

# Содержание

<b>Введение</b>	<b>15</b>
Для кого написана эта книга	16
Для преподавателей	17
Посвящение	19
Благодарности	19
Предупреждения	21
Ждем ваших отзывов!	21
<b>Глава 1. Что такое наука о данных?</b>	<b>23</b>
1.1. Информатика, наука о данных и реальная наука	24
1.2. Формирование интересных вопросов для данных	27
1.2.1. Бейсбольная энциклопедия	28
1.2.2. Интернет-база кинофильмов (IMDb)	30
1.2.3. N-граммы Google	33
1.2.4. Записи нью-йоркских такси	35
1.3. Свойства данных	38
1.3.1. Структурированные или неструктурированные данные	39
1.3.2. Количественные данные или качественные	39
1.3.3. Большие данные или небольшие	40
1.4. Классификация и регрессия	41
1.5. Видеоматериал: The Quant Shop	42
1.5.1. Конкурсы Kaggle	45
1.6. О случаях из жизни	45
1.7. Случай из жизни: ответ на правильный вопрос	47
1.8. Дополнительная информация	49
1.9. Упражнения	50
<b>Глава 2. Математические основы</b>	<b>53</b>
2.1. Вероятность	53
2.1.1. Вероятность против статистики	55
2.1.2. Составные события и независимость	56
2.1.3. Условная вероятность	58
2.1.4. Вероятностные распределения	59
2.2. Описательная статистика	61
2.2.1. Поиск центра	62
2.2.2. Поиск дисперсии	64

2.2.3. Интерпретация дисперсии	65
2.2.4. Характеристика распределений	67
2.3. Корреляционный анализ	68
2.3.1. Коэффициенты корреляции Пирсона и Спирмена	70
2.3.2. Сила и значение корреляции	72
2.3.3. Корреляция не означает причину!	74
2.3.4. Обнаружение автокорреляцией периодичности	75
2.4. Логарифмы	77
2.4.1. Логарифмы и умножение вероятностей	77
2.4.2. Логарифмы и соотношения	78
2.4.3. Логарифмы и нормализация асимметричных распределений	79
2.5. Случай из жизни: поиск дизайнерских генов	81
2.6. Дополнительная информация	83
2.7. Упражнения	83
<b>Глава 3. Манипулирование данными</b>	<b>89</b>
3.1. Языки для науки о данных	90
3.1.1. Важность окружения интерактивной оболочки	92
3.1.2. Стандартные форматы данных	93
3.2. Сбор данных	97
3.2.1. Охота на данные	97
3.2.2. Скрепинг данных	101
3.2.3. Регистрация	102
3.3. Очистка данных	103
3.3.1. Ошибки против артефактов	104
3.3.2. Совместимость данных	106
3.3.3. Как справиться с отсутствующими значениями	112
3.3.4. Обнаружение выброса	115
3.4. Случай из жизни: игры на фондовом рынке	116
3.5. Краудсорсинг	118
3.5.1. Демонстрация пенсов	119
3.5.2. Когда толпа проявляет мудрость?	120
3.5.3. Механизмы объединения	121
3.5.4. Службы краудсорсинга	122
3.5.5. Игрофикация	128
3.6. Дополнительная информация	129
3.7. Упражнения	130
<b>Глава 4. Оценки и ранги</b>	<b>135</b>
4.1. Индекс массы тела (BMI)	136
4.2. Разработка систем оценки	139

4.2.1. Золотые стандарты и прокси	139
4.2.2. Оценки или ранги	141
4.2.3. Выявление хороших функций оценки	143
4.3. Z-оценки и нормализация	145
4.4. Передовые методы ранжирования	146
4.4.1. Рейтинг Эло	147
4.4.2. Слияние рейтингов	150
4.4.3. Рейтинг на основе диграфа	152
4.4.4. Алгоритм PageRank	154
4.5. Случай из жизни: месть Клайда	154
4.6. Теорема Эрроу о невозможности	158
4.7. Случай из жизни: кто больше?	160
4.8. Дополнительная информация	163
4.9. Упражнения	164
<b>Глава 5. Статистический анализ</b>	<b>167</b>
5.1. Статистические распределения	168
5.1.1. Биномиальное распределение	169
5.1.2. Нормальное распределение	170
5.1.3. Значения нормального распределения	173
5.1.4. Распределение Пуассона	174
5.1.5. Распределение по степенному закону	176
5.2. Выборка из распределений	180
5.2.1. Случайная выборка вне одного измерения	181
5.3. Статистическая значимость	183
5.3.1. Значение значимости	184
5.3.2. Т-критерий: сравнение средних значений совокупностей	186
5.3.3. Критерий Колмогорова–Смирнова	188
5.3.4. Поправка Бонферрони	191
5.3.5. Частота ложных открытий	192
5.4. Случай из жизни: поиск фонтана молодости	193
5.5. Критерии перестановки и $p$ -значения	194
5.5.1. Создание случайных перестановок	197
5.5.2. Страйк хитов Ди Маджо	198
5.6. Байесовский вывод	200
5.7. Дополнительная информация	202
5.8. Упражнения	202
<b>Глава 6. Визуализация данных</b>	<b>207</b>
6.1. Исследовательский анализ данных	208
6.1.1. Противостояние новому набору данных	208

