящик». Поэтому мы предприняли попытку тщательно описать модель, идею, допущения и компромиссы, лежащие в основе каждого рассматриваемого нами метода.
3. *Несмотря на важность понимания функции, выполняемой каждым «винтиком», нет необходимости уметь конструировать саму машину, находящуюся внутри «ящика».* Поэтому мы минимизировали обсуждение технических деталей, имеющих отношение к процедурам подгонки моделей и теоретическим свойствам методов. Мы предполагаем, что читатель чувствует себя комфортно с простейшими математическими концепциями, но мы не ожидаем от него ученой степени в области математических наук. Например, мы почти полностью исключили использование матричной алгебры, и всю книгу можно понять без знания матриц и векторов.
4. *Мы предполагаем, что читатель интересуется применением методов статистического обучения для решения практических проблем.* Чтобы удовлетворить этот интерес и мотивировать к применению обсуждаемых методов, после каждой главы мы приводим раздел с лабораторными работами. В каждой лабораторной работе мы знакомим читателя с реалистичным практическим применением методов, рассмотренных в соответствующей главе. Когда мы преподавали этот материал в наших курсах, мы отводили на лабораторные работы примерно треть всего времени и нашли их чрезвычайно полезными. Многие студенты, которые поначалу испытывали затруднения при работе с командным интерфейсом R, усвоили необходимые навыки в течение семестра. Мы использовали R потому, что эта система является бесплатной и достаточно мощной для реализации всех рассмотренных в книге методов. Кроме того, она имеет расширения, которые можно загрузить для реализации буквально тысяч дополнительных методов. Но важнее всего то, что R предпочитают академические статистики, и новые методы часто становятся доступными в R на несколько лет раньше того, как они появляются в платных

программах. Тем не менее лабораторные работы в ВСО автономны, и их можно пропускать, если читатель желает использовать другое программное обеспечение или не намерен применять обсуждаемые методы к реальным проблемам.

## Кому следует прочесть эту книгу?

Эта книга предназначена для всех, кто интересуется применением современных статистических методов для моделирования и прогнозирования на основе данных. Эта группа читателей включает ученых, инженеров, финансовых аналитиков, а также людей с меньшей технической и математической подготовкой, которые имеют образование в таких областях, как социальные науки или бизнес. Мы ожидаем, что читатель прослушал как минимум один вводный курс по статистике. Знание линейной регрессии также полезно, но не обязательно, поскольку в главе 3 мы даем обзор ключевых концепций, лежащих в основе этого метода. Уровень математики в этой книге умеренный, и детальное знание матричных операций не требуется. Книга содержит введение в язык статистического программирования R. Предыдущий опыт программирования на другом языке, вроде MATLAB или Python, полезен, но не обязателен.

Мы успешно преподавали материал на этом уровне магистрантам и аспирантам, изучающим бизнес, компьютерные науки, биологию, науки о Земле, психологию и многие другие направления естественных и гуманитарных наук. Эта книга также могла бы оказаться подходящей для студентов последних курсов, которые уже прослушали курс по линейной регрессии. В контексте математически более строгого курса, где основным учебником является ОСО, ВСО можно было бы использовать в качестве дополнительного источника для преподавания вычислительных аспектов различных методов.

## Обозначения и простая матричная алгебра

Выбор системы обозначений для учебника — это всегда сложная задача. В большинстве случаев мы применяем те же условные обозначения, что и в ОСО.

Мы будем использовать  $n$  для обозначения числа отдельных значений, или наблюдений, в нашей выборке. При помощи  $p$  мы будем обозначать число имеющихся переменных, на основе которых можно делать предсказания. Например, набор данных `Wage` состоит из 12 переменных для 3000 людей, так что у нас есть  $n = 3000$  наблюдений и  $p = 12$  переменных, таких как `year`, `age`, `wage` и др. Заметьте, что на протяжении всей этой книги для обозначения имен переменных мы используем цветной шрифт: **Имя Переменной**.

В некоторых примерах  $p$  может быть довольно большим, порядка нескольких тысяч или даже миллионов; подобная ситуация достаточно часто возникает, например, при анализе современных биологических данных или данных по интернет-рекламе.

Обычно при помощи  $x_{ij}$  мы будем обозначать  $i$ -е значение  $j$ -й переменной, где  $i = 1, 2, \dots, n$ , а  $j = 1, 2, \dots, p$ . На протяжении этой книги

$i$  будет использоваться для индексирования выборок или отдельных наблюдений (от 1 до  $n$ ), а  $j$  — для индексирования переменных (от 1 до  $p$ ). С помощью  $\mathbf{X}$  мы обозначаем матрицу размером  $n \times p$ , чей  $(i, j)$ -й элемент — это  $x_{ij}$ . Другими словами,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Читателям, не знакомым с матрицами, полезно будет мысленно представлять  $\mathbf{X}$  в виде таблицы чисел с  $n$  строками и  $p$  столбцами.

В ряде случаев нам будут интересны строки матрицы  $\mathbf{X}$ , которые мы записываем как  $x_1, x_2, \dots, x_n$ . Здесь  $x_i$  представляет собой вектор длиной  $p$ , содержащий значения  $p$  переменных для  $i$ -го наблюдения. Другими словами,

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}. \quad (1.1)$$

(По определению, векторы представлены в виде столбцов.) Например, для набора данных `Wage`  $x_i$  — это вектор длиной 12, состоящий из значений `year`, `age`, `wage` и других переменных для  $i$ -го человека. В других случаях вместо строк нам будут интересны столбцы матрицы  $\mathbf{X}$ , которые мы записываем как  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ . Каждый из этих столбцов является вектором длиной  $n$ , т. е.

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

Например, в случае с данными `Wage`  $\mathbf{x}_1$  содержит  $n = 3000$  значений `year`.

Используя эту нотацию, матрицу  $\mathbf{X}$  можно записать как

$$\mathbf{X} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_p),$$

или

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

Символ  $^T$  обозначает *транспозицию* матрицы или вектора. Так, например,

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix},$$

тогда как

$$x_i^T = (x_{i1} \ x_{i2} \ \cdots \ x_{ip}).$$

Мы используем  $y_i$  для обозначения  $i$ -го наблюдения переменной, которую мы хотим предсказать (например, **wage**). Следовательно, в векторной форме мы записываем набор всех  $n$  наблюдений как

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Тогда наши наблюдаемые данные состоят из пар  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , где каждый элемент  $x_i$  — это вектор длиной  $p$ . Если  $p = 1$ , то  $x_i$  является просто скаляром.

В этой книге вектор длиной  $n$  всегда будет обозначаться при помощи *прописной буквы, выделенной жирным шрифтом*, т. е.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}.$$

Однако векторы, чья длина отличается от  $n$  (например, векторы признаков длиной  $p$ , как в (1.1)), будут обозначаться при помощи прописных букв, выполненных обычным шрифтом (например,  $a$ ). Скаляры также будут обозначаться при помощи таких *прописных букв*, т. е.  $a$ . В редких случаях, когда использование прописных букв с обычным шрифтом может привести к двусмысленности, мы будем пояснять, что имеется в виду. Матрицы будут обозначаться с использованием *заглавных букв, выполненных жирным шрифтом* (например, **A**). Случайные переменные будут обозначаться *заглавными буквами, выполненными обычным шрифтом* (например,  $A$ ), вне зависимости от их размерности.

Иногда у нас будет возникать необходимость указать размерность конкретного объекта. Чтобы показать, что объект является вектором, мы будем использовать нотацию  $a \in \mathbb{R}$ . Чтобы показать, что это вектор длиной  $k$ , мы будем использовать обозначение  $a \in \mathbb{R}^k$  (или  $a \in \mathbb{R}^n$ , если он имеет длину  $n$ ). Объекты, которые являются матрицами размером  $r \times s$ , мы будем обозначать как  $\mathbf{A} \in \mathbb{R}^{r \times s}$ .

Мы избегали использования матричной алгебры везде, где это было возможно. Однако в нескольких случаях полностью избежать ее становилось слишком обременительно. В этих редких случаях важно понимать концепцию умножения двух матриц. Предположим, что  $\mathbf{A} \in \mathbb{R}^{r \times d}$ ,

а  $\mathbf{B} \in \mathbb{R}^{d \times s}$ . Результат умножения  $\mathbf{A}$  на  $\mathbf{B}$  обозначается как  $\mathbf{AB}$ . Элемент  $(i, j)$  матрицы  $\mathbf{AB}$  вычисляется путем умножения каждой  $i$ -й строки  $\mathbf{A}$  на соответствующий элемент  $j$ -го столбца  $\mathbf{B}$ . Иначе говоря,  $(\mathbf{AB})_{ij} = \sum_{k=1}^d a_{ik}b_{kj}$ . В качестве примера представьте, что

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{и} \quad \mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}.$$

Тогда

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}.$$

Заметьте, что результатом этой операции является матрица размером  $r \times s$ .  $\mathbf{AB}$  можно вычислить только, если число столбцов в  $\mathbf{A}$  равно числу строк в  $\mathbf{B}$ .

## Структура этой книги

Глава 2 вводит основные термины и концепции, лежащие в основе статистического обучения. Эта глава описывает также классификатор на основе *K ближайших соседей* — очень простой метод, который удивительно хорошо работает для многих проблем. Главы 3 и 4 охватывают классические линейные методы регрессии и классификации. В частности, глава 3 дает обзор *линейной регрессии* — фундаментальной отправной точки всех регрессионных методов. В главе 4 мы обсуждаем два наиболее важных классических метода классификации — *логистическую регрессию* и *линейный дискриминантный анализ*.

Центральной проблемой всех случаев использования статистического обучения является выбор наилучшего метода для решения конкретной задачи. Поэтому в главе 5 мы знакомим читателя с *перекрестной проверкой* и *бутстрепом*, которые можно использовать для оценивания точности нескольких методов и выбора самого подходящего из них.


Многие недавние исследования в области статистического обучения были посвящены нелинейным методам. Однако линейные методы часто имеют преимущества по сравнению со своими нелинейными конкурентами в смысле интерпретируемости, а иногда и точности предсказаний. Поэтому в главе 6 мы рассматриваем целый ряд линейных методов, как классических, так и более современных, которые предлагают потенциальные улучшения по сравнению со стандартной линейной регрессией. Речь идет о *пошаговом отборе*, *гребневой регрессии*, *регрессии на главные компоненты*, методе *частных наименьших квадратов* и *лассо-регрессии*.

Остальные главы посвящены миру нелинейного статистического обучения. Сначала в главе 7 мы вводим несколько нелинейных методов, которые хорошо работают для проблем с одной входной переменной. Затем мы показываем, как эти методы можно использовать для построения нелинейных *аддитивных* моделей с несколькими входными переменными. В главе 8 мы исследуем методы, основанные на *решающих деревьях*, включая *бэггинг*, *бустинг* и *случайные леса*. *Метод опорных векторов*



— совокупность подходов для решения задач как линейной, так и нелинейной классификации — обсуждается в главе 9. Наконец, в главе 10 мы рассматриваем ситуацию, когда у нас имеются входные переменные, но нет выходной переменной. В частности, мы описываем *анализ главных компонент*, кластеризацию с помощью *метода K средних*, а также *иерархическую кластеризацию*.

В конце каждой главы мы приводим одну или несколько лабораторных работ с использованием R, в которых подробно разбираем примеры практического применения рассмотренных в этой главе методов. Эти лабораторные работы демонстрируют сильные и слабые стороны разных подходов, а также предоставляют полезный справочный материал по синтаксису, необходимому для реализации разных методов. Возможно, читатель предпочтет работать над этими лабораторными в своем собственном темпе, но это могут быть также и групповые сессии, являющиеся частью классной работы. В каждой лабораторной работе мы представляем результаты, которые получили на момент написания книги. Однако постоянно выходят новые версии R, и со временем используемые в лабораторных работах пакеты будут обновлены. Поэтому возможно, что в будущем представленные в соответствующих разделах результаты больше не будут в точности соответствовать результатам, полученным читателем. По мере необходимости мы будем обновлять лабораторные работы на сайте книги.

Мы используем значок  для обозначения разделов и упражнений повышенной сложности. Эти разделы могут быть легко пропущены читателями, которые не желают так глубоко погружаться в соответствующий материал или не имеют необходимой математической подготовки.

## Данные, использованные в лабораторных работах и упражнениях

В этом учебнике мы иллюстрируем методы статистического обучения на практических примерах из маркетинга, финансов, биологии и других областей. Пакет **ISLR**, доступный на сайте книги, содержит несколько наборов данных, которые необходимы для выполнения соответствующих лабораторных работ и упражнений. Один из таких наборов данных входит в состав библиотеки **MASS**, а еще один является частью базового дистрибутива R. Таблица 1.1 содержит сводную информацию по данным, необходимым для выполнения лабораторных работ и упражнений. Несколько таких наборов данных, используемых в главе 2, доступно также в виде текстовых файлов на сайте книги.

## Веб-сайт книги

Веб-сайт этой книги находится по адресу [www.StatLearning.com](http://www.StatLearning.com). Он содержит целый ряд ресурсов, включая связанный с книгой R-пакет и некоторые дополнительные наборы данных.

