

Оглавление

От переводчика	10
Предисловие	11
1 Введение	13
2 Статистическое обучение	27
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
3 Линейная регрессия	71
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
4	Классификация	143
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	174
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 прицепах	182
4.7	Упражнения	186
5	Методы создания повторных выборок	192
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
6	Отбор и регуляризация линейных моделей	221
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
7	Выходя за пределы линейности	288
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 сочленения	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 λ	302
7.6	Локальная регрессия	304
7.7	Обобщенные аддитивные модели	307
7.7.1	GAM для регрессионных задач	307
7.7.2	GAM для задач классификации	311
7.8	Лабораторная работа: нелинейные модели	311
7.8.1	Полиномиальная регрессия и ступенчатые функции	313
7.8.2	Сплайны	317
7.8.3	GAM	319
7.9	Упражнения	322
8	Методы, основанные на деревьях решений	328
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
9	Машины опорных векторов	364
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
10	Обучение без учителя	402
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
	Предметный указатель	450

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 верстки¹. 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.

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.

¹ 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 развилась 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)² — 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 Чулдечову, Патрика Данахера, Уилла Фитиана, Луэллу Фу, Сэма Гросса, Макса Гразьера Г'Селла, Кортни Паулсон, Ксингао Кьяо, Элизу Шенг, Ноа Симон, Кена Минга Тана и Ксина Лу Тана.

² 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 Хасте (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.

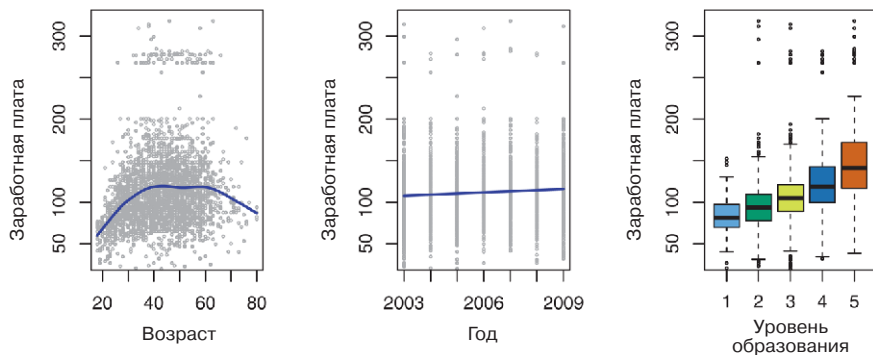


РИСУНОК 1.1. Таблица `Wage` с данными по заработной плате мужчин из центрально-атлантического региона США. Слева: `wage` как функция от `age`. В среднем `wage` увеличивается одновременно с `age` до возраста около 60 лет, после чего начинает снижаться. В центре: `wage` как функция от `year`. В период 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!

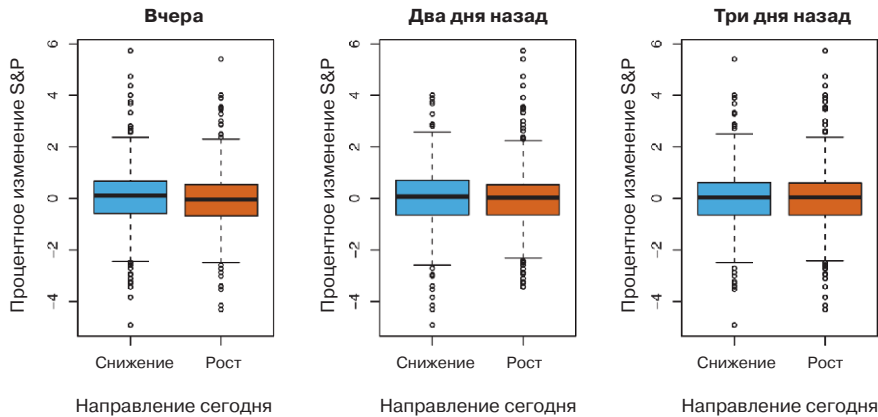


РИСУНОК 1.2. Слева: диаграмма размахов, отражающая процентное изменение индекса S&P по сравнению со вчерашним значением для дней, когда происходят рост или снижение рынка (по данным Smarket). В центре и справа: то же, но показаны процентные изменения по сравнению с двумя и тремя предыдущими днями соответственно

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).

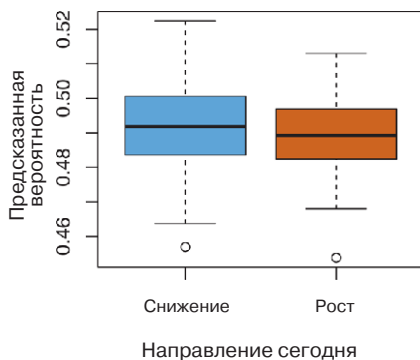


РИСУНОК 1.3. Мы подошли к квадратичной дискриминантной модели для части данных Smarket, соответствующей периоду с 2001 по 2004 г., и предсказали вероятность снижения рынка акций для данных за 2005 г. В среднем предсказанная вероятность снижения рынка выше для дней, когда снижение в действительности имело 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 поместили 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.

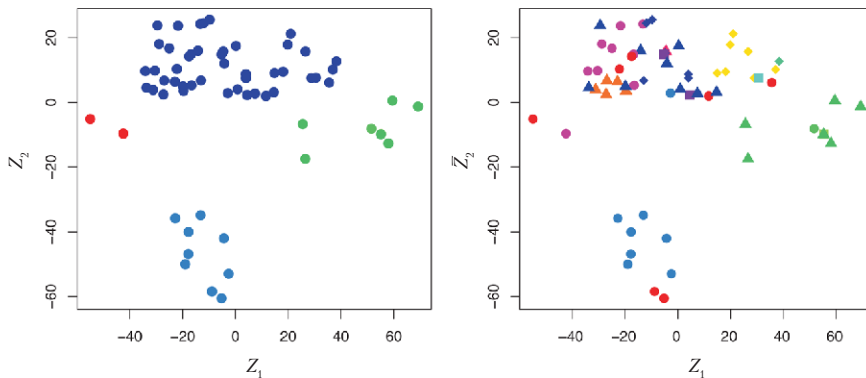


РИСУНОК 1.4. Слева: представление данных по уровню экспрессии генов *NCI60* в двумерном пространстве, образованном переменными Z_1 и Z_2 . Каждая точка соответствует одной из 64 клеточных линий. Клеточные линии образуют примерно четыре группы, которые мы представили разными цветами. Справа: то же, что 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 Веддербурн 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. Хасти 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 Хасти, 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

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

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, \mathbf{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 избежали 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 $\mathbf{A} \in \mathbb{R}^{r \times d}$,

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru