

УДК 519.25/.6:004.434Python
ББК 22.17с5
Д40

Джеймс Г., Уиттен Д., Хасты Т., Тибширани Р., Тейлор Дж.
Д40 Введение в статистическое обучение с примерами на языке Python / пер.
с англ. А. Ю. Гинько. – М.: ДМК Пресс, 2024. – 846 с.: ил.

ISBN 978-5-93700-217-4

В этой книге доступным языком описывается все разнообразие форм статистического обучения. Рассматриваются линейная регрессия, классификация, методы повторной выборки, отбор и регуляризация, полиномиальная регрессия, сплайны, локальная регрессия, обобщенные аддитивные модели, деревья решений, метод опорных векторов, кластеризация, а также нейронные сети, анализ выживаемости и множественная проверка гипотез. Теоретическая часть дополнена примерами из реальной практики и разборами решений на языке Python.

Издание предназначено не только для опытных специалистов в области статистики, но и для тех, кто желает попробовать применить продвинутые техники статистического обучения при анализе своих данных.

УДК 519.25/.6:004.434Python
ББК 22.17с5

First published in English under the title.

An Introduction to Statistical Learning; with Applications in Python by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor, edition: 1.

This edition has been translated and published under licence from Springer Nature Switzerland AG.

Springer Nature Switzerland AG takes no responsibility and shall not be made liable for the accuracy of the translation.

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.

ISBN 978-3-031-38746-3 (англ.)

ISBN 978-5-93700-217-4 (рус.)

© Springer Nature Switzerland AG
2023

© Перевод, оформление, издание,
ДМК Пресс, 2023

Содержание

От издательства	14
Предисловие	15
О переводчике	18
Глава 1. Введение	19
Общий обзор статистического обучения	19
Данные о зарплатах.....	19
Данные по рынку акций	21
Данные об экспрессии генов.....	22
Краткая история статистического обучения	24
О книге.....	25
Для кого предназначена эта книга?	28
Обозначения и матричная алгебра по-простому	29
Структура книги	32
Используемые в лабораторных работах и упражнениях наборы данных	33
Сайт книги.....	35
Источники	35
Глава 2. Статистическое обучение	36
2.1 Что такое статистическое обучение?	36
2.1.1 Зачем нужно оценивать f ?	38
2.1.2 Как оценивать f ?.....	42
2.1.3 Компромисс между точностью предсказаний и интерпретируемостью модели	46
2.1.4 Обучение с учителем и без учителя.....	49
2.1.5 Регрессия против классификации	51
2.2 Оценка точности модели	52
2.2.1 Оценка качества подгонки.....	52
2.2.2 Компромисс между смещением и дисперсией.....	58
2.2.3 Задачи классификации	61
2.3 Лабораторная работа: введение в Python.....	68
2.3.1 Подготовка.....	68
2.3.2 Основные команды	69
2.3.3 Введение в числовой Python.....	70
2.3.4 Графика	80
2.3.5 Последовательности и срезы	87
2.3.6 Индексирование данных.....	89

2.3.7	Загрузка данных	93
2.3.8	Циклы for	100
2.3.9	Дополнение про графики и числа.....	102
2.4	Упражнения.....	109
	Теоретические	109
	Практические.....	111
Глава 3. Линейная регрессия		115
3.1	Простая линейная регрессия	116
3.1.1	Оценка коэффициентов	117
3.1.2	Определение точности оценки коэффициентов	120
3.1.3	Определение точности оценки модели	126
3.2	Множественная линейная регрессия	129
3.2.1	Оценка регрессионных коэффициентов	130
3.2.2	Важные вопросы.....	133
3.3	Прочие факторы регрессионного моделирования.....	142
3.3.1	Качественные предикторы	142
3.3.2	Расширения линейной модели	147
3.3.3	Возможные проблемы	154
3.4	Маркетинговый план.....	166
3.5	Сравнение линейной регрессии и классификатора k-ближайших соседей	168
3.6	Лабораторная работа: линейная регрессия.....	174
3.6.1	Импорт библиотек.....	174
3.6.2	Простая линейная регрессия	176
3.6.3	Множественная линейная регрессия	185
3.6.4	Прелести многомерной подгонки	186
3.6.5	Эффекты взаимодействия.....	188
3.6.6	Нелинейные преобразования предикторов.....	188
3.6.7	Качественные предикторы	190
3.7	Упражнения.....	192
	Теоретические	192
	Практические.....	194
Глава 4. Классификация		201
4.1	Введение в классификацию	202
4.2	Почему не линейная регрессия?	203
4.3	Логистическая регрессия	205
4.3.1	Логистическая модель	206
4.3.2	Оценивание регрессионных коэффициентов.....	208
4.3.3	Предсказание	210
4.3.4	Множественная логистическая регрессия.....	211
4.3.5	Мультиномиальная логистическая регрессия	214
4.4	Обобщенные модели для классификации.....	215
4.4.1	Линейный дискриминантный анализ для $p = 1$	217
4.4	Линейный дискриминантный анализ для $p > 1$	220

4.4.3	Квадратичный дискриминантный анализ	229
4.4.4	Наивный байесовский классификатор.....	231
4.5	Сравнение методов классификации	236
4.5.1	Аналитическое сравнение	236
4.5.2	Практическое сравнение.....	240
4.6	Обобщенные линейные модели.....	244
4.6.1	Применение линейной регрессии к набору данных Vikeshare.....	244
4.6.2	Пуассоновская регрессия на наборе данных Vikeshare	247
4.6.3	Применимость обобщенных линейных моделей.....	251
4.7	Лабораторная работа: логистическая регрессия, LDA, QDA и KNN	252
4.7.1	Набор данных Smarket	252
4.7.2	Логистическая регрессия	254
4.7.3	Линейный дискриминантный анализ.....	261
4.7.4	Квадратичный дискриминантный анализ	264
4.7.5	Наивный байесовский классификатор.....	266
4.7.6	Классификатор k -ближайших соседей	268
4.7.7	Линейная и пуассоновская регрессия с набором данных Vikeshare.....	276
4.8	Упражнения.....	283
	Теоретические	283
	Практические.....	287
Глава 5. Методы повторной выборки		291
5.1	Перекрестная проверка	292
5.1.1	Метод проверочной выборки	292
5.1.2	Перекрестная проверка по отдельным наблюдениям.....	295
5.1.3	k -кратная перекрестная проверка.....	297
5.1.4	Компромисс между смещением и дисперсией применительно к k -кратной перекрестной проверке	300
5.1.5	Перекрестная проверка при решении задач классификации	301
5.2	Бутстреп.....	304
5.3	Лабораторная работа: перекрестная проверка и бутстреп.....	308
5.3.1	Метод проверочной выборки	309
5.3.2	Перекрестная проверка	312
5.3.3	Бутстреп	315
5.4	Упражнения.....	321
	Практические.....	322
Глава 6. Отбор и регуляризация линейных моделей		327
6.1	Отбор подмножества переменных	329
6.1.1	Отбор оптимального подмножества переменных	329
6.1.2	Пошаговый отбор	332
6.1.3	Выбор оптимальной модели.....	336
6.2	Методы сжатия	342

6.2.1	Гребневая регрессия.....	342
6.2.2	Лассо	347
6.2.3	Выбор гиперпараметра.....	357
6.3	Методы снижения размерности	359
6.3.1	Метод главных компонент	360
6.3.2	Метод частных наименьших квадратов	368
6.4	Размышляя о большой размерности.....	370
6.4.1	Данные большой размерности.....	370
6.4.2	Что не так с данными большой размерности?.....	371
6.4.3	Регрессия в условиях большой размерности	374
6.4.4	Интерпретация результатов в задачах большой размерности	375
6.5	Лабораторная работа: линейные модели и методы регуляризации	377
6.5.1	Методы отбора подмножеств переменных	378
6.5.2	Гребневая регрессия и лассо	387
6.5.3	Регрессия PCR и PLS	400
6.6	Упражнения.....	404
	Теоретические	404
	Практические.....	408
Глава 7. Выходим за рамки линейности		411
7.1	Полиномиальная регрессия.....	412
7.2	Ступенчатые функции.....	414
7.3	Базисные функции.....	417
7.4	Регрессионные сплайны.....	417
7.4.1	Кусочно-полиномиальная регрессия	417
7.4.2	Ограничения и сплайны.....	418
7.4.3	Представление сплайнов с помощью базисных функций	420
7.4.4	Выбор количества и расположения узлов.....	422
7.4.5	Сравнение с полиномиальной регрессией.....	424
7.5	Сглаживающие сплайны	425
7.5.1	Введение в сглаживающие сплайны.....	425
7.5.2	Выбор сглаживающего параметра λ	427
7.6	Локальная регрессия.....	429
7.7	Обобщенные аддитивные модели.....	432
7.7.1	GAM для регрессионных задач.....	432
7.7.2	GAM для задач классификации	436
7.8	Лабораторная работа: нелинейные модели	438
7.8.1	Полиномиальная регрессия и ступенчатые функции	438
7.8.2	Сплайны.....	446
7.8.3	Сглаживающие сплайны и GAM.....	450
7.8.4	Локальная регрессия	466
7.9	Упражнения.....	467
	Теоретические	467
	Практические.....	469

Глава 8. Методы на основе деревьев решений	473
8.1 Основы деревьев решений	473
8.1.1 Регрессионные деревья	474
8.1.2 Деревья классификации	482
8.1.3 Деревья против линейных моделей	485
8.1.4 Преимущества и недостатки деревьев.....	487
8.2 Бэггинг, случайные леса, бустинг и байесовские аддитивные регрессионные деревья	487
8.2.1 Бэггинг	488
8.2.2 Случайные леса.....	492
8.2.3 Бустинг	494
8.2.4 Байесовские аддитивные регрессионные деревья.....	497
8.2.5 Краткий вывод по ансамблевым методам, основанным на деревьях	501
8.3 Лабораторная работа: методы на основе деревьев.....	502
8.3.1 Построение деревьев классификации.....	502
8.3.2 Построение регрессионных деревьев	509
8.3.3 Бэггинг и случайный лес	511
8.3.4 Бустинг	514
8.3.5 Байесовские аддитивные регрессионные деревья.....	516
8.4 Упражнения.....	517
Теоретические	517
Практические	519
Глава 9. Метод опорных векторов	522
9.1 Классификатор с максимальным зазором	523
9.1.1 Что такое гиперплоскость?	523
9.1.2 Классификация с использованием разделяющей гиперплоскости.....	524
9.1.3 Классификатор с максимальным зазором	526
9.1.4 Построение классификатора с максимальным зазором.....	528
9.1.5 Случай с несуществующей разделяющей гиперплоскостью....	529
9.2 Классификаторы на опорных векторах	530
9.2.1 Введение в классификаторы на опорных векторах	530
9.2.2 Детали работы классификатора на опорных векторах.....	532
9.3 Метод опорных векторов	535
9.3.1 Классификация с использованием нелинейных решающих границ	535
9.3.2 Метод опорных векторов	537
9.3.3 Применение к данным о сердечных заболеваниях.....	541
9.4 SVM для случаев с несколькими классами	543
9.4.1 Классификация «один против одного»	543
9.4.2 Классификация «один против всех»	543
9.5 Связь с логистической регрессией.....	544
9.6 Лабораторная работа: метод опорных векторов.....	547
9.6.1 Классификатор на опорных векторах	547

9.6.2	Метод опорных векторов	555
9.6.3	ROC-кривые	560
9.6.4	SVM с несколькими классами	563
9.6.5	Применение на примере данных об экспрессии генов	565
9.7	Упражнения	567
	Теоретические	567
	Практические	568
Глава 10. Глубокое обучение	572	
10.1	Однослойные нейронные сети	573
10.2	Многослойные нейронные сети	576
10.3	Сверточные нейронные сети	581
	10.3.1 Сверточные слои	583
	10.3.2 Пулинговые слои	586
	10.3.3 Архитектура сверточной нейронной сети	586
	10.3.4 Аугментация данных	588
	10.3.5 Результаты использования обученного классификатора	589
10.4	Классификация документов	590
10.5	Рекуррентные нейронные сети	594
	10.5.1 Последовательные модели для классификации документов ...	597
	10.5.2 Прогнозирование временных рядов	600
	10.5.3 Резюме по рекуррентным нейронным сетям	604
10.6	Когда нужно использовать глубокое обучение	605
10.7	Обучение нейронных сетей	608
	10.7.1 Обратное распространение	610
	10.7.2 Регуляризация и стохастический градиентный спуск	611
	10.7.3 Метод прореживания	612
	10.7.4 Настройка нейронной сети	614
10.8	Интерполяция и двойной спуск	614
10.9	Лабораторная работа: глубокое обучение	619
	10.9.1 Однослойная нейронная сеть на наборе данных Hitters	622
	10.9.2 Многослойная нейронная сеть на наборе данных MNIST	632
	10.9.3 Сверточные нейронные сети	638
	10.9.4 Использование предварительно обученных сверточных моделей	644
	10.9.5 Классификация документов IMDB	647
	10.9.6 Рекуррентные нейронные сети	653
10.10	Упражнения	663
	Теоретические	663
	Практические	664
Глава 11. Анализ выживаемости и цензурированные данные	666	
11.1	Время выживаемости и цензурированное время	667
11.2	Понятие цензурирования	668
11.3	Кривая выживаемости по методу Каплана–Мейера	669
11.4	Логарифмический ранговый тест	672

11.5	Регрессионные модели с откликом о выживаемости.....	675
11.5.1	Функция риска.....	675
11.5.2	Пропорциональные риски.....	678
11.5.3	Пример: набор данных BrainCancer.....	681
11.5.4	Пример: набор данных Publication.....	682
11.6	Сжатие модели пропорциональных рисков Кокса.....	685
11.7	Дополнительные темы.....	687
11.7.1	Значение площади под кривой для анализа выживаемости....	687
11.7.2	Выбор временной шкалы.....	688
11.7.3	Предикторы, зависящие от времени.....	689
11.7.4	Проверка предположения о пропорциональных рисках.....	690
11.7.5	Деревья выживаемости.....	690
11.8	Лабораторная работа: анализ выживаемости.....	690
11.8.1	Набор данных BrainCancer.....	691
11.8.2	Набор данных Publication.....	698
11.8.3	Данные кол-центра.....	700
11.9	Упражнения.....	707
	Теоретические.....	707
	Практические.....	710
Глава 12. Методы обучения без учителя.....		712
12.1	Сложности, связанные с обучением без учителя.....	712
12.2	Анализ главных компонент.....	713
12.2.1	Что такое главные компоненты?.....	714
12.2.2	Другая интерпретация главных компонент.....	719
12.2.3	Доля объясненной дисперсии.....	721
12.2.4	Подробности анализа главных компонент.....	723
12.2.5	Другое применение главных компонент.....	726
12.3	Пропущенные значения и заполнение матрицы.....	726
12.4	Методы кластеризации.....	732
12.4.1	Кластеризация по методу k-средних.....	734
12.4.2	Иерархическая кластеризация.....	738
12.4.3	Практические сложности при применении кластеризации.....	748
12.5	Лабораторная работа: обучение без учителя.....	750
12.5.1	Анализ главных компонент.....	751
12.5.2	Заполнение матрицы.....	757
12.5.3	Кластеризация.....	761
12.5.4	Пример с набором данных NCI60.....	771
12.6	Упражнения.....	779
	Теоретические.....	779
	Практические.....	781
Глава 13. Множественная проверка гипотез.....		785
13.1	Краткий обзор проверки гипотез.....	786
13.1.1	Проверка гипотезы.....	787
13.1.2	Ошибки I и II рода.....	791