6.1.2. Сводная статистика и квартет Энскомба	212
6.1.3. Инструменты визуализации	214
6.2. Выработка эстетики визуализации	215
6.2.1. Максимизация соотношения данных и чернил	216
6.2.2. Минимизация фактора лжи	217
6.2.3. Минимизация неинформативных элементов	219
6.2.4. Правильные масштабы и ясные маркеры	221
6.2.5. Эффективное использование цвета	222
6.2.6. Сила повторения	223
6.3. Типы диаграмм	224
6.3.1. Табличные данные	226
6.3.2. Точечные и линейные графики	229
6.3.3. Диаграммы рассеяния	233
6.3.4. Гистограммы и круговые диаграммы	236
6.3.5. Гистограммы	240
6.3.6. Карты данных	244
6.4. Примеры правильной визуализации	246
6.4.1. Расписание поездов Маре	246
6.4.2. Карта распространения холеры Сноу	248
6.4.3. Карта погоды в Нью-Йорке	248
6.5. Чтение графиков	250
6.5.1. Соккрытие распределения	250
6.5.2. Переинтерпретация дисперсии	251
6.6. Интерактивная визуализация	252
6.7. Случай из жизни: текстовая карта мира	254
6.8. Дополнительная информация	256
6.9. Упражнения	257
<b>Глава 7. Математические модели</b>	<b>261</b>
7.1. Философия моделирования	261
7.1.1. Бритва Оккама	262
7.1.2. Дилемма смещения-дисперсии	263
7.1.3. Что бы сделал Нейт Силвер?	263
7.2. Классификация моделей	267
7.2.1. Линейные модели против нелинейных	267
7.2.2. Черные ящики против описательных моделей	267
7.2.3. Модели первого принципа против моделей управляемых данными	269
7.2.4. Стохастические модели против детерминированных	270
7.2.5. Плоские модели против иерархических	271

7.3. Базовые модели	272
7.3.1. Базовые модели для классификации	273
7.3.2. Базовые модели для прогнозирования значения	274
7.4. Оценка моделей	275
7.4.1. Оценка классификаторов	276
7.4.2. Кривые рабочей характеристики приемника (ROC)	282
7.4.3. Оценка мультиклассовых систем	284
7.4.4. Оценка моделей прогнозирования значений	287
7.5. Оценка среды	289
7.5.1. Гигиена данных для оценки	291
7.5.2. Усиление малых оценочных наборов	293
7.6. Случай из жизни: 100% корректности	295
7.7. Имитационные модели	297
7.8. Случай из жизни: вычисление ставок	298
7.9. Дополнительная информация	302
7.10. Упражнения	302
<b>Глава 8. Линейная алгебра</b>	<b>307</b>
8.1. Сила линейной алгебры	307
8.1.1. Интерпретация линейных алгебраических формул	309
8.1.2. Геометрия и векторы	310
8.2. Визуализация матричных операций	312
8.2.1. Сложение матриц	313
8.2.2. Умножение матриц	314
8.2.3. Применение матричного умножения	316
8.2.4. Единичные матрицы и инверсия	320
8.2.5. Инверсия матриц и линейные системы	321
8.2.6. Ранг матриц	323
8.3. Разложение матриц	324
8.3.1. Разложение матрицы признаков	325
8.3.2. Разложение LU матрицы и детерминанты	327
8.4. Собственные значения и собственные векторы	328
8.4.1. Свойства собственных значений	328
8.4.2. Вычисление собственных значений	329
8.5. Разложение по собственным значениям	330
8.5.1. Разложение по сингулярному значению	332
8.5.2. Анализ основных компонентов	334
8.6. Случай из жизни: человеческий фактор	336
8.7. Дополнительная информация	338
8.8. Упражнения	338