**ТАБЛИЦА 1.1.** Список наборов данных, необходимых для выполнения лабораторных работ и упражнений в этом учебнике. Все данные доступны в пакете **ISLR**, за исключением **Boston** и **USArrests** (они входят в состав базового дистрибутива **R**)

Название	Описание
<b>Auto</b>	Расход топлива, мощность двигателя и другая информация по автомобилям
<b>Boston</b>	Стоимость жилья и другая информация по пригородам Бостона
<b>Caravan</b>	Информация по людям, которым была предложена страховка для их жилых прицепов
<b>Carseats</b>	Информация по продажам автомобильных сидений в 400 магазинах
<b>College</b>	Демографические характеристики, стоимость обучения и другие данные по колледжам США
<b>Default</b>	Данные по должникам компании, выпускающей кредитные карты
<b>Hitters</b>	Данные по заработной плате бейсбольных игроков
<b>Khan</b>	Измерения генной экспрессии для четырех типов рака
<b>NCI60</b>	Измерения генной экспрессии для 64 раковых клеточных линий
<b>OJ</b>	Информация по продажам апельсинового сока марок Citrus Hill и Minute Made
<b>Portfolio</b>	Исторические значения финансовых активов, используемые для распределения инвестиционных средств
<b>Smarket</b>	Дневные удельные изменения доходности индекса S&P за 5-летний период
<b>USArrests</b>	Криминальная статистика в расчете на 100 000 жителей в 50 штатах США
<b>Wage</b>	Данные исследования доходов мужчин в центрально-атлантическом регионе США
<b>Weekly</b>	1098 значений недельной доходности рынка акций за 21 год

## Благодарности

Несколько графиков в этой книге было заимствовано из ОСО: рис. 6.7, 8.3 и 10.12. Все остальные графики в книге новые.

## Глава 2

# Статистическое обучение

### 2.1 Что такое статистическое обучение?

Рассмотрим простой пример, который поможет нам приступить к изучению статистического обучения. Представьте себе, что мы являемся консультантами-статистиками, нанятыми некоторым клиентом для разработки рекомендаций по повышению продаж определенного продукта. Набор данных **Advertising** включает сведения по продажам (**sales**) этого продукта в 200 различных регионах, а также по величине регионального бюджета на рекламу продукта в средствах массовой информации (СМИ) трех видов: телевидение (**TV**), радио (**radio**) и газеты (**newspaper**). Эти данные представлены на рис. 2.1. Наш клиент не имеет прямой возможности увеличить продажи. С другой стороны, он может контролировать затраты на рекламу в каждом из трех типов СМИ. Следовательно, если мы обнаружим связь между затратами на рекламу и продажами, то сможем дать рекомендации нашему клиенту по корректированию бюджета на рекламу, что опосредованно приведет к увеличению продаж. Другими словами, наша цель состоит в разработке модели, которую можно будет использовать для верных предсказаний продаж на основе данных по бюджету для трех типов СМИ.

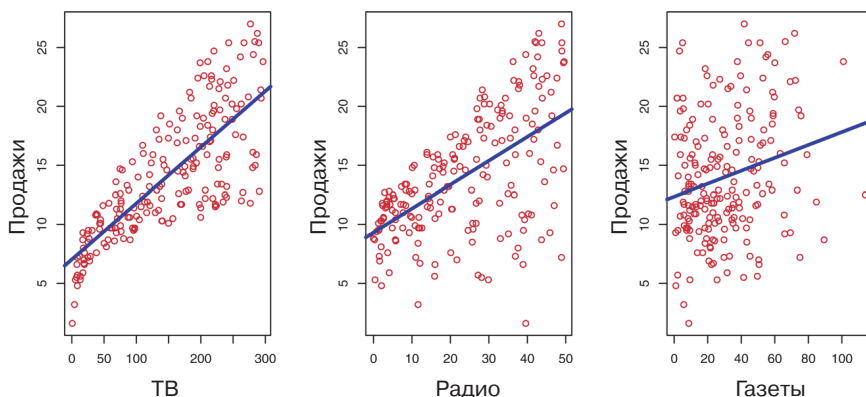
При таком сценарии уровни бюджета на рекламу в разных СМИ являются *входными переменными*, тогда как **sales** — *выходной переменной*. Входные переменные обычно обозначаются при помощи символа  $X$  с нижним индексом, позволяющим различать отдельные переменные. Так,  $X_1$  может обозначать бюджет **TV**,  $X_2$  — бюджет **radio**, а  $X_3$  — бюджет **newspaper**. Входные переменные известны под разными названиями — *предикторы*, *независимые переменные*, *признаки*, или иногда просто *переменные*. Выходную переменную — в данном случае **sales** — часто называют *откликом*, или *зависимой переменной*, и обычно обозначают при помощи символа  $Y$ . В этой книге мы будем использовать указанные термины попеременно.

входная и  
выходная  
переменные

независимая и  
зависимая  
переменные

предиктор  
признак  
отклик

В качестве более общего случая представьте, что мы наблюдаем некоторый количественный отклик  $Y$  и  $p$  отдельных предикторов  $X_1, X_2, \dots, X_p$ . Мы делаем предположение о том, что существует определенная связь между  $Y$  и  $X = (X_1, X_2, \dots, X_p)$ , которую в очень общей форме можно записать как



**РИСУНОК 2.1.** Набор данных *Advertising*. На рисунке показан объем продаж (*sales*, тыс. единиц) в зависимости от величины бюджета на рекламу (тыс. долларов) на телевидении (*TV*), радио (*radio*) и в газетах (*newspaper*) в 200 регионах

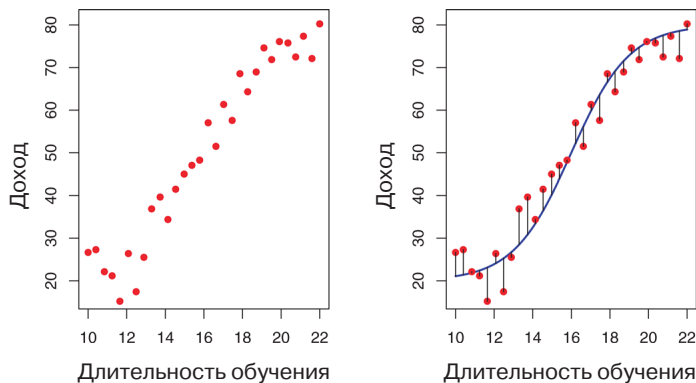
$$Y = f(X) + \epsilon. \quad (2.1)$$

Здесь  $f$  — это некоторая фиксированная, но не известная функция от  $X_1, \dots, X_p$ , а  $\epsilon$  — *ошибка*, которая не зависит от  $X$  и имеет нулевое среднее значение. В таком представлении  $f$  выражает *систематическую* информацию о  $Y$ , содержащуюся в  $X$ .

В качестве другого примера рассмотрим данные для 30 человек из таблицы *Income*, отражающие зависимость дохода (*income*) от количества лет, затраченных этими людьми на образование (*years*; см. рис. 2.2 слева). Этот график предполагает, что мы могли бы предсказать доход на основе длительности обучения. Однако функция, связывающая входную переменную с выходной переменной, в общем случае неизвестна. В такой ситуации мы должны оценить  $f$  на основе имеющихся наблюдений. Поскольку таблица *Income* содержит имитированные данные, то  $f$  известна и показана на рис. 2.2 справа в виде кривой голубого цвета. Вертикальные отрезки соответствуют ошибкам  $\epsilon$ . Заметьте, что некоторые из 30 наблюдений лежат выше голубой линии, а некоторые — ниже, но в целом среднее значение ошибок примерно равно нулю.

В общем случае функция  $f$  может включать более одной входной переменной. На рис. 2.3 мы изображаем *income* как функцию от длительности обучения и стажа (*seniority*). Здесь  $f$  представляет собой двумерную плоскость, которую мы должны оценить на основе имеющихся данных.

В сущности, под статистическим обучением понимают совокупность методов для оценивания  $f$ . В этой главе мы рассматриваем некоторые из ключевых теоретических концепций, применяемых при нахождении  $f$ , а также инструменты для определения качества полученных оценок.



**РИСУНОК 2.2.** Набор данных `Income`. Слева: красные точки соответствуют значениям переменных `income` (доход, десятки тыс. долларов) и `years` (длительность обучения, лет) у 30 человек. Справа: голубая кривая соответствует истинной функции связи между `income` и `years`, которая обычно неизвестна (в этом случае она известна, поскольку данные были имитированы). Черные вертикальные линии показывают ошибки соответствующих наблюдений. Заметьте, что некоторые ошибки положительны (когда наблюдение лежит выше голубой кривой), а некоторые — отрицательны (когда наблюдение находится ниже кривой). В целом среднее значение ошибок примерно равно нулю

### 2.1.1 Зачем оценивать $f$ ?

Существуют две основные причины, по которым мы хотели бы оценить  $f$ : *предсказание* и *статистический вывод*. Обсудим каждую из этих причин по порядку.

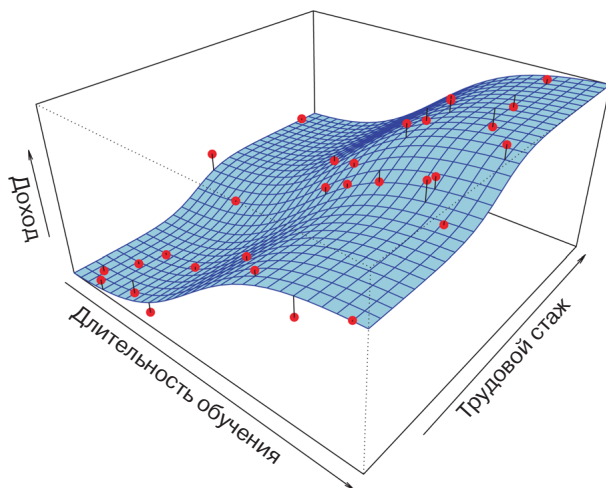
#### Предсказание

Во многих ситуациях набор входных переменных  $X$  легко доступен, однако получить выходную переменную  $Y$  не так просто. Благодаря тому, что ошибки имеют нулевое среднее значение, при таком сценарии мы можем предсказать  $Y$  с помощью

$$\hat{Y} = \hat{f}(X), \quad (2.2)$$

где  $\hat{f}$  представляет собой нашу оценку  $f$ , а  $\hat{Y}$  — предсказанное значение  $Y$ . В подобной ситуации  $\hat{f}$  часто рассматривают как *черный ящик*, в том смысле, что если  $\hat{f}$  обеспечивает верные предсказания  $Y$ , то точная форма этой функции для нас не важна.

В качестве примера предположим, что  $X_1, \dots, X_p$  являются характеристиками образца крови пациента, которые легко измерить в лаборатории, а  $Y$  — это переменная, отражающая риск того, что пациент проявит резко негативную реакцию на определенный лекарственный препарат. Естественным будет желание предсказать  $Y$  по  $X$ , поскольку таким образом



**РИСУНОК 2.3.** График показывает `income` как функцию от `years` и `seniority` (переменные из набора данных `Income`). Голубая плоскость отражает истинную зависимость `income` от `years` и `seniority`, которая известна, поскольку эти данные были имитированы. Красные точки показывают наблюдаемые значения для 30 человек

мы сможем избежать назначения данного препарата пациентам с высоким риском негативной реакции, т. е. пациентам с высокими значениями оценок  $Y$ .

Точность  $\hat{Y}$  в качестве предсказанного значения  $Y$  зависит от двух величин, которые мы будем называть *устраняемой ошибкой* и *неустраняемой ошибкой*. Обычно  $\hat{f}$  не будет идеальной оценкой  $f$ , и эта неточность приведет к возникновению некоторой ошибки. Такая ошибка является *устраняемой*, поскольку потенциально мы можем улучшить точность  $\hat{f}$ , используя более подходящий статистический метод для оценивания  $f$ . Но даже если бы имелась возможность достичь настолько идеальной оценки  $f$ , что  $\hat{Y} = f(X)$ , наше предсказанное значение все равно содержало бы в себе некоторую ошибку! Подобная ошибка известна как *неустраняемая*, поскольку как бы хорошо мы ни оценили  $f$ , мы не сможем снизить ошибку, привнесенную за счет  $\epsilon$ .

Почему же неустраняемая ошибка превышает нулевое значение? Величина  $\epsilon$  может включать неучтенные переменные, которые полезны для предсказания  $Y$ , но поскольку мы их не измеряем, то  $f$  не может использовать их для предсказания. Например, риск негативной реакции может варьировать у того или иного пациента в определенный день в зависимости от условий производства самого лекарственного препарата или от общего состояния здоровья пациента в тот день.

Представьте себе некоторую конкретную оценку  $\hat{f}$  и набор предикторов  $X$ , которые дают предсказание  $\hat{Y} = \hat{f}(X)$ . Допустите на мгновение, что и  $\hat{f}$ , и  $X$  являются фиксированными величинами. Тогда можно легко показать, что

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{устраняемая}} + \underbrace{\text{Var}(\epsilon)}_{\text{неустраняемая}}, \quad (2.3)$$

где  $E(Y - \hat{Y})^2$  представляет собой среднее, или *ожидаемое*, значение квадрата разности между предсказанным и истинным значением  $Y$ , а  $\text{Var}(\epsilon)$  — *дисперсию*, связанную с ошибкой  $\epsilon$ .

ожидаемое  
значение  
дисперсия

В этой книге сделан упор на методы, предназначенные для оценивания  $f$  с целью минимизации устранимой ошибки. Важно помнить, что неустраняемая ошибка всегда будет обеспечивать верхнюю границу точности нашего предсказания  $Y$ . На практике эта граница почти всегда неизвестна.

### Статистический вывод

Часто мы заинтересованы в понимании того, как изменение  $X_1, \dots, X_p$  влияет на  $Y$ . В такой ситуации мы хотим оценить  $f$ , но наша цель не обязательно заключается в получении предсказаний для  $Y$ . Вместо этого мы хотим понять взаимоотношение между  $X$  и  $Y$  или, более конкретно, понять функциональную связь между  $Y$  и  $X_1, \dots, X_p$ . В этом случае  $\hat{f}$  нельзя рассматривать в качестве черного ящика, поскольку нам нужно знать ее точную форму. При таком сценарии мы можем быть заинтересованы в ответе на следующие вопросы:

- *Какие предикторы связаны с откликом?* Часто только небольшая часть имеющихся в распоряжении предикторов тесно связана с  $Y$ . В зависимости от стоящей задачи нахождение ограниченного числа *важных* предикторов в большом наборе возможных переменных может оказаться чрезвычайно полезным.
- *Какова связь между откликом и каждым предиктором?* Некоторые предикторы могут иметь положительную связь с  $Y$  в том смысле, что увеличение предиктора вызывает возрастание значений  $Y$ . Другие предикторы могут оказывать противоположный эффект. В зависимости от сложности  $f$  связь между откликом и некоторым предиктором может зависеть также от значений других предикторов.
- *Можно ли связь между  $Y$  и каждым предиктором адекватно обобщить в виде линейного уравнения, или эта связь является более сложной?* Исторически сложилось так, что большинство методов для оценивания  $f$  подразумевали линейную форму. В некоторых ситуациях такое допущение является обоснованным или даже желательным. Однако часто истинная связь является более сложной и линейная модель не способна обеспечить адекватное представление зависимости между входными и выходными переменными.

В этой книге мы увидим несколько примеров, относящихся к ситуациям, когда требуется выполнение предсказаний, получение статистических выводов или комбинация обеих этих задач.

Например, представьте себе фирму, которая заинтересована в проведении персонализированной маркетинговой кампании. Цель заключается в использовании имеющихся в распоряжении фирмы демографических данных для нахождения людей, которые положительно ответят на высланное по почте предложение. В этом случае демографические переменные служат в качестве предикторов, а отклик на маркетинговую кампанию (положительный или отрицательный) является выходной переменной. Фирма не ставит задачей сформировать глубокое понимание взаимоотношений между каждым отдельным предиктором и откликом — вместо этого ей просто нужна точная модель для предсказания отклика на основе предикторов. Это пример построения модели для получения предсказаний.

С другой стороны, рассмотрим данные из таблицы **Advertising**, изображенные на рис. 2.1. Интерес могут представлять ответы на следующие вопросы:

- *Какие СМИ способствуют продажам?*
- *Какие СМИ вызывают наибольший всплеск продаж?*
- *Насколько тесно рост продаж связан с тем или иным увеличением затрат на телерекламу?*

Такая ситуация подпадает под парадигму статистических выводов. Другой пример включает моделирование бренда продукта, который клиент мог бы купить, исходя из таких переменных, как цена, местоположение магазина, уровни скидок, цена у конкурентов и т. д. В этой ситуации нас больше всего могло бы интересовать то, как каждая отдельная переменная влияет на вероятность покупки. Например, *какое влияние на продажи оказывает изменение цены?* Это пример моделирования с целью сделать статистический вывод.

Наконец, иногда моделирование выполняют для получения как предсказаний, так и статистических выводов. Например, в случае с недвижимым имуществом мы могли бы попытаться увязать значения стоимости домов с такими входными переменными, как уровень преступности, район, расстояние от реки, качество воздуха, доход соседей, размер домов и т. д. В этом случае нам может быть интересно то, как отдельные входные переменные влияют на цены: например, *насколько дороже будет дом с видом на реку?* Это проблема по получению статистического вывода. С другой стороны, нас могло бы интересовать просто предсказание стоимости дома на основе его характеристик: *недо- или переоценен этот дом?* Это проблема предсказания.

В зависимости от того, какова наша главная цель — предсказание, статистический вывод, или комбинация этих двух проблем, — для оценивания  $f$  подходящими могут оказаться разные методы. Например, *линейные модели* позволяют делать относительно простые и интерпретируемые статистические выводы, но в сравнении с другими подходами они могут давать менее точные предсказания. С другой стороны, некоторые из высоконелинейных подходов, обсуждаемых нами в последних главах этой книги, потенциально могут обеспечить очень точные предсказания для  $Y$ , но за счет менее интерпретируемой модели, по которой статистические выводы сделать сложнее.



### 2.1.2 Как мы оцениваем $f$ ?

В этой книге мы рассматриваем многие линейные и нелинейные подходы для оценивания  $f$ . Тем не менее часто эти методы имеют определенные общие характеристики. Мы приводим обзор этих общих черт в данном разделе. Во всех случаях предполагается, что у нас имеется набор из  $n$  отдельных наблюдений. Например, на рис. 2.2 у нас есть  $n = 30$  наблюдений. Эти наблюдения называются *обучающими данными*<sup>1</sup>, поскольку мы используем их для тренировки, или обучения, нашего метода тому, как оценить  $f$ . Пусть  $x_{ij}$  символизирует значение  $j$ -го предиктора, или входной переменной, у наблюдения  $i$ , где  $i = 1, 2, \dots, n$ , а  $j = 1, 2, \dots, p$ . Аналогичным образом пусть  $y_i$  обозначает отклик у  $i$ -го наблюдения. Обучающие данные тогда состоят из пар  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , где  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ .

обучающие данные

Наша цель заключается в применении некоторого метода статистического обучения к входным данным для нахождения неизвестной функции  $f$ . Другими словами, мы хотим найти такую функцию  $\hat{f}$ , что  $Y \approx \hat{f}(X)$  для любого наблюдения  $(X, Y)$ . В общих чертах большинство методов статистического обучения для решения этой задачи можно разделить на *параметрические* и *непараметрические*.

параметрические и непараметрические методы

#### Параметрические методы

Параметрические методы подразумевают основанную на модели процедуру из двух шагов.

1. Во-первых, мы делаем некоторое предположение о функциональной форме  $f$ . Например, одно из простых предположений заключается в том, что  $f$  является линейной функцией от  $X$ :

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p. \quad (2.4)$$

Это *линейная модель*, которая будет детально обсуждаться в главе 3. Как только мы допустили, что  $f$  имеет линейную форму, проблема оценивания  $f$  значительно упрощается. Вместо нахождения совершенно произвольной  $p$ -мерной функции  $f(X)$  нам нужно будет оценить лишь  $p + 1$  коэффициентов  $\beta_0, \beta_1, \dots, \beta_p$ .

2. После выбора модели нам потребуется процедура, которая использует обучающие данные для *подгонки*, или *обучения*, модели. В случае линейной модели (2.4) нам необходимо оценить параметры  $\beta_0, \beta_1, \dots, \beta_p$ . Другими словами, мы хотим найти такие значения этих параметров, при которых

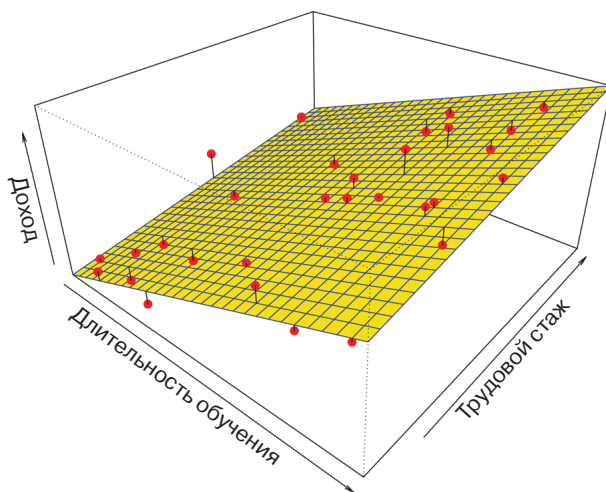
подгонка обучения

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

Наиболее распространенный метод для подгонки модели (2.4) известен как *метод наименьших квадратов*, и мы обсудим его в главе 3. Однако метод наименьших квадратов является лишь одним из многих возможных способов. В главе 6 мы рассмотрим другие подходы для нахождения параметров уравнения (2.4).

метод наименьших квадратов

<sup>1</sup> Синонимами этого термина являются «обучающая выборка» и «обучающее множество», которые также будут использоваться в дальнейшем. — *Прим. пер.*



**РИСУНОК 2.4.** Линейная модель, подогнанная по методу наименьших квадратов к данным *Income* (см. рис. 2.3). Выборочные значения представлены точками красного цвета, а желтая плоскость показывает подогнанную к данным модель

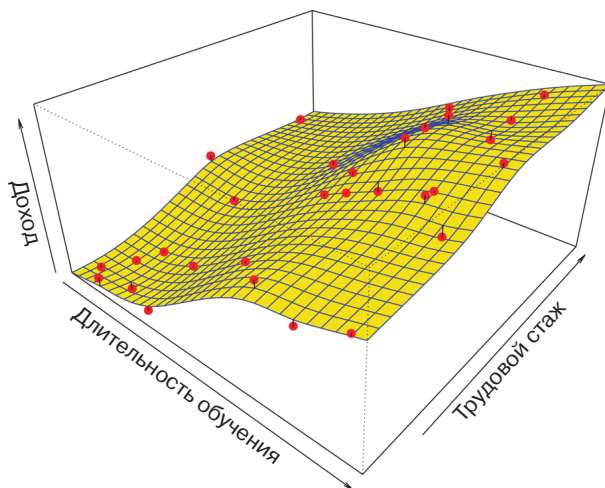
Описанный выше подход, основанный на модели, называется *параметрическим*; он сводит проблему оценивания  $f$  к проблеме оценивания некоторого набора параметров. Допущение о параметрической форме  $f$  упрощает проблему нахождения этой функции, поскольку, как правило, оценить набор параметров  $\beta_0, \beta_1, \dots, \beta_p$  линейной модели (2.4) гораздо проще, чем подогнать совершенно произвольную функцию  $f$ . Потенциальный недостаток параметрического подхода заключается в том, что выбираемая нами модель обычно не будет совпадать с истинной неизвестной формой  $f$ . Если выбранная модель слишком отличается от истинной, то наша оценка  $f$  будет плохой. Мы можем попытаться решить эту проблему, выбрав *гибкие* модели, которые способны описать многие из возможных функциональных форм  $f$ . В целом, однако, подгонка более гибкой модели требует оценивания большего числа параметров. Такие более сложные модели могут приводить к явлению, известному как *переобучение*, которое, в сущности, означает то, что эти модели начинают слишком близко аппроксимировать ошибками, или *шум*, в данных. Эти проблемы обсуждаются на протяжении всей книги.

Рисунок 2.4 демонстрирует пример применения параметрического метода к данным *Income* (см. рис. 2.3). Мы подогнали линейную модель вида

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}.$$

Поскольку мы допустили наличие линейной связи между откликом и этими двумя предикторами, то вся проблема по подгонке модели сводится к оцениванию  $\beta_0, \beta_1$  и  $\beta_2$ , которое мы выполняем с использованием линейной регрессии по методу наименьших квадратов. Сравнивая рис. 2.3 и 2.4, мы можем увидеть, что представленная на рис. 2.4 линейная модель не вполне верна: истинная функция  $f$  имеет некоторый изгиб, который

не отражен в этой линейной модели. Тем не менее линейная модель, похоже, хорошо справляется с задачей по описанию положительной зависимости между `education` и `income`, а также несколько менее выраженной положительной зависимости между `seniority` и `income`. Возможно, что при таком небольшом числе наблюдений это лучшее, что мы можем сделать.



**РИСУНОК 2.5.** Гладкий сплайн типа «тонкая пластина», подогнанный к данным `Income` (см. рис. 2.3), показан желтым цветом; выборочные наблюдения представлены точками красного цвета. Сплайны обсуждаются в главе 7

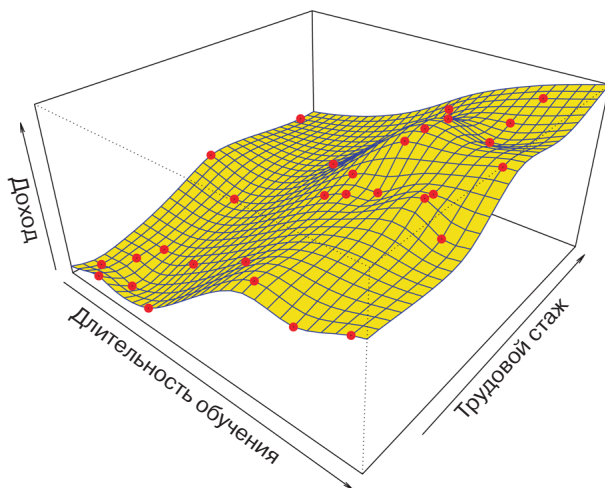
## Непараметрические методы

Непараметрические методы не делают явных предположений в отношении функциональной формы  $f$ . Вместо этого они выполняют поиск такой оценки  $f$ , которая приближается к данным максимально близко, не будучи при этом слишком грубой или слишком извилистой. Такие подходы могут иметь существенное преимущество по сравнению с параметрическими методами: избегая предположения о конкретной функциональной форме  $f$ , они способны точно описать более широкий ряд возможных форм  $f$ . Любой параметрический подход несет с собой возможность того, что используемая для оценивания  $f$  функциональная форма значительно отличается от истинной функции, вследствие чего итоговая модель будет плохо описывать данные. В то же время непараметрические методы полностью уходят от этой опасности, поскольку, по сути, не делается никакого предположения о форме  $f$ . Однако непараметрические методы страдают существенным недостатком: поскольку они не сводят проблему оценивания  $f$  к ограниченному набору параметров, то для получения точной оценки  $f$  требуется очень большое число наблюдений (намного больше, чем обычно необходимо для какого-либо параметрического метода).

Пример применения непараметрического метода для описания данных `Income` показан на рис. 2.5. Для оценивания  $f$  использован сплайн типа

сплайн  
типа  
«тонкая  
пластина»

«тонкая пластина». Этот метод не предписывает  $f$  какую-либо предварительно заданную модель. Вместо этого он пытается получить такую оценку  $f$ , которая максимально близко находилась бы к наблюдаемым данным, при условии что эта аппроксимация является *гладкой*. В рассматриваемом случае непараметрический метод дал удивительно точную оценку истинной функции  $f$ , показанной на рис. 2.3. Для подгонки сплайна типа «тонкая пластина» аналитик должен выбрать степень сглаживания. Рисунок 2.6 изображает результат подгонки такого же сплайна с меньшим уровнем сглаживания, который позволяет получить более неровную плоскость. Полученная функция в точности соответствует данным! Однако сплайн, показанный на рис. 2.6, намного более изменчив в сравнении с истинной функцией  $f$  из рис. 2.3. Это — пример переобучения, о котором мы говорили ранее. Такая ситуация нежелательна, поскольку полученная модель не будет давать точных предсказаний по новым наблюдениям, не входившим в состав исходного набора данных. Мы обсуждаем методы подбора *корректной* степени сглаживания в главе 5. Сплайны обсуждаются в главе 7.



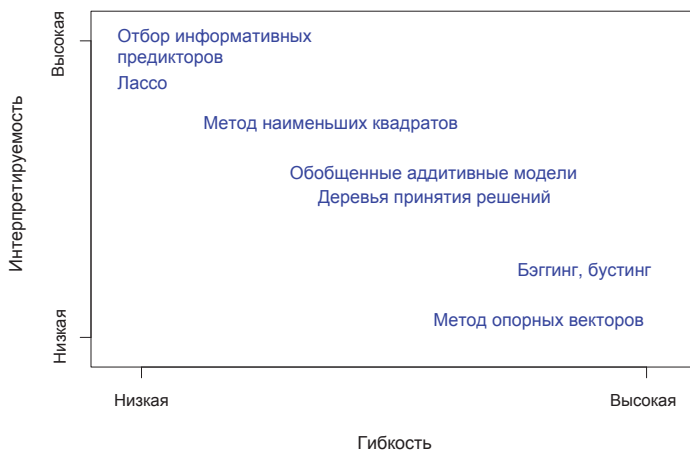
**РИСУНОК 2.6.** Грубый сплайн типа «тонкая пластина», подогнанный к данным *Income* (рис. 2.3). Эта модель не делает ошибок на обучающих данных

Как мы выяснили, параметрические и непараметрические методы имеют свои преимущества и недостатки. На протяжении этой книги мы рассматриваем оба подхода.

### 2.1.3 Компромисс между точностью предсказаний и интерпретируемостью модели

Некоторые из методов, рассматриваемых нами в этой книге, являются менее гибкими, или более ограниченными, в том смысле, что они способны породить лишь относительно небольшой набор функциональных форм

для оценивания  $f$ . Например, линейная регрессия является относительно негибким подходом, поскольку она может порождать только линейные функции вроде линий на рис. 2.1 или плоскости на рис. 2.4. Другие методы, такие как сплайны типа «тонкая пластина», приведенные на рис. 2.5 и 2.6, являются намного более гибкими, так как они могут порождать гораздо более широкий спектр возможных форм для оценивания  $f$ .



**РИСУНОК 2.7.** Один из способов представить компромисс между гибкостью и интерпретируемостью разных методов статистического обучения. В целом при возрастании гибкости метода его интерпретируемость снижается

Резонно мог бы возникнуть следующий вопрос: зачем нам вообще выбирать более ограниченный метод вместо очень гибкого метода? Имеется несколько причин предпочесть более ограниченную модель. Если мы преимущественно заинтересованы в статистических выводах, то ограниченные модели являются гораздо более интерпретируемыми. Например, когда целью является статистический вывод, линейная модель может быть хорошим выбором, поскольку довольно легко будет понять взаимоотношения между  $Y$  и  $X_1, X_2, \dots, X_p$ . В то же время очень гибкие методы, такие как сплайны, обсуждаемые в главе 7 и показанные на рис. 2.5 и 2.6, а также методы бустинга, обсуждаемые в главе 8, могут приводить к настолько сложным оценкам  $f$ , что будет трудно понять, как любой отдельно взятый предиктор связан с откликом.

Рисунок 2.7 иллюстрирует компромисс между гибкостью и интерпретируемостью некоторых методов, рассматриваемых нами в этой книге. Обсуждаемая в главе 3 линейная регрессия по методу наименьших квадратов является относительно негибкой, но при этом вполне интерпретируемой. Метод *лассо*, обсуждаемый в главе 6, основан на линейной модели (2.4), но использует альтернативную процедуру для оценивания коэффициентов  $\beta_1, \beta_2, \dots, \beta_p$ . Эта новая процедура накладывает более жесткие ограничения при нахождении оценок коэффициентов и приравнивает некоторые из них в точности к нулю. Следовательно, в этом смысле лассо является менее гибким подходом, чем линейная регрессия. Кроме того, этот ме-

лассо

обобщен-  
ные  
аддитив-  
ные  
модели

тод легче поддается интерпретации по сравнению с линейной регрессией, поскольку в конечной модели переменная–отклик будет зависеть только от небольшого подмножества предикторов, а именно тех, чьи оценки коэффициентов оказались отличными от нуля. В то же время *обобщенные аддитивные модели* (GAM)<sup>2</sup>, обсуждаемые в главе 7, расширяют (2.4), позволяя моделировать определенные нелинейные зависимости. Как следствие GAM являются более гибкими, чем линейная регрессия. Они также несколько менее интерпретируемы, чем линейная регрессия, поскольку связь между каждым предиктором и откликом в них моделируется с использованием кривой. Наконец, полностью нелинейные методы, такие как

бэггинг  
бустинг  
метод  
опорных  
векторов

*бэггинг*, *бустинг* и *метод опорных векторов* с нелинейными ядрами, обсуждаемые в главах 8 и 9, являются чрезвычайно гибкими подходами, которые трудно интерпретировать.

Мы выяснили, что в случаях, когда целью анализа является статистический вывод, есть явные преимущества в использовании простых и относительно негибких методов статистического обучения. Однако в некоторых ситуациях мы заинтересованы только в предсказании, а интерпретируемость предсказательной модели нам просто неинтересна. Например, если мы пытаемся разработать алгоритм для предсказания цены акций, то единственным нашим требованием к алгоритму будет точность прогноза — интерпретируемость значения не имеет. В такой ситуации мы могли бы ожидать, что лучше всего будет использовать наиболее гибкую из имеющихся моделей. Удивительно, но это не всегда так! Часто мы будем получать более точные предсказания с помощью менее гибкого метода. Это явление, которое на первый взгляд может показаться нелогичным, объясняется склонностью более гибких методов к переобучению. Мы видели пример переобучения на рис. 2.6. Подробнее мы будем обсуждать эту важную концепцию в разделе 2.2 и в других частях книги.

### 2.1.4 Обучение с учителем и без учителя

с учителем  
и без  
учителя

Большинство проблем статистического обучения попадает в одну из двух категорий: *обучение с учителем* и *без учителя*. Все рассмотренные до сих пор примеры из этой главы относятся к категории обучения с учителем. Для каждого измеренного значения предиктора  $x_i$ , где  $i = 1, \dots, n$ , имеется соответствующее значение отклика  $y_i$ . Мы желаем построить модель, которая описывает связь между откликом и предикторами с целью точного предсказания отклика для будущих наблюдений (прогнозирование) или для лучшего понимания взаимоотношений между откликом и предикторами (статистические выводы). Многие классические методы статистического обучения, такие как линейная регрессия и *логистическая регрессия* (глава 4), а также более современные подходы, вроде GAM, бустинга и метода опорных векторов, относятся к категории обучения с учителем. Таким проблемам посвящена большая часть этой книги.

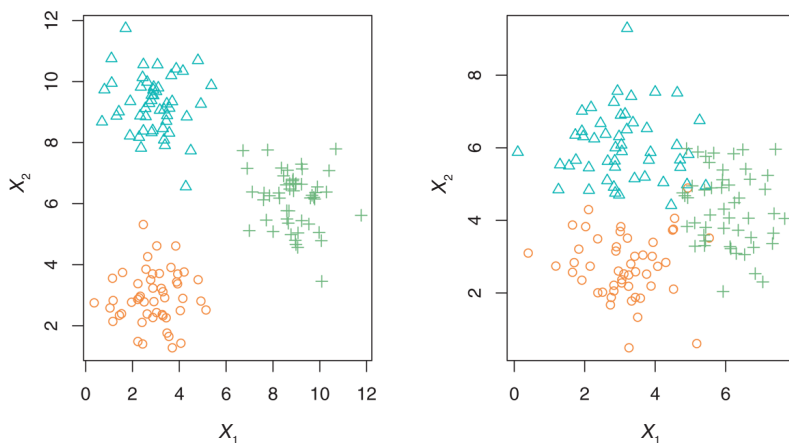
логи-  
стическая  
регрессия

Обучение без учителя, в свою очередь, описывает несколько более сложную ситуацию, в которой для каждого наблюдения  $i = 1, \dots, n$  мы имеем вектор измерений  $x_i$ , но без соответствующего отклика  $y_i$ . Подогнать линейную регрессионную модель невозможно, поскольку подлежащая предсказанию зависимая переменная отсутствует. При таком сценарии

<sup>2</sup> Аббревиатура от «generalized additive models». — Прим. пер.

рии мы в некотором смысле работаем вслепую: подобную ситуацию называют обучением *без учителя*, поскольку у нас нет зависимой переменной, которая «руководила» бы нашим анализом. Какой статистический анализ все же возможен? Мы можем попытаться понять взаимоотношения между переменными или между наблюдениями. Один из статистических инструментов, который мы можем использовать в такой ситуации, — это *кластерный анализ*, или кластеризация. Цель кластерного анализа заключается в установлении того, разделяются ли наблюдения  $x_1, \dots, x_n$  на относительно четко выраженные группы. Например, в исследовании по сегментированию рынка мы могли бы учесть различные характеристики (переменные) для потенциальных клиентов, такие как почтовый индекс, семейный доход и покупательские привычки. Мы могли бы ожидать, что клиенты образуют разные группы, такие как «тратящие много» и «тратящие мало». Если бы информация по стоимости покупок была доступна для каждого клиента, тогда было бы возможным обучение с учителем. Однако такая информация отсутствует, т. е. мы не знаем покупательную способность потенциальных клиентов. При таком сценарии мы можем попытаться выполнить кластерный анализ на основе измеренных переменных с целью определить обособленные группы клиентов. Обнаружение подобных групп может представлять интерес в связи с тем, что они могут различаться в отношении некоторой интересной характеристики, такой как типичные траты на покупки.

кластер-  
ный  
анализ



**РИСУНОК 2.8.** Кластеризация данных, содержащих три группы. Каждая группа показана с использованием разных цветных символов. Слева: три группы хорошо разделены. В такой ситуации кластерный анализ успешно их обнаружит. Справа: группы в определенной степени перекрываются. Здесь задача кластеризации является более сложной

На рис. 2.8 представлен простой пример проблемы кластеризации. Мы изобразили 150 наблюдений со значениями двух переменных —  $X_1$  и  $X_2$ . Каждое наблюдение соответствует одной из трех отдельных групп. В целях демонстрации члены каждой группы обозначены нами в виде разноцветных символов. Однако на практике групповые принадлежности неиз-



вестны, и задача заключается именно в определении группы, к которой относится каждое наблюдение. На рис. 2.8 слева эта задача довольно проста, поскольку группы хорошо разделены. В то же время рисунок справа иллюстрирует более трудную задачу, в которой имеется некоторое перекрытие между группами. Нельзя ожидать, что в этом случае метод кластеризации правильно отнесет все перекрывающиеся точки к их соответствующим группам (голубой, зеленой или оранжевой).