13.2	Трудности множественной проверки гипотез.....	793
13.3	Групповая вероятность ошибки.....	795
13.3.1	Что такое групповая вероятность ошибки	795
13.3.2	Способы контроля групповой вероятности ошибки.....	797
13.3.3	Компромисс между групповой вероятностью ошибки и мощностью.....	804
13.4	Ожидаемая доля ложных отклонений гипотез.....	805
13.4.1	Представление ожидаемой доли ложных отклонений гипотез	805
13.4.2	Метод Бенджамини–Хохберга.....	807
13.5	Метод повторной выборки применительно к p -значениям и ожидаемой доле ложных отклонений гипотез.....	810
13.5.1	Метод повторной выборки для p -значений.....	811
13.5.2	Метод повторной выборки для ожидаемой доли ложных отклонений гипотез	814
13.5.3	Когда бывают полезны методы повторной выборки?	817
13.6	Лабораторная работа: множественная проверка гипотез	818
13.6.1	Обзор проверки гипотез.....	818
13.6.2	Групповая вероятность ошибки	820
13.6.3	Ожидаемая доля ложных отклонений гипотез	824
13.6.4	Метод повторной выборки	827
13.7	Упражнения.....	831
	Теоретические	831
	Практические.....	833

Предисловие

Статистическое обучение подразумевает использование набора инструментов, позволяющих извлечь ценные сведения из сложно организованных данных. В последние годы мы стали свидетелями невероятного роста масштаба и охвата собираемых данных практически во всех областях науки и промышленности. В результате использование инструментов статистического обучения приобрело критически важный смысл для тех, кто хочет понять подноготную данных, а поскольку сегодня работа с данными охватывает все большее количество отраслей, получается, что статистическое обучение теперь нужно едва ли не каждому!

Одна из первых книг, посвященных статистическому обучению, – «Основы статистического обучения» (*The Elements of Statistical Learning* в соавторстве Тренора Хасты, Роберта Тибширани и Джерома Фридмана), – была опубликована в 2001 году, а второе издание увидело свет в 2009-м. Эта книга приобрела широкую популярность не только в области статистики, но также и во многих смежных областях. Одной из причин такой популярности была доступность изложения материала. В то же время для ее чтения было необходимо обладать достаточными математическими знаниями.

Книга «Введение в статистическое обучение с примерами на языке R»¹, выдержавшая два издания – в 2013 и 2021 годах, – была призвана сделать освещение основных аспектов статистического обучения более простым и менее техническим. В дополнение к линейной регрессии в книге описываются многие из наиболее значимых на сегодняшний день подходов в статистике и машинном обучении, включая методы повторной выборки, разреженные методы классификации и регрессии, обобщенные аддитивные модели, методы на основе деревьев решений, метод опорных векторов, глубокое обучение, анализ выживаемости, кластеризацию и множественную проверку гипотез.

С момента своей первой публикации книга «Введение в статистическое обучение с примерами на языке R» закрепились в качестве одного из основных учебных пособий для бакалавров и магистров статистики по всему миру и базового справочника для специалистов в области науки о данных. Ключом к такому успеху стала практическая направ-

¹ Джеймс Г., Уиттон Д., Хасты Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. М.: ДМК Пресс, 2016. URL: <https://dmkpress.com/catalog/computer/statistics/978-5-97060-495-3/>.

ленность книги, в которой, начиная со второй главы, каждая глава сопровождается лабораторной работой на языке R, демонстрирующей реализацию соответствующих методов статистического обучения, что позволяет развить практические навыки у обучающихся.

Однако в последние годы большую популярность в среде науки о данных приобрел язык программирования Python, что привело к росту потребности в альтернативной версии книги с уклоном в этот язык. Так и появилась книга, которую вы держите в руках. В ней было сохранено и дополнено содержание глав, а все практические примеры были переписаны с R на Python, чем мы обязаны нашему новому соавтору Джонатану Тейлору (Jonathan Taylor). В некоторых лабораторных работах мы будем использовать библиотеку `ISLP Python`, специально разработанную для облегчения реализации методов статистического обучения на Python. Эти практические упражнения будут полезны как для новичков в Python, так и для опытных разработчиков.

При написании обеих книг мы главным образом делали упор на реализацию обсуждаемых статистических методов на практике, а не на математических предположках, так что они идеально подойдут для студентов старших курсов бакалавриата и магистратуры, изучающих статистику и другие точные науки, а также для тех, кто хочет воспользоваться инструментами статистического обучения для выявления зависимостей в своих данных. Книгу, которую вы держите в руках, можно использовать в качестве учебного пособия в курсе по статистике, состоящем из двух семестров.

Мы безмерно благодарны читателям, приславшим свои ценные замечания после выхода первого издания книги, и хотели бы перечислить их поименно: Паллави Басу (Pallavi Basu), Александра Чулдохова (Alexandra Chouldechova), Патрик Данахер (Patrick Danaher), Уилл Фитьян (Will Fithian), Луэлла Фу (Luella Fu), Сэм Гросс (Sam Gross), Макс Гразьер G'Sell (Max Grazier G'Sell), Кортни Паулсон (Courtney Paulson), Синхэо Цяо (Xinghao Qiao), Элиза Шенг (Elisa Sheng), Ноа Симон (Noah Simon), Кин Минг Тан (Kean Ming Tan), Синь Лу Тан (Xin Lu Tan). Мы также хотим сказать спасибо тем, кто внес непосредственный вклад в написание второго издания книги: Алан Агрести (Alan Agresti), Иэн Кармайкл (Iain Carmichael), Икун Чен (Yiqun Chen), Эрин Крейг (Erin Craig), Дэйзи Динг (Daisy Ding), Люси Гао (Lucy Gao), Исмаэл Лемхадри (Ismael Lemhadri), Брайан Мартин (Bryan Martin), Анна Ньюфелд (Anna Neufeld), Джеофф Тимс Geoff Tims, Карстен Фолкманн (Carsten Voelkmann), Стив Ядловски (Steve Yadlowsky) и Джеймс Цзоу (James Zou). Кроме того, мы выражаем искреннюю благодарность Баласубраманиану «Нарас» Нарасимхану (Balasubramanian «Naras» Narasimhan) за его помощь в подготовке обеих книг.

Для нас большая честь наблюдать за тем, какое существенное влияние книга «Введение в статистическое обучение с примерами на языке R» оказала на изучение статистических методов на практике – как

в учебных заведениях, так и в сфере самообразования. Мы надеемся, что в нашей новой книге нынешние и будущие специалисты в области прикладной статистики смогут найти подходящие инструменты для анализа данных.

Прогнозы – дело нелегкое, особенно если они касаются будущего.

– Йоги Берра

О переводчике

Александр Гинько, обладающий богатым опытом работы в сфере ИТ и более десяти лет посвятивший переводам книг и статей на самые разные темы, в последние годы специализируется на переводе книг в области бизнес-аналитики и программирования для издательства «ДМК Пресс» по направлениям Python, R, SQL, Power BI, DAX, Excel, Power Query, Tableau... На данный момент в активе Александра уже более 20 книг, включая одну авторскую, и он продолжает плодотворно работать над переводом новых книг.

Помимо перевода книг, Александр ведет свой канал в Telegram (https://t.me/alexanderginko_books), на котором вы можете из первых уст получить ответы на все интересующие вас вопросы об уже переведенных книгах, находящихся в работе и запланированных на будущее. Также на канале можно найти промокоды на все книги Александра для покупки книг на сайте издательства «ДМК Пресс» с большими скидками.

Глава 1

Введение

Общий обзор статистического обучения

Статистическое обучение (statistical learning) представляет собой обширный набор инструментов для лучшего понимания сущности данных. Эти инструменты можно условно разбить на две большие группы: *обучение с учителем* (supervised) и *обучение без учителя* (unsupervised). Первая из них в общем смысле предполагает построение *статистических моделей* (statistical model) для предсказания, или оценивания, некой *выходной переменной* (output) на основе одной или нескольких *входных переменных* (input). С задачами такого рода можно столкнуться в самых разных сферах жизнедеятельности, включая бизнес, медицину, астрофизику и общественную политику. Что касается обучения без учителя, здесь также присутствуют входные переменные, но нет контролируемого выхода. Несмотря на это, мы можем изучить связи и структуру на основе представленных данных. Чтобы вы лучше понимали области применения статистического обучения, давайте рассмотрим три реальных набора данных, с которыми мы будем работать в этой книге.

Данные о зарплатах

В этом примере (к данным которого мы будем обращаться на протяжении книги как к набору *wage*) мы займемся оценкой некоторого количества факторов на предмет их влияния на зарплаты группы мужчин из Атлантического региона США. В частности, нам бы хотелось понять зависимости между возрастом (*age*), уровнем образования (*education*) человека и текущим годом (*year*) с одной стороны и его зарплатой (*wage*) – с другой. Давайте рассмотрим левый график на рис. 1.1, на котором отображена связь между возрастом и зарплатой для всех людей в нашем наборе данных. Исходя из графика, вполне очевидно, что в среднем зарплата увеличивается пропорционально возрасту до определенного момента (около 60 лет), после чего наблюдается неко-

торый спад. Синяя линия, соответствующая оценке средней зарплаты людей в зависимости от возраста, подтверждает нашу догадку. Таким образом, если мы знаем возраст гипотетического сотрудника, мы можем использовать построенную кривую для *предсказания* (predict) его зарплаты. В то же время мы наблюдаем значительную вариативность в оценке зарплаты этим методом, а значит, одной входной переменной в виде возраста человека недостаточно для точного предсказания его зарплаты.

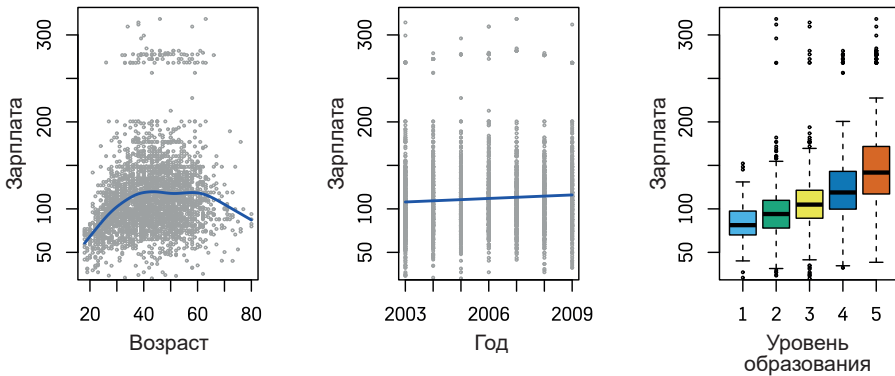


РИС. 1.1 Набор данных Wage, содержащий исследовательскую информацию по мужчинам, проживающим в центральной части Атлантического региона США. Слева: зарплата (wage) как функция от возраста (age). В среднем зарплата растет с увеличением возраста примерно до 60-летней отметки, после чего начинается спад. В середине: зарплата (wage) как функция от года (year). В этом разрезе наблюдается малозначительный, но устойчивый рост зарплаты примерно на 10000 долл. в период с 2003 года по 2009-й. Справа: диаграмма размаха, отражающая зарплату (wage) как функцию от уровня образования (education), где 1 соответствует низшему уровню (аттестат об окончании средней школы), а 5 – высшему (ученая степень). В среднем, как видно по графику, зарплата прямо пропорционально связана с уровнем образования

Также у нас есть информация относительно связи между двумя оставшимися входными переменными – годом (year) и уровнем образования (education) – и зарплатой. По центру и справа на рис. 1.1 как раз показаны графики зависимости зарплаты от этих двух переменных соответственно. Как видно, и год получения зарплаты, и уровень образования человека так или иначе влияют на его достаток. В интервале между 2003 и 2009 годами средняя зарплата исследуемых мужчин в среднем выросла на 10 000 долл., а темпы роста оказались едва заметными на фоне высокой изменчивости данных.

Что касается уровня образования (справа), то здесь также наблюдается прямая зависимость, означающая, что в среднем зарплата напрямую зависит от образования. Очевидно, что наиболее точный прогноз зарплаты можно получить, проанализировав все три указанные

выше входные переменные. В главе 3 мы будем говорить о линейной регрессии, которая может с успехом применяться для предсказания зарплаты на основании представленного набора данных. В идеале наш прогноз должен учитывать нелинейный характер связи между возрастом и зарплатой. В главе 7 мы рассмотрим серию подходов, позволяющих справиться с этой задачей.

Данные по рынку акций

В примере с набором данных `Wage` мы имели дело с предсказанием *непрерывных* (continuous), или *количественных* (quantitative), значений выходной переменной. Такой вид анализа часто называют *задачей восстановления регрессии* (regression problem). Однако иногда требуется предсказывать нечисловые значения, и в таких случаях мы говорим о *категориальных* (categorical), или *качественных* (qualitative), выходных переменных. В главе 4, например, мы будем исследовать набор данных `Smarket` по рынку акций, в котором содержатся дневные изменения индекса Standard & Poor's 500 (S&P) за пятилетний период с 2001 по 2005 год. Целью в данном случае будет являться предсказание того, покажет ли индекс рост или падение в конкретный день, на основании колебаний индекса за последние пять дней. Как видите, здесь не идет речи о предсказании числовых значений, вместо этого мы задаемся вопросом о том, в какую корзину (рост или снижение) попадет показатель индекса в конкретный день. Такой вид анализа относится к классу *задач классификации* (classification problem). Пользу от модели, способной точно прогнозировать направление движения рынка на основе прочих факторов, невозможно переоценить!