<b>Глава 9. Линейная и логистическая регрессии</b>	<b>341</b>
9.1. Линейная регрессия	342
9.1.1. Линейная регрессия и двойственность	342
9.1.2. Ошибка в линейной регрессии	344
9.1.3. Нахождение оптимального соответствия	344
9.2. Лучшие регрессионные модели	346
9.2.1. Удаление выбросов	346
9.2.2. Поиск соответствия нелинейных функций	347
9.2.3. Функция и целевое масштабирование	349
9.2.4. Работа с сильно коррелирующими признаками	352
9.3. Случай из жизни: водитель такси	353
9.4. Регрессия как подбор параметров	355
9.4.1. Выпуклые пространства параметров	356
9.4.2. Поиск с градиентным спуском	358
9.4.3. Какова правильная скорость обучения?	360
9.4.4. Стохастический градиентный спуск	362
9.5. Упрощение моделей с помощью регуляризации	363
9.5.1. Гребневая регрессия	364
9.5.2. Регрессия LASSO	365
9.5.3. Компромисс между точностью соответствия и сложностью	366
9.6. Классификация и логистическая регрессия	367
9.6.1. Регрессия для классификации	368
9.6.2. Границы принятия решений	369
9.6.3. Логистическая регрессия	370
9.7. Проблемы логистической классификации	374
9.7.1. Сбалансированные учебные классы	374
9.7.2. Мультиклассовая классификация	376
9.7.3. Иерархическая классификация	378
9.7.4. Функции разбиения и полиномиальная регрессия	379
9.8. Дополнительная информация	381
9.9. Упражнения	381
<b>Глава 10. Методы измерения расстояний и сетей</b>	<b>385</b>
10.1. Измерение расстояний	385
10.1.1. Метрики расстояния	386
10.1.2. Метрика расстояния $L_k$	387
10.1.3. Работа в более высоких размерностях	389
10.1.4. Размерный эгалитаризм	390
10.1.5. Точки или векторы	391
10.1.6. Расстояния между вероятностными распределениями	393

10.2. Классификация ближайших соседей	394
10.2.1. В поисках хороших аналогий	396
10.2.2. $k$ ближайших соседей	397
10.2.3. Поиск ближайших соседей	399
10.2.4. Локальное хеширование	401
10.3. Графы, сети и расстояния	404
10.3.1. Взвешенные графы и индуцированные сети	405
10.3.2. Классификация графов	406
10.3.3. Теория графов	408
10.4. PageRank	410
10.5. Кластеризация	414
10.5.1. Кластеризация методом $k$ -средних	417
10.5.2. Агломерационная кластеризация	423
10.5.3. Сравнение кластеров	429
10.5.4. Подобие графов и кластеризация на основе сегментации	430
10.6. Случай из жизни: кластерная бомбардировка	433
10.7. Дополнительная информация	435
10.8. Упражнения	435
<b>Глава 11. Машинное обучение</b>	<b>441</b>
11.1. Наивный байесовский классификатор	444
11.1.1. Формулировка	445
11.1.2. Как справиться с нулевым счетом (дисконтирование)	447
11.2. Классификаторы дерева решений	449
11.2.1. Построение деревьев решений	451
11.2.2. Реализация исключающего ИЛИ	453
11.2.3. Ансамбли деревьев решений	454
11.3. Бустинг и ансамблевое обучение	455
11.3.1. Голосование с классификаторами	456
11.3.2. Алгоритмы бустинга	457
11.4. Метод опорных векторов	460
11.4.1. Линейные SVM	462
11.4.2. Нелинейные SVM	463
11.4.3. Ядра	465
11.5. Степени контроля	465
11.5.1. Обучение с учителем	466
11.5.2. Обучение без учителя	467
11.5.3. Обучение с частичным привлечением учителя	469
11.5.4. Проектирование признаков	470

11.6. Глубокое обучение	472
11.6.1. Сети и глубина	474
11.6.2. Обратное распространение	478
11.6.3. Векторное представление слов и графов	479
11.7. Случай из жизни: игра имен	482
11.8. Дополнительная информация	485
11.9. Упражнения	485
<b>Глава 12. Большие данные: достижение крупного масштаба</b>	<b>489</b>
12.1. Что такое большие данные?	490
12.1.1. Большие данные — плохие данные	491
12.1.2. Три V	493
12.2. Случай из жизни: вопросы инфраструктуры	494
12.3. Алгоритмы для больших данных	496
12.3.1. Анализ большого O	497
12.3.2. Хеширование	499
12.3.3. Использование иерархии хранилищ	501
12.3.4. Поточковые и однопроходные алгоритмы	503
12.4. Фильтрация и выборка	505
12.4.1. Детерминированные алгоритмы выборки	506
12.4.2. Случайная и потоковая выборка	507
12.5. Параллелизм	508
12.5.1. Один, два, много	509
12.5.2. Параллелизм данных	511
12.5.3. Сеточный поиск	512
12.5.4. Службы облачных вычислений	513
12.6. MapReduce	513
12.6.1. Программирование MapReduce	515
12.6.2. MapReduce под капотом	517
12.7. Социальные и этические последствия	519
12.8. Дополнительная информация	523
12.9. Упражнения	524
<b>Глава 13. Заключение</b>	<b>527</b>
13.1. Получить работу!	527
13.2. Пойти в аспирантуру!	528
13.3. Профессиональные консалтинговые услуги	529
<b>Глава 14. Список литературы</b>	<b>531</b>
<b>Предметный указатель</b>	<b>539</b>