В примерах, показанных на рис. 2.8, есть только две переменные, и поэтому для обнаружения кластеров можно просто визуально изучить диаграммы рассеяния. Однако на практике мы часто сталкиваемся с данными, которые содержат намного больше, чем две переменные. В таких случаях представить наблюдения на графике непросто. Например, если в нашем наборе данных есть  $p$  переменных, то можно будет построить  $p(p-1)/2$  отдельных диаграмм рассеяния, и их визуальное инспектирование для обнаружения кластеров будет просто невыполнимо. По этой причине важную роль играют методы автоматической кластеризации. Мы обсуждаем кластеризацию и другие методы обучения без учителя в главе 10.

Многие проблемы естественным образом распадаются на категории обучения с учителем и без учителя. Однако иногда вопрос о том, к какой из этих двух категорий следует отнести конкретный анализ, менее ясен. Предположим, например, что у нас есть набор из  $n$  наблюдений. Для  $m$  наблюдений, где  $m < n$ , у нас имеются измерения как для предикторов, так и для отклика. Для остальных  $n - m$  наблюдений у нас есть измерения предикторов, но нет отклика. Подобный сценарий может возникать, когда измерение предикторов является относительно дешевым, но сбор информации по соответствующим откликам сопряжен с гораздо большими затратами. Мы называем такую ситуацию проблемой *обучения смешанного типа*. В этом случае желательно применить такой метод статистического обучения, который включает как те  $m$  наблюдений, для которых измерения отклика доступны, так и те  $n - m$  наблюдений, для которых такие измерения отсутствуют. Хотя это очень интересная тема, она лежит за рамками данной книги.

обучение  
смешанно-  
го типа

### 2.1.5 Различия между проблемами регрессии и классификации

типы пере-  
менных

класс

регрессия  
и класси-  
фикация

Переменные можно охарактеризовать либо как *количественные*, либо как *качественные* (последние известны также как «*категориальные*»). Количественные переменные принимают числовые значения. В качестве примеров можно привести возраст, рост и доход человека, стоимость дома и цену акций. В свою очередь, качественные переменные принимают значения, соответствующие одному из  $K$  различных *классов*, или категорий. Примеры качественных переменных: пол человека (мужской или женский), бренд купленного продукта (А, В или С), переход в категорию неплательщиков по долгам (да или нет), а также диагноз ракового заболевания (острая миелоидная лейкемия, острая лимфобластическая лейкемия или отсутствие лейкемии). Обычно проблемы, связанные с количественным откликом, называют проблемами *регрессии*, тогда как проблемы, связанные с качественным откликом, часто называют проблемами *классификации*.



ции. Однако это различие не всегда бывает таким четким. Регрессия по методу наименьших квадратов (глава 3) используется для количественного отклика, тогда как логистическая регрессия обычно применяется для качественного *бинарного* отклика (т. е. переменной с двумя классами). Поэтому логистическая регрессия часто используется как метод классификации. Но поскольку этот метод оценивает вероятности классов, о нем можно думать также и как о методе регрессии. Некоторые статистические методы, такие как метод  $K$  ближайших соседей (главы 2 и 4) и бустинг (глава 8), могут применяться как для количественных, так и для качественных откликов.

Обычно мы выбираем метод на основе того, является ли отклик количественным или качественным; например, мы могли бы использовать линейную регрессию для количественного отклика и логистическую регрессию — для качественного. В то же время тип предикторов (т. е. являются ли они количественными или качественными) обычно считается менее важным. Большинство обсуждаемых в этой книге методов статистического обучения может применяться вне зависимости от типа предикторов, при условии что перед выполнением анализа все качественные предикторы должным образом *закодированы*.

## 2.2 Описание точности модели

Одна из ключевых целей этой книги заключается в том, чтобы представить читателю широкий круг методов статистического обучения, выходящих далеко за рамки стандартного подхода линейной регрессии. Зачем описывать столько разных методов статистического обучения вместо одного, *самого лучшего* метода? В статистике *бесплатный сыр бывает только в мышеловке*: ни один из методов не доминирует над другими для всех возможных наборов данных. На некотором конкретном наборе данных тот или иной метод может оказаться оптимальным, тогда как другие методы могут сработать лучше на похожих или отличающихся данных. Поэтому выбор метода, который дает наилучшие результаты для имеющегося набора данных, представляет собой важную задачу. Выбор такого метода может оказаться одним из самых трудных аспектов применения статистического обучения на практике.

В этом разделе мы обсуждаем некоторые из наиболее важных концепций, возникающих при выборе процедуры статистического обучения для того или иного набора данных. По мере дальнейшего изложения мы будем объяснять, как представленные здесь концепции можно применять на практике.

### 2.2.1 Измерение качества модели

Чтобы оценить эффективность некоторого метода статистического обучения в отношении имеющегося набора данных, нам необходим способ для измерения того, насколько хорошо полученные при помощи этого метода предсказания совпадают с наблюдаемыми данными. Другими словами, нам нужно количественно выразить степень того, насколько предсказанное значение отклика близко к истинному значению отклика у соответ-

средне-  
квадра-  
тичная  
ошибка

ствующего наблюдения. В регрессионном анализе наиболее часто используемой для этого мерой является *среднеквадратичная ошибка* (MSE)<sup>3</sup>, вычисляемая как

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2, \quad (2.5)$$

где  $\hat{f}(x_i)$  — это предсказанное значение, которое  $\hat{f}$  дает для  $i$ -го наблюдения. MSE будет низкой, если предсказанные значения отклика очень близки к истинным значениям, и высокой, если для некоторых наблюдений предсказанные и истинные значения существенно разнятся.

ошибка на  
обучаю-  
щей  
выборке  
контроль-  
ная  
выборка

MSE в (2.5) вычисляется на основе обучающих данных, которые были использованы для подгонки модели, и поэтому точнее будет называть ее *ошибкой на обучающей выборке*<sup>4</sup>. Однако обычно нас не очень заботит то, как хорошо метод работает на обучающих данных. Вместо этого *мы заинтересованы в точности предсказаний, которые мы получаем, когда применяем наш метод к не использованным ранее контрольным данным*. Почему это нас заботит? Представьте, что мы заинтересованы в разработке алгоритма предсказания цены акций на основе их доходности в прошлом. Мы можем обучить некоторый метод, используя значения доходности акций за последние 6 месяцев. Однако нам не очень интересно, насколько хорошо наш метод предсказывает значения цены акций для прошлой недели. Вместо этого нас интересует то, как хорошо он будет предсказывать цену на завтра или для следующего месяца. В том же ключе представьте, что у нас есть измерения клинических показателей (например, вес, давление крови, рост, возраст, семейная история болезни) для нескольких пациентов, а также информация о том, болен ли каждый пациент диабетом. Мы можем использовать этих пациентов, чтобы обучить статистический метод для предсказания риска диабета на основе клинических измерений. В действительности мы хотим, чтобы этот метод верно предсказывал риск диабета для *будущих пациентов* на основе их клинических измерений. Нам не очень интересно, насколько точно метод предсказывает риск диабета для пациентов, использованных для обучения модели, поскольку мы уже знаем, кто из этих пациентов болен.

ошибка на  
контроль-  
ной  
выборке

Чтобы выразить это математически, представьте, что мы подгоняем нашу статистическую модель к обучающим наблюдениям  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  и получаем оценку  $\hat{f}$ . Далее мы можем вычислить  $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$ . Если полученные значения примерно равны  $y_1, y_2, \dots, y_n$ , то MSE на обучающей выборке, рассчитанная по (2.5), окажется низкой. Однако нам совсем нет дела то того, что  $\hat{f}(x_i) \approx y_i$ ; вместо этого мы хотим знать, является ли величина  $\hat{f}(x_0)$  примерно равной  $y_0$ , где  $(x_0, y_0)$  — *неизвестное ранее проверочное наблюдение, которое не использовалось при обучении статистической модели*. Мы хотим выбрать такой метод, который дает минимальную *ошибку на контрольной выборке*<sup>5</sup>, а не минимальную ошибку на обучающей выборке. Другими словами,

<sup>3</sup> Аббревиатура от «mean squared error». — Прим. пер.

<sup>4</sup> Синонимами являются также термины «ошибка на обучающих данных», «ошибка на обучающем множестве» и «ошибка обучения». — Прим. пер.

<sup>5</sup> Синонимами являются также термины «ошибка на проверочной выборке», «ошибка на тестовой выборке» и «ошибка на проверочном множестве». — Прим. пер.

располагая большим количеством наблюдений, мы могли бы вычислить

$$\text{Ave}(y_0 - \hat{f}(x_0))^2, \quad (2.6)$$

то есть среднеквадратичную ошибку предсказаний для этих проверочных наблюдений  $(x_0, y_0)$ . Мы хотели бы выбрать такую модель, для которой это среднее значение — среднеквадратичная ошибка на контрольной выборке — является как можно меньшим.

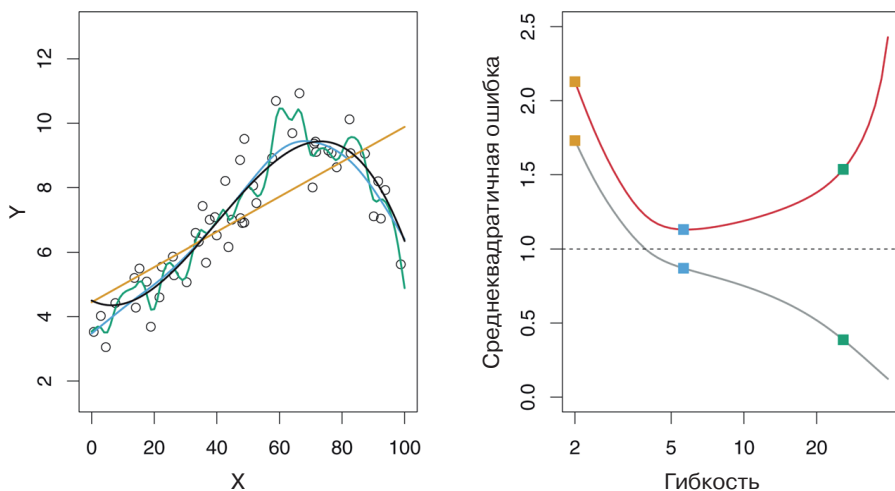
Каким же образом нам выбрать метод, минимизирующий MSE на контрольной выборке? В некоторых случаях у нас в распоряжении может оказаться набор контрольных данных, т. е. мы можем иметь доступ к набору наблюдений, которые не были использованы для обучения статистической модели. Тогда мы можем просто вычислить величину из (2.6) по этим проверочным наблюдениям и выбрать метод статистического обучения с наименьшей контрольной MSE. Но как быть, если проверочные наблюдения недоступны? В этом случае теоретически можно было бы выбрать метод статистического обучения, который минимизирует MSE на обучающей выборке. Это могло бы показаться разумным подходом, поскольку MSE на обучающей и контрольной выборках выглядят тесно связанными между собой. К сожалению, у этой стратегии есть фундаментальная проблема: нет гарантии, что метод с минимальной MSE на обучающих данных также будет иметь минимальную MSE на контрольных данных. Грубо говоря, проблема заключается в том, что многие статистические методы специально оценивают коэффициенты с целью минимизации MSE на обучающей выборке. Ошибка обучения у этих методов может быть довольно низкой, но MSE на контрольных данных часто гораздо выше.

Рисунок 2.9 иллюстрирует этот феномен на простом примере. Слева на этом рисунке приведены наблюдения, которые были сгенерированы нами на основе (2.1), а истинная функция  $f$  показана кривой черного цвета. Оранжевая, голубая и зеленая кривые иллюстрируют три возможные оценки  $\hat{f}$ , полученные при помощи методов с возрастающим уровнем гибкости. Оранжевая линия представляет собой линейную регрессионную модель, которая относительно негибка. Голубая и зеленая кривые были получены при помощи обсуждаемых в главе 7 *сглаживающих сплайнов* с разными уровнями гладкости. Хорошо видно, что по мере увеличения гибкости кривые все ближе подступают к наблюдениям. Зеленая кривая является наиболее гибкой и описывает данные очень хорошо; однако мы видим, что она слабо напоминает истинную функцию  $f$  (показана черным) в силу своей извилистости. Изменяя уровень гибкости сглаживающего сплайна, мы можем подогнать к этим данным много разных моделей.

