
Оглавление

Об авторах.....	13
Предисловие.....	15
Чего ожидать.....	15
Условные обозначения, принятые в книге.....	15
Использование примеров кода.....	16
Благодарности.....	16
Комментарий переводчика.....	17
Глава 1. Разведочный анализ данных.....	19
Элементы структурированных данных.....	20
Дополнительные материалы для чтения.....	22
Прямоугольные данные.....	23
Кадры данных и индексы.....	24
Непрямоугольные структуры данных.....	25
Дополнительные материалы для чтения.....	26
Оценки центрального положения.....	26
Среднее.....	27
Медиана и робастные оценки.....	28
Выбросы.....	29
Пример: оценки центрального положения численности населения и уровня убийств.....	30
Дополнительные материалы для чтения.....	31
Оценки вариабельности.....	31
Стандартное отклонение и связанные с ним оценки.....	33
Оценки на основе процентилей.....	35
Пример: оценки вариабельности населения штатов.....	36
Дополнительные материалы для чтения.....	37
Обследование распределения данных.....	37
Процентили и коробчатые диаграммы.....	38
Частотная таблица и гистограммы.....	39
Оценки плотности.....	41
Дополнительные материалы для чтения.....	43
Обследование двичных и категориальных данных.....	43
Мода.....	45
Математическое ожидание.....	45
Дополнительные материалы для чтения.....	46
Корреляция.....	46
Диаграммы рассеяния.....	49
Дополнительные материалы для чтения.....	50

Исследование двух или более переменных.....	51
Шестиугольная сетка и контуры (отображение числовых данных против числовых).....	51
Две категориальных переменных.....	54
Категориальные и числовые данные.....	55
Визуализация многочисленных переменных.....	56
Дополнительные материалы для чтения.....	58
Решите.....	58
Глава 2. Распределение данных и выборки.....	59
Случайный отбор и смещенная выборка.....	60
Смещение.....	62
Пространственный выбор.....	63
Размер против качества: когда размер имеет значение?.....	64
Выборочное среднее против популяционного среднего.....	65
Дополнительные материалы для чтения.....	66
Систематическая ошибка отбора.....	66
Регрессия к среднему.....	67
Дополнительные материалы для чтения.....	69
Выборочное распределение статистики.....	69
Центральная предельная теорема.....	72
Стандартная ошибка.....	72
Дополнительные материалы для чтения.....	73
Бутстрап.....	74
Повторный отбор против бутстрапирования.....	77
Дополнительные материалы для чтения.....	77
Доверительные интервалы.....	77
Дополнительные материалы для чтения.....	80
Нормальное распределение.....	80
Стандартное нормальное распределение и квантили, квантильные графики.....	82
Длиннохвостые распределения.....	84
Дополнительные материалы для чтения.....	85
<i>t</i> -Распределение Стьюдента.....	86
Дополнительные материалы для чтения.....	88
Биномиальное распределение.....	88
Дополнительные материалы для чтения.....	90
Распределение Пуассона и другие с ним связанные распределения.....	90
Распределение Пуассона.....	91
Экспоненциальное распределение.....	92
Оценка интенсивности отказов.....	92
Распределение Вейбулла.....	93
Дополнительные материалы для чтения.....	94
Решите.....	94
Глава 3. Статистические эксперименты и проверка значимости.....	95
<i>A/B</i> -тестирование.....	95
Зачем нужна контрольная группа?.....	98
Почему только <i>A/B</i> ? Почему не <i>C, D</i> ?.....	99
Дополнительные материалы для чтения.....	100
Проверка статистических гипотез.....	100
Нулевая гипотеза.....	102

Альтернативная гипотеза.....	102
Односторонняя и двухсторонняя проверки гипотез.....	103
Дополнительные материалы для чтения.....	104
Повторный отбор.....	104
Перестановочный тест.....	105
Пример: признатливость веб-страниц.....	105
Исчерпывающий и бутстрэповский перестановочные тесты.....	108
Перестановочные тесты: сухой остаток для науки о данных.....	109
Дополнительные материалы для чтения.....	109
Статистическая значимость и <i>p</i>-значения.....	110
<i>p</i> -Значение.....	112
Альфа.....	112
Чему равно <i>p</i> -значение?.....	113
Ошибки 1-го и 2-го рода.....	114
Наука о данных и <i>p</i> -значения.....	114
Дополнительные материалы для чтения.....	115
Проверка на основе <i>t</i>-статистики.....	115
Дополнительные материалы для чтения.....	117
Множественное тестирование.....	117
Дополнительные материалы для чтения.....	121
Степени свободы.....	121
Дополнительные материалы для чтения.....	122
ANOVA.....	123
<i>F</i> -статистика.....	126
Двухсторонняя процедура ANOVA.....	127
Дополнительные материалы для чтения.....	127
Проверка на основе статистики хи-квадрат.....	128
Проверка χ^2 : подход на основе повторного отбора.....	128
Проверка χ^2 : статистическая теория.....	130
Точная проверка Фишера.....	131
Актуальность проверок для науки о данных.....	133
Дополнительные материалы для чтения.....	134
Алгоритм многоаружного бандита.....	134
Дополнительные материалы для чтения.....	137
Мощность и размер выборки.....	138
Размер выборки.....	140
Дополнительные материалы для чтения.....	141
Решома.....	142
Глава 4. Регрессия и предсказание.....	143
Простая линейная регрессия.....	143
Уравнение регрессии.....	144
Подогнанные значения и остатки.....	146
Наименьшие квадраты.....	148
Предсказание против объяснения (профессирование).....	149
Дополнительные материалы для чтения.....	150
Множественная линейная регрессия.....	150
Пример: данные о жилом фонде округа Кинг.....	151
Диагностика модели.....	152

