

УДК 519.25/.6:004.434R

ББК 22.17с5

Д40

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

Д40 Введение в статистическое обучение с примерами на языке R. Изд. второе, испр. Пер. с англ. С. Э. Мاستицкого – М.: ДМК Пресс, 2017. – 456 с.: ил.

ISBN 978-5-97060-495-3

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

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

УДК 519.25/.6:004.434R

ББК 22.17с5

Translation from the English language edition:

An Introduction to Statistical Learning

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

Copyright © Springer Science+Business Media New York 2013

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

All Rights Reserved.

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

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

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

ISBN 978-1-4614-7137-0 (англ.)

ISBN 978-5-97060-495-3 (рус.)

Copyright © Springer Science+Business Media New York, 2013

© Издание, оформление, перевод, ДМК Пресс, 2017

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 прицепам	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