сглажива-  
ющий  
сплайн

Перейдем теперь к графику, приведенному на рис. 2.9 справа. Серая линия показывает среднее значение MSE на обучающей выборке в зависимости от уровня гибкости, или, более формально, от *числа степеней свободы*, для нескольких сглаживающих сплайнов. Число степеней свободы представляет собой величину, которая обобщает гибкость кривой; более подробно это понятие обсуждается в главе 7. Оранжевые, голубые и зеленые квадраты показывают ошибки обучения, связанные с соответствующими кривыми на рисунке слева. Более ограниченная и, следовательно, более гладкая кривая имеет меньше степеней свободы, чем извилистая кривая (заметьте, что на рис. 2.9 линейная регрессия имеет две степени свободы и является наиболее ограниченной). MSE на обучающей выборке

число  
степеней  
свободы

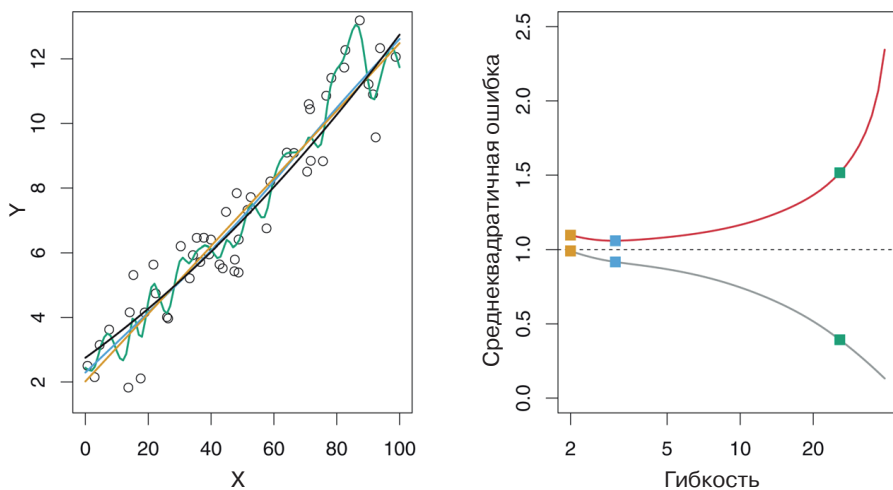


**РИСУНОК 2.9.** Слева: данные, имитированные на основе  $f$ , показаны полными точками черного цвета. Представлены три способа подгонки  $f$ : линейная регрессия (оранжевая линия) и две модели гладких сплайнов (голубая и зеленая линии). Справа: среднеквадратичная ошибка на обучающих (серая линия) и контрольных данных (красная линия), а также минимально возможное значение ошибки для контрольных данных (пунктирная линия). Квадратные символы соответствуют ошибкам на обучающих и контрольных выборках, которые были получены для трех моделей, изображенных на графике слева

монотонно снижается по мере возрастания гибкости. В этом примере истинная функция  $f$  нелинейна, в связи с чем линейная модель (оранжевый цвет) недостаточно гибка для получения удовлетворительной оценки  $f$ . Из всех трех методов зеленая кривая имеет наименьшую MSE на обучающих данных, поскольку она соответствует наиболее гибкой из трех моделей, представленных на графике слева.

В этом примере истинная функция  $f$  нам известна, и поэтому мы также можем вычислить MSE для очень большой контрольной совокупности в зависимости от гибкости модели. (Конечно, как правило,  $f$  неизвестна, и такие вычисления будут невозможны.) MSE на контрольной выборке показана на рис. 2.9 справа в виде красной кривой. Как и в случае с MSE на обучающих данных, MSE на контрольной выборке сначала снижается по мере увеличения гибкости. Однако в какой-то момент MSE на контрольной выборке снова начинает возрастать. Как следствие оранжевая и зеленая кривые имеют самые высокие ошибки на контрольных данных. Голубая кривая имеет минимальную MSE на контрольной выборке, что не должно удивлять, поскольку визуально эта модель выглядит как наилучшая оценка  $f$  (см. рис. 2.9 слева). Горизонтальная прерывистая линия показывает  $\text{Var}(\epsilon)$  — неустранимую ошибку из (2.3), которая соответствует наименьшей достижимой MSE на контрольных данных для всех возможных методов. Следовательно, сглаживающий сплайн, показанный в виде голубой линии, близок к оптимуму.

Как видно на рис. 2.9 (справа), по мере возрастания гибкости метода статистического обучения происходит монотонное снижение MSE на обучающих данных и  $U$ -образное изменение MSE на контрольных данных. Это является фундаментальным свойством статистического обучения, которое остается справедливым для любого имеющегося набора данных и для любого применяемого статистического метода. По мере увеличения гибкости модели MSE на обучающей выборке будет снижаться, но MSE на контрольной выборке необязательно будет вести себя тем же образом. Ситуацию, когда некоторый метод обеспечивает небольшую MSE на обучающих данных и высокую MSE на проверочных данных, называют *переобучением* модели. Это происходит потому, что наша процедура статистического обучения слишком усердно пытается найти закономерности в обучающих данных и в результате может обнаружить некоторые закономерности, которые просто случайны и никак не связаны с истинными свойствами неизвестной функции  $f$ . При переобучении модели MSE на контрольной выборке будет большой потому, что предполагаемые закономерности, найденные нашим методом в обучающих данных, в проверочных данных просто не существуют. Заметьте, что вне зависимости от того, случилось ли переобучение, мы почти всегда ожидаем, что MSE на обучающих данных будет ниже, чем MSE на контрольных данных, поскольку большинство методов статистического обучения так или иначе пытаются минимизировать ошибку обучения. К переобучению относят также частный случай, когда менее гибкая модель обеспечивает меньшую MSE на контрольных данных.



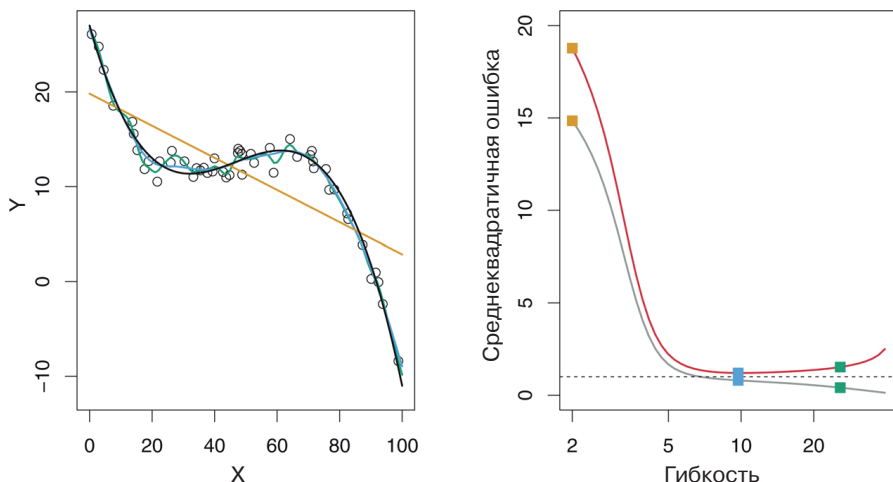
**РИСУНОК 2.10.** То же, что на рис. 2.9, но с истинной функцией  $f$ , которая намного ближе к линейной. В этой ситуации линейная регрессия очень хорошо описывает данные

На рис. 2.10 приведен пример, в котором истинная функция  $f$  приблизительно линейна. Как и раньше, мы наблюдаем монотонное снижение MSE на обучающих данных по мере возрастания гибкости модели, а кривая MSE на контрольных данных имеет  $U$ -образную форму. Одна-

ко в связи с тем, что истинная зависимость близка к линейной, MSE на контрольных данных лишь незначительно снижается перед последующим возрастанием, и поэтому оранжевая линия, подогаданная по методу наименьших квадратов, подходит для описания данных гораздо лучше, чем очень гибкая зеленая кривая. Наконец, рис. 2.11 иллюстрирует пример, в котором функция  $f$  в значительной мере нелинейна. Кривые MSE для обучающих и контрольных выборок по-прежнему демонстрируют те же общие закономерности, но теперь имеет место быстрое снижение обеих этих кривых, перед тем как MSE на контрольных данных начинает медленно возрастать.

перекрестная  
проверка

На практике вычислить MSE по обучающим данным относительно легко, однако оценить MSE на контрольных данных гораздо труднее, поскольку они обычно отсутствуют. Как показывают предыдущие три примера, уровень гибкости, соответствующий модели с наименьшей контрольной MSE, может значительно варьировать в зависимости от свойств данных. В этой книге мы обсуждаем целый ряд подходов, которые можно использовать на практике получения оценки этого минимального значения. Одним из важных методов, предназначенных для оценивания MSE на контрольных данных, является *перекрестная проверка*<sup>6</sup> (глава 5).



**РИСУНОК 2.11.** То же, что на рис. 2.9, но с истинной функцией  $f$ , существенно отличной от линейной. В этой ситуации линейная регрессия описывает данные очень плохо

### 2.2.2 Компромисс между смещением и дисперсией

Оказывается, что U-образная форма кривых, описывающих MSE на контрольных данных (рис. 2.9–2.11), является результатом двух конкурирующих свойств методов статистического обучения. Хотя математическое доказательство выходит за рамки этой книги, можно показать, что для

<sup>6</sup> Используются также термины «кросс-проверка», «кросс-валидация» и «скользящий контроль». — Прим. пер.

некоторого контрольного значения  $x_0$  ожидаемую MSE всегда можно разложить на сумму трех фундаментальных величин: *дисперсии*  $\hat{f}(x_0)$ , квадрата *смещения*  $\hat{f}(x_0)$  и дисперсии остатков  $\epsilon$ . Другими словами<sup>7</sup>:

дисперсия  
смещение

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon). \quad (2.7)$$

Здесь  $E\left(y_0 - \hat{f}(x_0)\right)^2$  обозначает *математическое ожидание* ошибки на контрольной выборке и представляет собой среднее значение MSE, которое мы получили бы при многократном повторном оценивании  $f$  на основе большого числа обучающих выборок и вычислении ошибки для каждого контрольного значения  $x_0$ . Общую ожидаемую MSE на контрольной выборке можно вычислить путем усреднения  $E\left(y_0 - \hat{f}(x_0)\right)^2$  для всех возможных проверочных значений  $x_0$ .

математи-  
ческое  
ожидание  
ошибки

Уравнение (2.7) говорит нам о том, что для минимизации ожидаемой ошибки на проверочных данных мы должны выбрать такой метод статистического обучения, который одновременно обеспечивает *низкую дисперсию* и *низкое смещение*. Заметьте, что дисперсия по определению является положительной величиной, равно как и квадрат смещения. В итоге мы видим, что математическое ожидание MSE на контрольной выборке никогда не может быть ниже неустранимой ошибки  $\text{Var}(\epsilon)$  из уравнения (2.3).

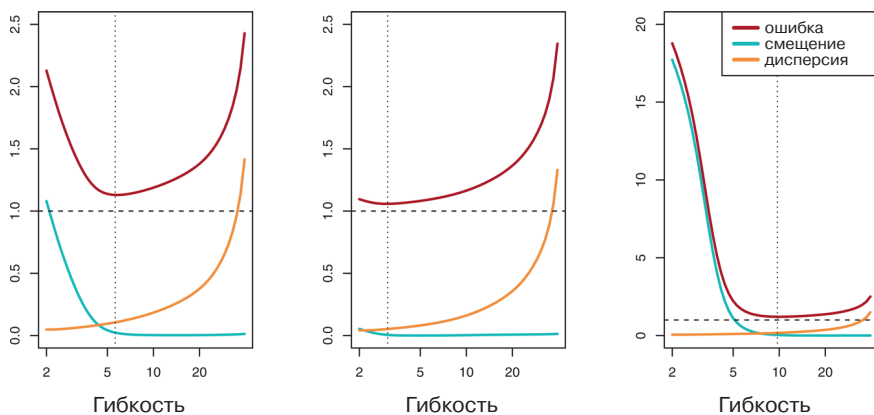
Что мы понимаем под *дисперсией* и *смещением* метода статистического обучения? *Дисперсия* означает величину, на которую  $\hat{f}$  изменилась бы при оценивании этой функции с использованием другой обучающей выборки. Поскольку для подгонки модели используются обучающие данные, то разные обучающие выборки будут приводить к разным  $\hat{f}$ . В идеале оценки  $f$ , полученные на разных выборках, должны варьировать незначительно. Однако если некоторый метод обладает высокой дисперсией, то небольшие изменения в обучающих данных могут привести к большим изменениям в  $\hat{f}$ . В целом более гибкие статистические методы имеют более высокую дисперсию. Рассмотрим зеленую и оранжевую кривые на рис. 2.9. Гибкая зеленая кривая очень близко следует за отдельными наблюдениями. Она обладает высокой дисперсией, поскольку изменение любого наблюдения может вызвать значительное изменение оценки  $\hat{f}$ . В то же время подогнанная по методу наименьших квадратов оранжевая кривая является относительно негибкой и имеет низкую дисперсию, поскольку сдвиг любого наблюдения приведет, скорее всего, лишь к небольшому сдвигу положения этой кривой.