В левой части рис. 1.2 показаны две *диаграммы размаха* (boxplot), также называемые *ящичками с усами*, построенные на основании изменений индекса по сравнению со вчерашним днем: в 648 случаях наблюдался рост индекса в последующий день, в оставшихся 602 случаях – падение. Как видите, диаграммы получились практически идентичными – это можно интерпретировать так, что вчерашнее колебание индекса не оказывает решающего влияния на его следующее изменение. Два следующих графика подтверждают нашу догадку, показывая, что текущее изменение индекса также не зависит и от его колебаний два и три дня назад соответственно. Разумеется, примерно такого результата и следовало ожидать – если бы между соседними изменениями индекса наблюдалась строгая корреляция, можно было бы очень просто выработать прибыльную биржевую стратегию. И все же в главе 4 мы применим к этим исходным данным некоторые методы статистического обучения, которые помогут нам выявить в них слабые тренды и покажут, что, по крайней мере, для этого пятилетнего периода мы можем предсказывать направление движения индекса с вероятностью около 60% (рис. 1.3).

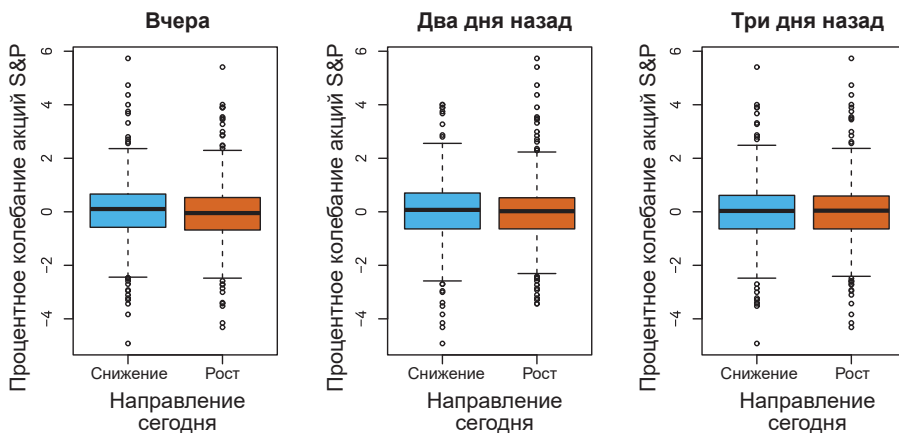


РИС. 1.2 Слева: диаграммы размаха по дням, в которые наблюдался рост или падение индекса, на основании вчерашних изменений. В центре и справа – те же выходные данные, но по изменениям, наблюдавшимся два и три дня назад соответственно

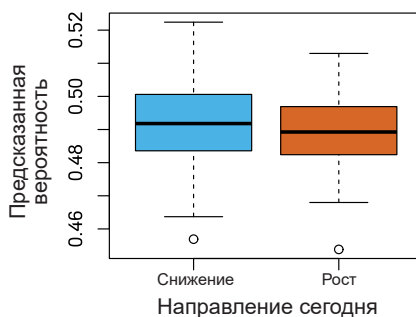


РИС. 1.3 Мы применили модель квадратичного дискриминантного анализа к данным из набора *Smarket* за период с 2001 по 2004 год и предсказали вероятность снижения рынка акций для данных за 2005 год. В среднем предсказанная вероятность снижения рынка оказалась выше для тех дней, когда действительно наблюдался спад. Основываясь на этих результатах, мы можем спрогнозировать направление движения акций на рынке в 60% случаев

Данные об экспрессии генов

В предыдущих двух примерах были проиллюстрированы наборы данных, в которых присутствовали как входные, так и выходные переменные. В то же время существует класс задач, в которых мы имеем дело исключительно со входными переменными без соответствующих им выходных. К примеру, при проведении анализа рынка мы можем располагать демографическими сведениями о некотором количестве существующих или потенциальных покупателей. У нас может возникнуть желание узнать, какие группы покупателей похожи друг на друга, и для этого мы можем сгруппировать их по определенным характе-

ристикам. Данная ситуация известна как задача кластеризации. В отличие от предыдущих примеров, здесь мы не пытаемся предсказать какую-либо выходную переменную.

В главе 12 мы обратимся к методам статистического обучения, предназначенным для задач, в которых отсутствуют естественные выходные переменные. В частности, рассмотрим набор данных NCI60, содержащий 6830 значений уровня экспрессии генов для 64 линий раковых клеток. Вместо предсказания значений выходных переменных нас будет больше интересовать вопрос объединения исследуемых клеточных линий в группы, или кластеры, на основании полученных измерений экспрессии генов. Это довольно сложная задача, поскольку для каждой клеточной линии есть тысячи измерений, что затрудняет процесс визуализации.

В левой части рис. 1.4 эта задача решается путем вывода всех 64 клеточных линий с использованием всего двух числовых критериев: Z_1 и Z_2 . Они представляют собой первые две *главные компоненты* (principal components) данных, с помощью которых информация о 6830 измерениях для каждой клеточной линии сводится к двум числам, или *измерениям* (dimensions). И хотя такое сокращение количества измерений привело к потере части информации, мы, по крайней мере, получили возможность визуально проанализировать данные на предмет образования кластеров. Решение о том, на каком количестве кластеров остановиться, может оказаться непростым. В левой части на рис. 1.4 мы визуально можем выделить по меньшей мере четыре группы клеточных линий, которые отметили на диаграмме разными цветами.

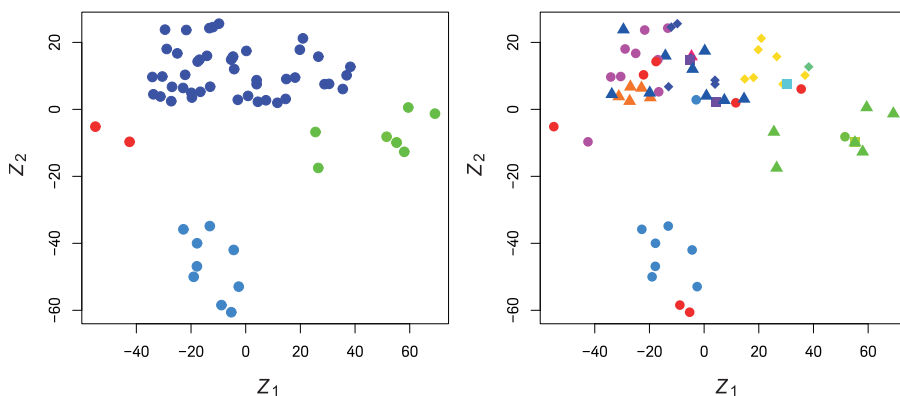


РИС. 1.4 Слева: представление набора данных NCI60 в двумерном пространстве переменных Z_1 и Z_2 . Каждая точка на графике соответствует одной из 64 клеточных линий. В ходе анализа было выявлено четыре группы линий, которые на графике помечены разными цветами. Справа: тот же график, но с добавлением информации о конкретных типах рака. Клеточные линии, соответствующие одному типу заболевания, стремятся объединиться в отчетливо заметные группы в двумерном пространстве

В нашем конкретном наборе данных клеточные линии соответствуют 14 разным типам рака (однако эта информация не была использована при построении левого графика на рис. 1.4). Справа на рис. 1.4 представлен тот же график, что и слева, за исключением того, что каждому типу рака здесь соответствует метка определенной формы и цвета. При взгляде на него вполне очевидно, что клеточные линии, соответствующие одному и тому же типу рака, по большей части располагаются на этом двумерном графике близко друг к другу. Отметим также, что, несмотря на игнорирование информации о типе рака при построении левого графика, полученные в результате кластеры обладают довольно большим сходством с группами, показанными справа, где тип рака был учтен. В некоторой степени этот факт является независимым свидетельством правильности проведенного нами кластерного анализа.

Краткая история статистического обучения

Несмотря на то что сам термин *статистическое обучение* является относительно новым, многие концепции, лежащие в его основе, были разработаны достаточно давно. На заре XIX века был введен в практику *метод наименьших квадратов* (method of least squares), ставший предтечей направления, в наши дни известного как *линейная регрессия* (linear regression). Впервые этот метод был успешно применен в области астрономии. Линейная регрессия используется для предсказания значений количественных переменных, таких как уровень зарплаты конкретного индивида. Для предсказания значений качественных переменных (выживет пациент или нет, будет ли зафиксирован рост или падение рынка акций и т. д.) в 1936 году был предложен метод *линейного дискриминантного анализа* (linear discriminant analysis). В 1940-х годах сразу несколько авторов в качестве альтернативы предложили использовать для подобных расчетов метод *логистической регрессии* (logistic regression). В начале 1970-х появился термин *обобщенная линейная модель* (generalized linear model), который описывал целый класс методов статистического обучения, включая линейную и логистическую регрессию в виде особых случаев.

К концу 1970-х годов свет увидел целый ряд техник для изучения данных. Однако почти все они базировались на линейных методах, поскольку в то время было еще недостаточно вычислительных ресурсов для качественной обработки нелинейных методов. В 1980-х годах компьютерные технологии сделали заметный шаг вперед, что позволило относительно недорого выполнять нелинейный анализ. В результате к середине 1980-х появились *деревья регрессии и классификации* (clas-

sification and regression trees), а следом за этим методом свет увидели *обобщенные аддитивные модели* (generalized additive models). Также в 1980-е приобрели популярность *нейронные сети* (neural networks), а в 1990-е появился и *метод опорных векторов* (support vector machines).

С тех пор статистическое обучение было выделено в обособленную ветвь статистики и главным образом сконцентрировалось на вопросах создания моделей с учителем и без учителя, а также на прогнозировании. В последние годы отрасль статистического обучения добилась заметного прогресса в первую очередь за счет появления мощных и достаточно дружественных инструментов, одним из которых является популярный и бесплатный язык программирования Python. И в этом отношении у статистического обучения есть все возможности для дальнейшего развития и перехода от набора техник, разрабатываемых и используемых подготовленными специалистами в области статистики и науки о данных, к инструментарию, которым смогут пользоваться широкие массы.

О книге

Книга «Основы статистического обучения», написанная Хасти (Hastie), Тибширани (Tibshirani) и Фридманом (Friedman), впервые была опубликована в 2001 году. С тех пор она стала незаменимым справочным изданием по основам статистического машинного обучения. Своим успехом эта книга отчасти была обязана полноценной глубокой проработке важнейших тем в области статистического обучения и относительной легкости подачи материала по сравнению со многими другими трудами тех лет в области статистики. Но главным фактором огромной популярности книги стала сама ее тема. В то время было ощущение, что интерес к статистическому обучению вот-вот взорвется. И книга «Основы статистического обучения» стала одним из первых источников знаний эту тему, написанных понятным человеческим языком.

С момента публикации книги область статистического обучения продолжила развиваться, при этом ее развитие происходило по двум направлениям. Первое касалось разработки новых более развитых подходов и методов статистического обучения, призванных ответить на наиболее насущные вопросы из самых разных областей. Параллельно с этим ширилась и аудитория заинтересованных в статистическом обучении. В 1990-е годы стремительный рост доступности вычислительных ресурсов обусловил повышение интереса к этой области у людей, далеких от статистики, но желающих использовать передовые инструменты для анализа своих данных. К сожалению, исключитель-

ная техническая направленность статистических подходов в то время ограничивала приток новых людей в эту область, основу которой составляли эксперты в области статистики, компьютерных наук и смежных технологий, обладавшие достаточным опытом (и временем) для понимания и реализации этих подходов.

В последние годы широкое распространение получили программные пакеты, существенно облегчающие и берущие на себя реализацию рутинных задач, свойственных для методов статистического обучения. Одновременно с этим представители все большего количества отраслей, включая бизнес, здравоохранение, генетику, социальные науки и пр., начали осознавать пользу от практического применения мощных инструментов статистического обучения. В результате эта область превратилась из исключительно академической и узконаправленной в массовую, потенциально доступную для широкой аудитории. И эта тенденция вряд ли ослабится в ближайшее время, особенно с учетом появления все большего числа источников данных и программного обеспечения для их анализа.

Целью книги, которую вы держите в руках, является содействие облегчению перехода статистического обучения из академического поля в массовую культуру. Она не призвана заменить собой книгу «Основы статистического обучения», в которой дается более всеобъемлющий материал с более глубоким погружением в тему. Мы рассматриваем книгу «Основы статистического обучения» как важное дополнение для специалистов, обладающих необходимым уровнем образования в сфере статистики, машинного обучения и смежных дисциплин, в деле понимания технических подробностей, лежащих в основе подходов статистического обучения. В то же время нельзя не отметить наметившегося роста сообщества, к которому примкнули люди с другими интересами и фоновыми знаниями. Таким образом, освободилось место для менее технически наполненной и более дружелюбной версии книги «Основы статистического обучения».

Посвятив преподаванию этих тем не один год, мы пришли к выводу о том, что этими дисциплинами интересуются магистранты и аспиранты из таких разных сфер, как бизнес-администрирование, биология и компьютерные науки, и студенты старших курсов, изучающие количественные методы. Для такой разношерстной публики важно понимать сами модели и их сущность, а также плюсы и минусы различных подходов. В то же время многие технические аспекты, лежащие в основе методов статистического обучения, такие как алгоритмы оптимизации и теоретические свойства, могут не входить в их сферу интересов. Мы уверены, что студентам нет необходимости постигать все технические тонкости и нюансы, чтобы активно использовать различные методологии и тем самым приносить пользу своей отрасли.

Книга «Введение в статистическое обучение» основывается на четырех предпосылках.

1. *Многие методы статистического обучения могут с успехом применяться в широком спектре академических и прочих дисциплин и не ограничиваются лишь статистической наукой.* Мы уверены, что многие современные процедуры статистического обучения должны получить и получают широкое распространение и применение по примеру того, как сегодня используются классические методы вроде линейной регрессии. Потому мы решили не распылять внимание на все существующие подходы (охватить все в любом случае не получится), а сосредоточиться на основных методах, которые нам кажутся наиболее полезными и применимыми на практике.
2. *Статистическое обучение не стоит рассматривать как последовательность черных ящиков.* Не существует единого подхода, идеального для всех ситуаций. Без понимания того, как именно крутятся шестеренки внутри ящика и как они друг с другом соединены, невозможно выбрать нужный ящик. Поэтому мы постарались тщательно описать модель, ее смысл, предпосылки и компромиссы, лежащие в основе рассматриваемых методов.
3. *Хотя знать, как крутятся шестеренки внутри ящика, очень важно, нет никакой необходимости обладать умениями собирать сам механизм внутри ящика самостоятельно.* Таким образом, мы сведем к минимуму технические подробности, связанные с процедурами обучения моделей и техническими свойствами. Мы полагаем, что читатель обладает некой математической базой, но при этом не требуем наличия ученой степени в этой области. К примеру, мы почти полностью избавились от использования в книге матричной алгебры, в связи с чем вы сможете комфортно читать ее, даже не обладая глубокими познаниями в области матриц и векторов.
4. *Мы предполагаем, что читатель заинтересован в применении методов статистического обучения на практике.* И чтобы облегчить ему эту задачу, а также мотивировать на применение обсуждаемых техник, в конце каждой главы мы включили лабораторные работы. В каждой из них мы демонстрируем применение изучаемых методов на реалистичных примерах. В процессе преподавания этого материала на курсах мы примерно треть времени отводили на решение лабораторных задач и нашли такой подход исключительно полезным. Студенты, которые не были тесно связаны с компьютерными науками и опасались практических занятий, уже в течение одной четверти или семестра осваивались и входили в нужный ритм. В первых изданиях этой книги в лабораторных работах использовался язык программирования R. С тех пор большое распространение в области науки

о данных приобрел язык Python, и в этой версии книги мы решили полностью перейти на него. Количество доступных библиотек на Python растет с каждым месяцем, и по секциям с импортом в начале каждой лабораторной работы вы сможете понять, что мы используем только наиболее подходящие из них. Также мы собрали дополнительный код и функциональность в отдельном пакете *ISLP*, которым вы можете пользоваться. В то же время лабораторные работы являются необязательными и могут быть пропущены при чтении, если вы хотите использовать другое программное обеспечение или вовсе не собираетесь применять на практике полученные знания.

Для кого предназначена эта книга?

Эту книгу следует прочесть тем, кто заинтересован в использовании современных статистических методов при моделировании и прогнозировании на основе данных. В эту группу могут входить ученые, инженеры, аналитики данных, специалисты в области науки о данных, специалисты по количественному анализу, а также менее технически подкованные читатели, не обладающие специальным математическим образованием и занимающиеся, к примеру, бизнесом или социальными науками. При этом мы предполагаем, что у наших читателей за плечами есть как минимум один вводный курс по статистике. Начальные знания в области линейной регрессии тоже будут весьма полезны, хоть и не обязательны, поскольку в главе 3 мы обсудим все ключевые концепции, относящиеся к этому методу. Что касается математического уровня книги, то мы считаем его умеренным, без особых требований – читателю даже не потребуются дополнительные знания в области матричных операций. Также в книге мы приведем краткий экскурс по языку Python. Знание других языков программирования, таких как MATLAB и R, приветствуется, но также не является обязательным требованием для чтения книги.

Первое издание этой книги использовалось в качестве пособия при обучении магистров и аспирантов в области бизнеса, экономики, компьютерных наук, биологии, геологии, психологии и многих других естественных и социальных наук. Также книга была использована при обучении студентов старших курсов бакалавриата, которые до этого уже проходили дисциплину, связанную с линейной регрессией. На курсах с более серьезной математической подачей, где основным учебным пособием служит книга «Основы статистического обучения», данная книга может использоваться в качестве дополнительной литературы при изучении вычислительных аспектов различных методов.

Обозначения и матричная алгебра по-простому

Выбирать терминологию и нотацию для учебника всегда очень непросто. По большей части при написании этой книги мы решили придерживаться условных обозначений, принятых в книге «Основы статистического обучения».

Мы будем использовать букву n для обозначения количества различающихся точек данных, или наблюдений, в нашей выборке. Буквой p мы обозначаем количество переменных, которые могут быть использованы для предсказаний. К примеру, в наборе данных `Wage` содержится 11 переменных для 3000 человек. Таким образом, n в нашем случае будет равно 3000, а $p = 11$ (переменные `year`, `age`, `gase` и др.). Обратите внимание, что на протяжении всей книги для обозначения имен переменных мы будем использовать цветной шрифт: *Имя переменной*.

В некоторых примерах число p может оказаться достаточно большим и исчисляться в тысячах или даже миллионах. И такие ситуации – далеко не редкость, например при анализе современных биологических данных или рекламных данных в веб-аналитике.

В основном мы будем обозначать как x_{ij} значение j -й переменной для i -го наблюдения, где $i = 1, 2, \dots, n$, а $j = 1, 2, \dots, p$. На протяжении этой книги буквой i будет обозначаться выборка или конкретное наблюдение (от 1 до n), а буквой j – номер переменной (от 1 до p). С помощью \mathbf{X} мы будем обозначать матрицу $n \times p$, в которой элемент с индексом (i, j) будет x_{ij} . Получим следующую матрицу:

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

Тем, кто не очень хорошо знаком с матрицами, будет легче представить себе \mathbf{X} в виде таблицы с числами, состоящей из n строк и p столбцов.

Иногда нас будут отдельно интересовать строки из матрицы \mathbf{X} , которые записываются в виде последовательностей x_1, x_2, \dots, x_n . Здесь x_i представлен вектором длины p и содержит p значений переменных для i -го наблюдения, как показано ниже:

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

(Векторы по умолчанию представляются в виде колонок.) К примеру, в наборе данных `Wage` элемент x_i представлен вектором длины 11, содержащим значения переменных `year`, `age`, `race` и др. для i -го наблюдения. Во всех остальных случаях нас будут интересовать колонки матрицы \mathbf{X} , которые мы записываем как $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Каждый из этих векторов обладает длиной n , т. е.:

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

Например, в наборе данных `Wage` в векторе \mathbf{x}_1 содержится $n = 3000$ значений переменной `year`. Используя эту нотацию, матрица \mathbf{X} может быть записана как

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

или

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

Символ T означает *транспонирование* (transpose) матрицы или вектора. Например,

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

тогда как

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

Мы используем обозначение y_i для указания на i -е наблюдение переменной, которую мы хотим предсказать, например `wage`. В векторной форме набор из всех n наблюдений можно записать так:

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

Таким образом, наблюдаемые нами данные состоят из пар $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, где каждый x_i представлен вектором длины p . Если $p = 1$, то x_i – простое скалярное значение.

В данной книге вектор длины n всегда будет обозначаться буквой в нижнем регистре, выделенной жирным шрифтом, как показано ниже:

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

В то же время векторы длиной, отличной от n , такие как векторы с длиной p , как на (1.1), будут обозначаться буквами в нижнем регистре с обычным шрифтом, например a . Скаляры также будут обозначаться обычным шрифтом, т. е. a . В редких случаях, когда они будут использоваться совместно, мы будем отдельно упоминать, что именно имеется в виду. Матрицы будут обозначаться жирным шрифтом в верхнем регистре: \mathbf{A} . Случайные переменные мы будем писать в верхнем регистре обычным шрифтом, например A , вне зависимости от их размерности.

Иногда нам нужно будет продемонстрировать размерность какого-либо объекта. Для обозначения того, что объект представляет скалярную величину, мы будем использовать нотацию $a \in \mathbb{R}$. Чтобы показать, что это вектор длины k , мы будем писать $a \in \mathbb{R}^k$ (или $a \in \mathbb{R}^n$, если речь идет о векторе длины n). Если объект представляет собой матрицу $r \times s$, будем обозначать его так: $a \in \mathbb{R}^{r \times s}$.