Перекрестная проверка	154
Отбор модели и шагговая регрессия	155
Внешенная регрессия	157
Предсказание на основе регрессии	158
Опасности экстраполяции	159
Доверительный и предсказательный интервалы	159
Факторные переменные в регрессии	161
Представление фиктивных переменных	162
Многоуровневые факторные переменные	164
Порядковые факторные переменные	165
Интерпретация уравнения регрессии	166
Коррелированные предикторы	167
Мультиколлинеарность	168
Искажающие переменные	169
Взаимодействия и главные эффекты	170
Проверка допущений: диагностика регрессии	172
Выбросы	173
Влиятельные значения	174
Гетероскедастичность, ненормальность и коррелированные ошибки	177
Графики частных остатков и нелинейность	179
Нелинейная регрессия	181
Параболическая регрессия	182
Сплайновая регрессия	183
Обобщенные аддитивные модели	185
Дополнительные материалы для чтения	187
Решение	187
Глава 5. Классификация	189
Важный байесовский алгоритм	190
Почему точная байесовская классификация непрактична?	191
Важное решение	192
Числовые предикторные переменные	194
Дополнительные материалы для чтения	194
Дискриминантный анализ	195
Ковариационная матрица	196
Линейный дискриминант Фишера	196
Простой пример	197
Дополнительные материалы для чтения	199
Логистическая регрессия	199
Функция логистического отклика и логит-преобразование	200
Логистическая регрессия и обобщенная линейная модель	202
Обобщенные линейные модели	203
Предсказанные значения в логистической регрессии	203
Интерпретация коэффициентов и отношений шансов	204
Линейная и логистическая регрессии: сходства и различия	205
Подгонка модели	205
Диагностика модели	206
Дополнительные материалы для чтения	209
Оценивание моделей классификации	210
Матрица несоответствий	211

Проблема редкого класса	213
Прецизионность, полнота и специфичность	213
ROC-кривая	214
Метрический показатель AUC	216
Лифт	217
Дополнительные материалы для чтения	218
Стратегии в отношении несбалансированных данных	219
Понижающий отбор	220
Повышающий отбор и понижающая/повышающая перенесенка	220
Генерация данных	221
Стоимостно-ориентированная классификация	222
Обследование предсказаний	222
Дополнительные материалы для чтения	224
Резюме	224
Глава 6. Статистическое машинное обучение	225
 K ближайших соседей	226
Небольшой пример: предсказание неустарата ссуды	227
Метрические показатели расстояния	229
Кодировщик с одним активным состоянием	230
Стандартизация (нормализация, z-оценки)	231
Выбор K	233
Метод KNN как конструктор признаков	234
 Древовидные модели	235
Простой пример	237
Алгоритм рекурсивного сегментирования	238
Измерение однородности или разнородности	240
Остановка роста дерева	241
Предсказывание непрерывной величины	243
Каким образом деревья используются	243
Дополнительные материалы для чтения	244
 Бэггинг и случайный лес	244
Бэггинг	246
Случайный лес	246
Важность переменных	249
Гиперпараметры	251
 Бустинг	252
Алгоритм бустинга	253
XGBoost	254
Регуляризация: предотвращение переобучения	256
Гиперпараметры и перекрестная проверка	259
 Резюме	261
Глава 7. Обучение без учителя	263
 Анализ главных компонент	264
Простой пример	265
Вычисление главных компонент	267
Интерпретация главных компонент	267
Дополнительные материалы для чтения	270

Кластеризация на основе K средних	270
Простой пример	271
Алгоритм K средних	272
Интерпретация кластеров	273
Выбор количества кластеров	275
Иерархическая кластеризация	277
Простой пример	277
Дендограмма	278
Агломеративный алгоритм	279
Меры расстояния	280
Модельно-ориентированная кластеризация	281
Многомерное нормальное распределение	282
Смеси нормальных распределений	283
Выбор количества кластеров	285
Дополнительные материалы для чтения	287
Шкалирование и категориальные переменные	287
Шкалирование переменных	288
Доминантные переменные	289
Категориальные данные и расстояние Говера	290
Проблемы кластеризации смешанных данных	293
Резюме	294
Библиография	295
Предметный указатель	297