В свою очередь, *смещение* означает ошибку, вводимую за счет аппроксимирования проблемы из реального мира, которая может оказаться чрезвычайно сложной, при помощи гораздо более простого метода. Например, линейная регрессия предполагает, что между  $Y$  и  $X_1, X_2, \dots, X_p$  имеется линейная зависимость. Маловероятно, что какая-либо проблема из реального мира действительно описывается такой простой зависимостью, в связи с чем применение линейной регрессии несомненно приведет к определенному смещению оценки  $f$ . На рис. 2.11 истинная функция  $f$  явно нелинейна, и поэтому не важно, как много наблюдений у нас есть

<sup>7</sup> В соответствии с англоязычными терминами  $\text{Var}$  и  $\text{Bias}$  в приведенной формуле обозначают дисперсию и смещение соответственно. — *Прим. пер.*

в распоряжении — верную оценку при помощи линейной регрессии получить здесь будет невозможно. Другими словами, линейная регрессия в этом примере вызывает большое смещение. Однако на рис. 2.10 истинная функция  $f$  очень близка к линейной, и поэтому при наличии достаточного объема данных линейная регрессия сможет дать верную оценку. В целом более гибкие методы вызывают меньшее смещение.

Как правило, при использовании более гибких методов дисперсия будет возрастать, а смещение — снижаться. Относительная скорость изменения этих двух величин определяет направление изменения MSE на контрольной выборке. По мере увеличения гибкости некоторого класса методов снижение смещения вначале обычно происходит быстрее, чем рост дисперсии. В результате этого математическое ожидание MSE на контрольной выборке снижается. Однако в определенный момент возрастающая гибкость перестает оказывать влияние на смещение, но при этом вызывает рост дисперсии. Когда это происходит, MSE на контрольной выборке возрастает. Заметьте, что мы уже наблюдали такое начальное снижение ошибки на контрольных данных и последующее ее возрастание (см. графики, представленные справа на рис. 2.9–2.11).



**РИСУНОК 2.12.** Квадрат смещения (голубая кривая), дисперсия (оранжевая кривая),  $\text{Var}(\epsilon)$  (пунктирная линия) и MSE на контрольной выборке (красная кривая) для трех наборов данных, показанных на рис. 2.9–2.11. Вертикальная пунктирная линия показывает уровень гибкости, соответствующий наименьшей MSE

Три графика на рис. 2.12 иллюстрируют свойства уравнения (2.7) на примере данных, показанных на рис. 2.9–2.11. В каждом случае сплошная синяя линия показывает квадрат смещения для разных уровней гибкости, а оранжевая кривая соответствует дисперсии. Горизонтальная прерывистая линия отражает  $\text{Var}(\epsilon)$  — неустранимую ошибку. Наконец, красная кривая, соответствующая MSE на контрольной выборке, является суммой этих трех величин. Во всех трех случаях по мере увеличения гибкости метода дисперсия возрастает, а смещение снижается. Однако уровень гибкости, соответствующий оптимальной MSE на контрольной выборке, значительно различается между этими тремя наборами данных, поскольку



квадрат смещения и дисперсия в каждом случае изменяются с разными скоростями. На рис. 2.12 слева уровень смещения поначалу быстро падает, вызывая резкое начальное снижение математического ожидания MSE на контрольной выборке. С другой стороны, на центральном графике рис. 2.12 истинная функция  $f$  близка к линейной, в связи с чем имеет место лишь небольшое снижение уровня смещения по мере увеличения гибкости, а MSE на контрольной выборке незначительно снижается перед началом быстрого роста по мере увеличения дисперсии. Наконец, на рисунке 2.12 справа с увеличением гибкости наблюдается значительное падение уровня смещения, поскольку истинная функция  $f$  существенно нелинейна. Имеется также очень небольшое возрастание дисперсии по мере роста гибкости. В результате MSE на контрольной выборке значительно падает, перед тем, как снова начать медленно возрастать по мере увеличения гибкости.

Связь между смещением, дисперсией и MSE на контрольной выборке, описанная в уравнении 2.7 и показанная на рис. 2.12, известна как *компромисс между смещением и дисперсией*<sup>8</sup>. Хороший результат, достигаемый методом статистического обучения на контрольных данных, требует как низкой дисперсии, так и низкого квадрата смещения. Это называют компромиссом, поскольку можно легко получить метод с чрезвычайно низким смещением, но высокой дисперсией (например, нарисовав кривую, которая проходит через каждую точку обучающей выборки), или метод с очень низкой дисперсией, но высоким смещением (путем подгонки горизонтальной линии к данным). Трудность заключается в нахождении метода, у которого малы как дисперсия, так и квадрат смещения. Данный компромисс является одной из наиболее важных тем этой книги.

компро-  
мисс

В практической ситуации, когда функция  $f$  неизвестна, выполнить непосредственный расчет MSE на контрольных данных, уровня смещения и дисперсии для некоторого метода статистического обучения обычно невозможно. Тем не менее следует всегда помнить о компромиссе между смещением и дисперсией. В этой книге мы рассматриваем методы, которые являются чрезвычайно гибкими и поэтому могут фактически полностью устранить смещение. Однако это не гарантирует того, что они будут работать лучше более простого метода, такого как линейная регрессия. В качестве экстремального примера предположим, что истинная функция  $f$  является линейной. В этой ситуации линейная регрессия будет обладать нулевым смещением, в связи с чем более сложному методу будет сложно конкурировать. Однако если истинная функция  $f$  в значительной мере нелинейна и у нас есть много обучающих наблюдений, тогда, возможно, нам стоит воспользоваться более гибким методом, как показано на рис. 2.11. В главе 5 мы обсуждаем перекрестную проверку, которая представляет собой способ оценки контрольной MSE с использованием обучающих данных.

### 2.2.3 Задачи классификации

До сих пор при обсуждении качества моделей мы фокусировались на задачах регрессии. Однако многие из встретившихся нам концепций, таких как компромисс между смещением и дисперсией, переносятся на задачи классификации лишь с небольшими модификациями, обусловленны-

<sup>8</sup> В оригинале используется термин «bias–variance trade–off». — Прим. пер.

ми тем, что  $y_i$  больше не являются количественными значениями. Предположим, что мы пытаемся оценить  $f$  на основе обучающих наблюдений  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , где  $y_1, \dots, y_n$  теперь представляют собой значения качественной переменной. Наиболее распространенный способ количественного описания точности нашей оценки  $\hat{f}$  заключается в расчете частоты ошибок на обучающей выборке, т. е. доли ошибок, допущенных при применении нашей оцененной функции  $\hat{f}$  к обучающим наблюдениям:

частота  
ошибок

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i). \quad (2.8)$$

индикаторная  
переменная

Здесь  $\hat{y}_i$  представляет собой метку класса, предсказанную для  $i$ -го наблюдения при помощи  $\hat{f}$ .  $I(y_i \neq \hat{y}_i)$  — это индикаторная переменная<sup>9</sup>, равная 1 при  $y_i \neq \hat{y}_i$  и 0 при  $y_i = \hat{y}_i$ . Если  $I(y_i \neq \hat{y}_i) = 0$ , то  $i$ -е наблюдение было предсказано нашим методом классификации правильно; иначе оно классифицировано неверно. Следовательно, уравнение (2.8) позволяет вычислить долю неправильно классифицированных случаев.

ошибка  
обучения

Величину из уравнения 2.8 называют частотой ошибок на обучающей выборке, поскольку она рассчитывается на основе данных, используемых для обучения нашего классификатора. Как и в задачах регрессии, нам более всего интересны частоты ошибок, получаемые в результате применения нашего классификатора к контрольным наблюдениям, которые не использовались в ходе обучения модели. Частота ошибок на контрольной выборке, связанная с набором проверочных наблюдений вида  $(x_0, y_0)$ , вычисляется как<sup>10</sup>

ошибка на  
контрольной  
выборке

$$\text{Ave}(I(y_0 \neq \hat{y}_0)), \quad (2.9)$$

где  $\hat{y}_0$  представляет собой метку класса, полученную в результате применения классификатора к проверочному наблюдению с вектором предикторов  $x_0$ . Хорошим является тот классификатор, у которого ошибка (2.9) минимальна.

## Байесовский классификатор

Можно показать (хотя математическое доказательство выходит за рамки данной книги), что средняя частота ошибок на контрольных данных, приведенная в (2.9), минимизируется очень простым классификатором, который присваивает каждому наблюдению наиболее вероятный класс с учетом соответствующих значений предикторов. Другими словами, нам следует отнести проверочное наблюдение с вектором предикторов  $x_0$  к такому классу  $j$ , для которого вероятность

$$\Pr(Y = j | X = x_0) \quad (2.10)$$

условная  
вероятность

максимальна. Заметьте, что (2.10) представляет собой условную вероятность<sup>11</sup>, т. е. вероятность того, что  $Y = j$  при наблюдаемом векторе предикторов  $x_0$ . Этот очень простой классификатор называют байесовским

<sup>9</sup> В оригинале используется термин «indicator variable». В русскоязычных источниках применяется также термин «фиктивная переменная». — Прим. пер.

<sup>10</sup> Ave означает усреднение (от «average»). — Прим. пер.

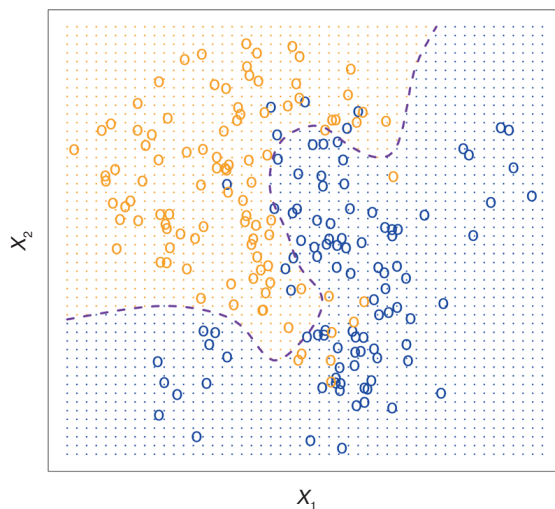
<sup>11</sup> В оригинале используется термин «conditional probability». — Прим. пер.

классификатором. В проблеме с двумя классами, где имеется только два возможных значения отклика, например *класс 1* или *класс 2*, байесовский классификатор предсказывает класс 1, если  $\Pr(Y = 1|X = x_0) > 0.5$ , и класс 2 в остальных случаях.

байесов-  
ский  
классифи-  
катор

На рис. 2.13 приведен пример с использованием имитированного набора данных в двумерном пространстве, образованном предикторами  $X_1$  и  $X_2$ . Оранжевые и синие кружки соответствуют обучающим наблюдениям, которые принадлежат к двум разным классам. Для каждого значения  $X_1$  и  $X_2$  имеется своя вероятность того, что отклик относится к «оранжевому» или «синему классу». Поскольку это имитированные данные, то мы знаем, как они были получены, и мы можем вычислить условные вероятности для каждой пары значений  $X_1$  и  $X_2$ . Закрашенная оранжевым цветом область отражает множество точек, для которых  $\Pr(Y = \text{оранжевый}|X)$  превышает 50%, тогда как область, закрашенная синим цветом, показывает множество точек, для которых эта вероятность ниже 50%. Фиолетовая пунктирная линия соответствует точкам, для которых вероятность в точности равна 50%. Эта линия называется *байесовской решающей границей*<sup>12</sup>. Предсказание байесовского классификатора определяется байесовской решающей границей: наблюдение, попадающее в оранжевую область, будет отнесено к «оранжевому классу», а наблюдение, попадающее в синюю область, — к «синему классу».

байесов-  
ская  
решающая  
граница



**РИСУНОК 2.13.** Набор имитированных данных, состоящий из 100 наблюдений в каждой группе (обозначены синим и оранжевым цветами). Фиолетовая пунктирная линия соответствует байесовской решающей границе. Фоновая сетка оранжевого цвета обозначает область, в которой наблюдение из контрольной выборки будет отнесено к «оранжевому» классу, а сетка синего цвета соответствует области, в которой контрольное наблюдение будет отнесено к «синему» классу

<sup>12</sup> В оригинале используется термин «Bayes decision boundary». — Прим. пер.

байесов-  
ская  
частота  
ошибок

Байесовский классификатор обеспечивает наименьшую возможную частоту ошибок на контрольной выборке, называемую *байесовской частотой ошибок*. Поскольку байесовский классификатор всегда выберет класс, для которого величина (2.10) максимальна, частота ошибок при  $X = x_0$  составит  $1 - \max_j \Pr(Y = j|X = x_0)$ . В целом общая байесовская частота ошибок вычисляется как

$$1 - E \left( \max_j \Pr(Y = j|X) \right), \quad (2.11)$$

где математическое ожидание есть средняя вероятность, рассчитанная по всем возможным значениям  $X$ . Для наших имитированных данных байесовская частота ошибок составляет 0.1304. Это больше 0, поскольку классы в генеральной совокупности перекрываются и  $\Pr(Y = j|X = x_0) < 1$  для некоторых значений  $x_0$ . Байесовская частота ошибок аналогична обсуждавшейся ранее неустраняемой ошибке.

### Метод $K$ ближайших соседей

Теоретически для предсказания качественного отклика всегда желательно было бы использовать байесовский классификатор. Однако для реальных данных мы не знаем условного распределения  $Y$  при заданном  $X$ , и поэтому вычисление байесовского классификатора невозможно. Следовательно, байесовский классификатор служит в качестве недостижимого золотого стандарта, с которым сравниваются другие методы. Многие методы пытаются оценить условное распределение  $Y$  при заданном  $X$  и затем относят то или иное наблюдение к классу с наибольшей *оцененной* вероятностью. Одним из них является *метод  $K$  ближайших соседей* (KNN)<sup>13</sup>. Для некоторого положительного целого числа  $K$  и контрольного наблюдения  $x_0$  классификатор KNN сначала определяет  $K$  наблюдений из обучающей выборки (обозначаются как  $\mathcal{N}_0$ ), которые находятся максимально близко к  $x_0$ . Затем он оценивает условную вероятность для класса  $j$  как долю примеров в  $\mathcal{N}_0$ , у которых значение отклика равно  $j$ :

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j). \quad (2.12)$$

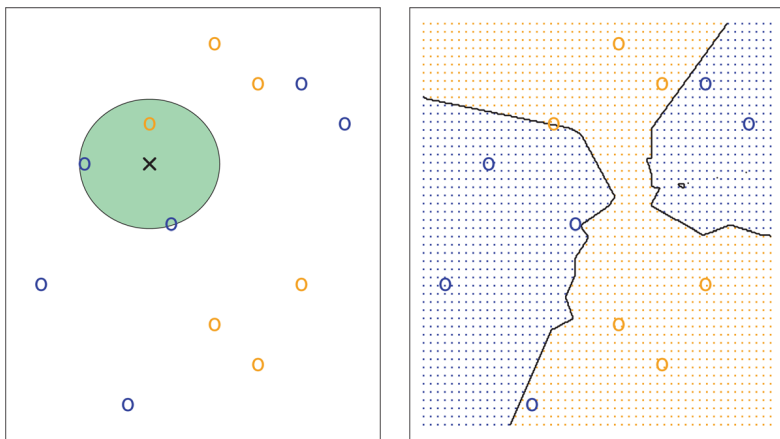
Наконец, KNN применяет теорему Байеса и относит проверочное наблюдение  $x_0$  к классу с наибольшей вероятностью.

На рис. 2.14 приведен пример, поясняющий метод KNN. На графике слева мы изобразили небольшой обучающий набор данных, состоящий из шести голубых и шести оранжевых точек. Наша цель — сделать предсказание для точки, обозначенной черным крестиком. Предположим, что мы выбрали  $K = 3$ . Тогда KNN сначала определит три наблюдения, расположенных к крестику ближе всего. Эта близлежащая область обозначена кругом. Она содержит две синие точки и одну оранжевую, что приводит к вероятности  $2/3$  для «синего класса» и  $1/3$  для «оранжевого класса». Следовательно, KNN предскажет, что черный крестик принадлежит к «синему классу». На рис. 2.14 справа мы применили метод KNN с  $K = 3$

<sup>13</sup> Аббревиатура от «K-nearest neighbors». — Прим. пер.

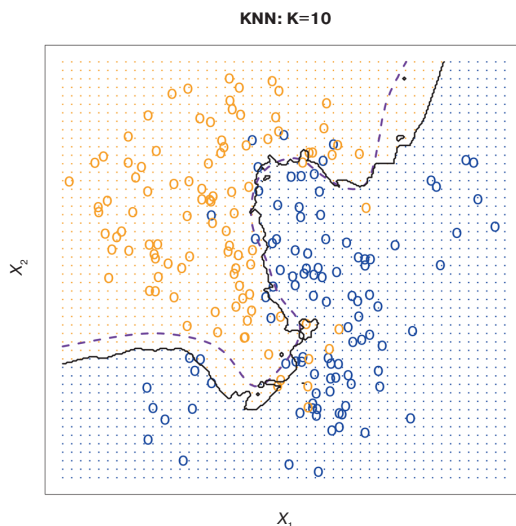
для всех возможных значений  $X_1$  и  $X_2$  и изобразили соответствующую решающую границу.

Несмотря на то что это очень простой подход, KNN часто может приводить к созданию классификаторов, которые удивительно близки к оптимальному байесовскому классификатору. На рис. 2.15 показана решающая граница метода KNN с  $K = 10$ , примененного к большему по размеру набору имитированных данных из рис. 2.13. Обратите внимание: даже несмотря на то, что истинная решающая граница классификатору KNN неизвестна, его решающая граница очень похожа на границу байесовского классификатора. Частота ошибок на контрольной выборке при использовании KNN составляет 0.1363, что близко к байесовской частоте ошибок — 0.1304.



**РИСУНОК 2.14.** Метод KNN с  $K = 3$  проиллюстрирован на простом примере с шестью «синими» и шестью «оранжевыми» наблюдениями. Слева: подлежащее классификации наблюдение из контрольной выборки показано черным крестиком. Определены три ближайшие к этому наблюдению точки, в результате чего оно отнесено к наиболее часто встречающемуся среди соседей классу — в данном случае к «синему классу». Справа: решающая граница KNN из этого примера показана линией черного цвета. Синяя сетка соответствует области, в которой наблюдение из контрольной выборки будет отнесено к «синему классу», а сетка оранжевого цвета — области, в которой наблюдение из контрольной выборки будет отнесено к «оранжевому классу»

Выбор  $K$  имеет огромный эффект на итоговый классификатор KNN. На рис. 2.16 приведены две модели KNN, подогнанные к имитированным данным из рис. 2.13 с использованием  $K = 1$  и  $K = 100$ . При  $K = 1$  решающая граница чрезмерно гибка и обнаруживает закономерности в данных, которые не согласуются с байесовской решающей границей. Это соответствует классификатору с низким смещением и очень высокой дисперсией. По мере увеличения  $K$  метод становится менее гибким и формирует решающую границу, похожую на прямую линию. Это соответствует классификатору с низкой дисперсией и высоким смещением. Для этих

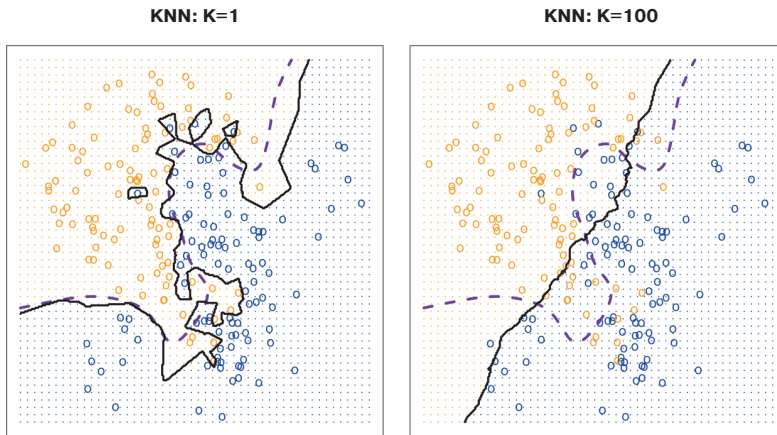


**РИСУНОК 2.15.** Черная линия показывает решающую границу KNN для данных из рис. 2.13, найденную с использованием  $K = 10$ . Байесовская решающая граница представлена в виде фиолетовой пунктирной линии. Эти две решающие границы очень похожи

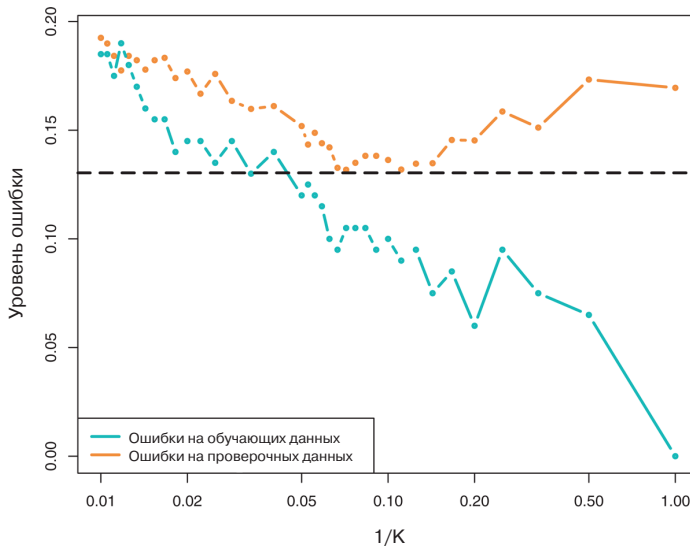
имитированных данных ни  $K = 1$ , ни  $K = 100$  не дают хороших предсказаний: частоты ошибок на контрольных данных составляют 0.1695 и 0.1925 соответственно.

Как и в случае с регрессией, сильной связи между ошибкой на обучающих данных и ошибкой на контрольных данных нет. При  $K = 1$  ошибка обучения KNN равна 0, однако ошибка на контрольных данных может быть довольно высокой. Как правило, при использовании более гибких методов частота ошибки на обучающей выборке будет снижаться, но частота ошибок на контрольной выборке не обязательно будет вести себя тем же образом. На рис. 2.17 мы изобразили ошибки KNN на обучающих и проверочных данных в зависимости от  $1/K$ . При увеличении  $1/K$  этот метод становится более гибким. Подобно регрессионным моделям, при увеличении гибкости частота ошибок на обучающих данных равномерно снижается. Однако ошибка на контрольных данных проявляет характерную  $U$ -образную форму, поначалу снижаясь (с минимумом примерно на отметке  $K = 10$ ), а затем снова возрастаая, когда метод становится чрезмерно гибким и приводит к переобучению.

При решении задач как регрессии, так и классификации выбор правильного уровня гибкости является критическим для успеха любого метода статистического обучения. Компромисс между смещением и дисперсией и вытекающая из него  $U$ -образная форма кривой ошибки на контрольных данных могут сделать этот выбор тяжелой задачей. В главе 5 мы вернемся к данной теме и обсудим различные подходы для оценивания частот ошибок на контрольных выборках, а тем самым и нахождения оптимального уровня гибкости для того или иного метода статистического обучения.



**РИСУНОК 2.16.** Сравнение решающих границ KNN (черные сплошные линии), полученных по данным из рис. 2.13 с использованием  $K = 1$  и  $K = 100$ . При  $K = 1$  решающая граница получается чрезмерно гибкой, тогда как при  $K = 100$  она недостаточно гибка. Байесовская решающая граница показана в виде фиолетовой пунктирной линии



**РИСУНОК 2.17.** Ошибки KNN-классификатора на обучающей (синяя кривая, 200 наблюдений) и контрольной выборках (оранжевая линия, 5000 наблюдений) для данных из рис. 2.13, показанные в зависимости от возрастающей гибкости ( $1/K$ ) или, что эквивалентно, в зависимости от снижающегося числа соседей  $K$ . Черная пунктирная линия показывает байесовскую частоту ошибок. Извилистость кривых обусловлена малым размером обучающей выборки



## 2.3 Лабораторная работа: введение в R

В этой лабораторной работе мы представим некоторые простые команды R. Лучший способ изучения нового языка заключается в экспериментировании с его командами. R можно загрузить с сайта

<http://cran.r-project.org/>

### 2.3.1 Основные команды

функция Для выполнения тех или иных операций R использует *функции*. Для запуска функции с именем `funcname` мы набираем `funcname(input1, input2)`, где входные параметры, или *аргументы*, `input1` и `input2` сообщают R, как именно следует исполнить эту функцию. Например, для создания вектора с несколькими числами мы используем функцию *конкатенации* `c()`<sup>14</sup>. Следующая команда говорит R объединить числа 1, 3, 2 и 5 и сохранить их в виде вектора с именем `x`. Когда мы наберем `x`, то в ответ получим этот вектор.

```
> x <- c(1, 3, 2, 5)
> x
[1] 1 3 2 5
```

Обратите внимание на то, что `>` не является частью команды — R выводит этот знак просто, чтобы показать свою готовность к выполнению следующей команды. Мы можем сохранять объекты не только с помощью `<-`, но также и `=`:

```
> x = c(1, 6, 2)
> x
[1] 1 6 2
> y = c(1, 4, 3)
```

Многократное нажатие клавиши со стрелкой *вверх* приведет к показу предыдущих команд, которые можно отредактировать. Это полезно, поскольку необходимость повторения похожих команд возникает часто. Кроме того, ввод команды `?funcname` всегда откроет новое окно со справочным файлом, содержащим дополнительную информацию по функции `funcname`.

Мы можем попросить R выполнить сложение двух чисел. В этом случае программа сначала добавит первое число из `x` к первому числу из `y` и т. д. Однако `x` и `y` должны быть одинаковой длины. Мы можем проверить их длину при помощи функции `length()`.

```
> length(x)
[1] 3
> length(y)
[1] 3
> x + y
[1] 2 10 5
```

<sup>14</sup> «Concatenate» значит «соединять», «объединять». — Прим. пер.