
Содержание

Предисловие	12
Структура книги	13
Как использовать книгу	14
Интернет-ресурсы	17
Благодарности	17
Обозначения	19
Условные обозначения	19
Глава 1. Методы машинного обучения для аналитического прогнозирования	23
1.1. Что такое аналитическое прогнозирование	23
1.2. Что такое машинное обучение	25
1.3. Как работает машинное обучение	29
1.4. Что плохого может произойти при машинном обучении	35
1.5. Жизненный цикл проекта по аналитическому прогнозированию: CRISP-DM	37
1.6. Инструменты для аналитического прогнозирования	40
1.7. Дальнейшие перспективы	41
1.8. Упражнения	43
Глава 2. Данные — выводы — решения	45
2.1. Преобразование бизнес-проблем в аналитические решения	45
2.1.1. Пример: мошенничество с автострахованием	47
2.2. Оценка осуществимости	48
2.1.1. Пример: мошенничество с автострахованием	50
2.3. Разработка базовой аналитической таблицы	52
2.3.1. Пример: мошенничество с автострахованием	56
2.4. Проектирование и реализация	57
2.4.1. Различные типы данных	59
2.4.2. Различные типы признаков	60
2.4.3. Время обработки	63
2.4.4. Юридические вопросы	66
2.4.5. Реализация признаков	69
2.4.6. Пример: мошенничество с автострахованием	69
2.5. Резюме	74
2.6. Дальнейшее чтение	76
2.7. Упражнения	78
Глава 3. Изучение данных	81
3.1. Отчет о качестве данных	82
3.1.1. Пример: мошенничество с автострахованием	83

3.2. Ознакомление с данными	88
3.2.1. Нормальное распределение	91
3.2.2. Пример: мошенничество с автострахованием	93
3.3. Определение проблем, связанных с качеством данных	93
3.3.1. Пропущенные значения	94
3.3.2. Нерегулярная мощность	95
3.3.3. Выбросы	96
3.3.4. Пример: мошенничество с автострахованием	98
3.4. Решение проблем, связанных с качеством данных	101
3.4.1. Обработка пропущенных значений	101
3.4.2. Обработка выбросов	103
3.5. Углубленное исследование данных	105
3.5.1. Визуализация отношений между признаками	105
3.5.2. Вычисление ковариации и корреляции	115
3.6. Подготовка данных	122
3.6.1. Нормализация	122
3.6.2. Статистическое группирование	124
3.6.3. Выборочные методы	128
3.7. Резюме	131
3.8. Дальнейшее чтение	132
3.9. Упражнения	133
Глава 4. Информационное обучение	149
4.1. Основная идея	149
4.2. Основы	152
4.2.1. Деревья решений	153
4.2.2. Модель энтропии Шеннона	156
4.2.3. Прирост информации	160
4.3. Стандартный подход: алгоритм ID3	166
4.3.1. Реальный пример: прогнозирование распределения растительности	170
4.4. Обобщения и варианты	178
4.4.1. Альтернативный выбор признаков и показатели неоднородности	179
4.4.2. Обработка непрерывных описательных признаков	184
4.4.3. Прогнозирование непрерывных целевых признаков	188
4.4.4. Усечение деревьев	194
4.4.5. Ансамбли моделей	199
4.5. Резюме	204
4.6. Дальнейшее чтение	206
4.6. Упражнения	207
Глава 5. Обучение на основе сходства	217
5.1. Основная идея	217
5.2. Основы	218
5.2.1. Пространство признаков	219
5.2.2. Измерение сходства с помощью расстояния	221