Мы будем избегать использования матричной алгебры всегда, когда это возможно. Однако иногда полностью от нее отказаться будет просто невозможно. В этих редких случаях для понимания происходящего от вас потребуется знание концепции перемножения двух матриц. Предположим, $A \in \mathbb{R}^{r \times d}$, а $B \in \mathbb{R}^{d \times s}$. Произведение этих двух матриц будет обозначаться как \mathbf{AB} . (i, j) -й элемент итоговой матрицы вычисляется путем перемножения каждого элемента i -й строки матрицы \mathbf{A} на соответствующий элемент j -го столбца матрицы \mathbf{B} . То есть $(\mathbf{AB})_{ij} = \sum_{k=1}^d a_{ik} b_{kj}$. Рассмотрим в качестве примера две матрицы:

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

В этом случае

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

Обратите внимание, что в результате мы получим матрицу размером $r \times s$. При этом вычислить произведение \mathbf{AB} возможно только в случае, если количество столбцов в матрице \mathbf{A} соответствует количеству строк в матрице \mathbf{B} .

Структура книги

В главе 2 мы познакомимся с базовыми терминами и концепциями, лежащими в основе статистического обучения. В этой главе мы также рассмотрим классификатор k -ближайших соседей, представляющий собой простейший метод, с успехом справляющийся с самыми разными задачами. В главах 3 и 4 мы углубимся в классические линейные методы регрессии и классификации. В частности, в главе 3 мы рассмотрим линейную регрессию, являющуюся основой для всех методов регрессионного анализа. В главе 4 мы взглянем на два наиболее важных метода классической классификации: логистическую регрессию и линейный дискриминантный анализ.

Центральной проблемой всех случаев использования статистического обучения является выбор наиболее подходящего метода для конкретной ситуации. Таким образом, в пятой главе книги мы познакомимся с перекрестной проверкой (кросс-валидацией) и бутстрепом, которые могут быть использованы для оценки точности применения разных методов с целью выбора наиболее подходящего.


Большинство последних исследований в области статистического обучения сконцентрированы вокруг нелинейных методов. Однако их линейные аналоги зачастую превосходят нелинейные в плане интерпретируемости, а иногда и точности. Так что всю шестую главу мы посвятили набору линейных методов, как классических, так и более современных, предлагающих определенные улучшения по сравнению с линейной регрессией. Речь пойдет о пошаговом отборе, гребневой регрессии, регрессии на главные компоненты и лассо-регрессии.

Оставшиеся главы книги перенесут нас в мир нелинейных методов статистического обучения. Для начала в главе 7 мы рассмотрим несколько нелинейных методов, хорошо себя зарекомендовавших при работе с одной входной переменной. После этого мы посмотрим, как эти методы могут быть использованы для построения нелинейных аддитивных моделей с более чем одной входной переменной. В главе 8 мы исследуем методы на основе деревьев решений, включая бэггинг,

бустинг и случайные леса. Глава 9 будет посвящена методу опорных векторов, представляющему совокупность подходов для выполнения как линейной, так и нелинейной классификации. В главе 10 мы обратимся к теме глубокого обучения – подхода для нелинейной регрессии и классификации, получившего в последние годы широкое распространение. В главе 11 мы уделим внимание анализу выживаемости, представляющему собой особый вид регрессионного подхода для ситуаций, когда выходная переменная является цензурированной, т. е. содержит неполную информацию.

В главе 12 мы рассмотрим методы обучения без учителя, когда у нас есть входные переменные, но нет выходных. В частности, мы познакомимся с анализом главных компонент, кластеризацией методом k -средних и иерархической кластеризацией. Наконец, в главе 13 мы коснемся важной темы, связанной с множественной проверкой гипотез.

В конце каждой главы вас ждет одна или несколько лабораторных работ на языке Python, в которых вы сможете проверить на практике все изученные в главе методы статистического анализа. В этих работах будут выявляться сильные и слабые стороны разных подходов, а также вы освоите синтаксис команд для реализации того или иного метода. Читатель может выполнять лабораторные работы в своем собственном темпе, кроме того, эти работы могут стать предметом совместных обсуждений на групповых занятиях. В каждой работе мы будем демонстрировать результаты, полученные во время их выполнения на момент написания книги. Но со временем реализации интерпретатора Python претерпевают изменения, и библиотеки, которые мы будем использовать в лабораторных, также не стоят а месте в плане развития. Таким образом, представленные в книге результаты в какой-то момент могут начать отличаться от того, что получите на практике вы. По возможности и при необходимости мы будем выкладывать на сайте книги обновления к лабораторным работам.

Символом  мы будем помечать разделы или упражнения повышенной сложности. Эти материалы могут быть пропущены читателями, не желающими глубоко погружаться в предмет или не обладающими достаточными знаниями в области математики.

Используемые в лабораторных работах и упражнениях наборы данных

В этой книге мы будем демонстрировать примеры применения методов статистического обучения в самых разных областях, включая маркетинг, финансы, биологию и др. В пакете ISLP содержатся все необходимые наборы данных для выполнения предложенных в книге

упражнений и лабораторных работ. Единственный набор данных, которого в пакете нет, – это `USArrests`, он располагается в дистрибутиве R, и в разделе 12.5.1 мы покажем, как можно получить к нему доступ из языка Python. В табл. 1.1 содержится сводная информация о наборах данных, используемых в этой книге. Несколько из приведенных наборов располагаются также на сайте книги в виде текстовых файлов для использования в главе 2.

ТАБЛИЦА 1.1. *Список наборов данных, необходимых для выполнения лабораторных работ и упражнений из этой книги. Все наборы данных доступны в пакете `ISLP`, за исключением набора `USArrests`, который является частью дистрибутива R, но доступен также и в Python*

Набор данных	Описание
Auto	Расход бензина, мощность и прочая информация о машинах
Bikeshare	Почасовая информация о программе проката велосипедов в Вашингтоне
Boston	Стоимость объектов недвижимости и прочая информация по районам переписи Бостона
BrainCancer	Информация о выживаемости пациентов с диагнозом рак мозга
Caravan	Информация об индивидуальном страховании трейлеров
Carseats	Данные о продаже автокресел в 400 магазинах
College	Демографическая, образовательная и прочая информация о колледжах в США
Credit	Информация о задолженности по кредитным картам для 400 клиентов
Default	Данные о возможном невыполнении обязанностей по долгам для компании, выпускающей кредитные карты
Fund	Информация о работе 2000 управляющих хеджевых фондов за 50 месяцев
Hitters	Статистическая информация и данные о зарплате по игрокам в бейсбол
Khan	Данные об экспрессии генов по четырем типам рака
NCI60	Данные об экспрессии генов по 64 клеточным линиям рака
NYSE	Доходность, волатильность и объемы на Нью-Йоркской фондовой бирже
OJ	Данные о продажах апельсинового сока марок Citrus Hill и Minute Maid
Portfolio	Информация о прошлой стоимости финансовых активов для распределения портфелей
Publication	Данные о публикации информации о 244 клинических исследованиях
Smarket	Данные о дневных изменениях индекса S&P 500 за пятилетний период
USArrests	Статистика преступности по 100 000 резидентам из 50 штатов США
Wage	Данные о доходах мужчин, проживающих в центральной части Атлантического региона США
Weekly	Данные о работе фондовой биржи за 1089 недель в течение 21 года

Сайт книги

Официальный сайт данной книги: <https://www.statlearning.com>. На этом сайте содержится ряд вспомогательных ресурсов, включая наш пакет на Python, используемых в книге, а также дополнительные наборы данных.

Источники

Некоторые диаграммы, приведенные в данной книге, были взяты из книги «Основы статистического обучения». Это рис. 6.7, 8.3 и 12.14. Все остальные графики были взяты из книги «Введение в статистическое обучение с примерами на языке R», за исключением рис. 13.10, который был получен с помощью инструментов Python.