5.3. Стандартный подход: алгоритм ближайшего соседа	224
5.3.1. Реальный пример	225
5.4. Обобщения и варианты	229
5.4.1. Обработка зашумленных данных	229
5.4.2. Поиск, эффективный с точки зрения памяти	234
5.4.3. Нормализация данных	244
5.4.4. Прогнозирование непрерывных целевых признаков	250
5.4.5. Другие меры сходства	254
5.4.6. Выбор признаков	268
5.5. Резюме	278
5.6. Дальнейшее чтение	281
5.7. Эпилог	282
5.8. Упражнения	283
Глава 6. Вероятностное обучение	291
6.1. Основная идея	291
6.2. Основы	294
6.2.1. Теорема Байеса	297
6.2.2. Байесовское прогнозирование	300
6.2.3. Условная независимость и факторизация	307
6.3. Стандартный подход: наивная модель Байеса	312
6.3.1. Практический пример	314
6.4. Обобщения и варианты	318
6.4.1. Сглаживание	318
6.4.2. Непрерывные признаки: функции плотности вероятности	323
6.4.3. Непрерывные характеристики: группирование	337
6.4.4. Байесовские сети	341
6.4.4.1. Построение байесовских сетей	347
6.4.4.2. Использование байесовской сети для прогнозирования	354
6.4.4.3. Прогнозирование с отсутствующими описательными значениями признаков	355
6.5. Резюме	361
6.6. Дальнейшее чтение	364
6.7. Упражнения	365
Глава 7. Обучение на основе ошибок	371
7.1. Основная идея	371
7.2. Основы	372
7.2.1. Простая линейная регрессия	372
7.2.2. Измерение ошибки	375
7.2.3. Поверхности ошибок	379
7.3. Стандартный подход: множественная линейная регрессия с градиентным спуском	381
7.3.1. Множественная линейная регрессия	381

7.3.2. Градиентный спуск	383
7.3.3. Выбор скорости обучения и начальных весов	390
7.3.4. Практический пример	393
7.4. Обобщения и варианты	396
7.4.1. Интерпретация моделей множественной линейной регрессии	397
7.4.2. Определение скорости обучения с использованием сокращения весов	399
7.4.3. Обработка категориальных описательных признаков	401
7.4.4. Обработка категориальных целевых признаков: логистическая регрессия	404
7.4.5. Моделирование нелинейных зависимостей	418
7.4.6. Мультиномиальная логистическая регрессия	425
7.4.7. Метод опорных векторов	429
7.5. Резюме	435
7.6. Дальнейшее чтение	439
7.7. Упражнения	439
Глава 8. Оценивание	449
8.1. Основная идея	449
8.2. Основы	451
8.3. Стандартный подход: доля ошибок классификации на тестовом множестве	452
8.4. Обобщения и варианты	457
8.4.1. Проектирование оценочных экспериментов	458
8.4.2. Показатели эффективности: категориальные целевые признаки	467
8.4.3. Показатели эффективности: оценки прогноза	477
8.4.3.2. Статистика Колмогорова–Смирнова	486
8.4.4. Показатели эффективности: мультиномиальные цели	495
8.4.5. Показатели эффективности: непрерывные целевые признаки	498
8.4.6. Оценка моделей после развертывания	503
8.5. Резюме	512
8.6. Дальнейшее чтение	513
8.7. Упражнения	514
Глава 9. Тематический пример: отток клиентов	521
9.1. Понимание бизнеса	521
9.2. Понимание данных	525
9.3. Подготовка данных	530
9.4. Моделирование	538
9.5. Оценка	540
9.6. Развертывание	543
Глава 10. пример: классификация галактик	545
10.1. Понимание бизнеса	546
10.1.1. Свободное владение ситуацией	548
10.2. Понимание данных	550

10.3. Подготовка данных	559
10.4. Моделирование	564
10.4.1 Базовые модели	564
10.4.2. Выбор признака	568
10.4.3. Пятиуровневая модель	569
10.5. Оценка	573
10.6. Развертывание	574
Глава 11. Искусство машинного обучения для аналитического прогнозирования	577
11.1. Различные перспективы для моделей прогнозирования	579
11.2. Выбор метода машинного обучения	585
11.2.1. Согласование подходов к машинному обучению	588
11.2.2. Согласование метода машинного обучения и данных	590
11.3. Следующие шаги	591
Приложение А. Описательная статистика и визуализация данных для машинного обучения	593
A.1. Описательная статистика для непрерывных признаков	593
A.1.1. Среднее значение	593
A.1.2. Разброс	595
A.2. Описательные статистики для категориальных признаков	598
A.3. Генеральные совокупности и выборки	600
A.4. Визуализация данных	602
A.4.1. Столбчатые диаграммы	603
A.4.2. Гистограммы	604
A.4.3. Блочные диаграммы	607
Приложение Б. Введение в теорию вероятностей	609
Б.1. Основы теории вероятностей	609
Б.2. Распределение вероятностей и суммирование	614
Б.3. Некоторые полезные правила вероятностей	616
Б.4. Сводка	618
Приложение В. Правила дифференцирования	619
В.1. Производные непрерывных функций	620
В.2. Правило дифференцирования сложных функций	623
В.3. Частные производные	623
Библиография	
Список рисунков	631
Список таблиц	645
Предметный указатель	652