

# Содержание

Об авторах	15
<b>Предисловие</b>	18
<b>Глава 1. Наделение компьютеров способностью обучения на данных</b>	27
Построение интеллектуальных машин для трансформирования данных в знания	28
Три типа машинного обучения	28
Выработка прогнозов о будущем с помощью обучения с учителем	28
Решение интерактивных задач с помощью обучения с подкреплением	33
Обнаружение скрытых структур с помощью обучения без учителя	34
Введение в основную терминологию и обозначения	36
Дорожная карта для построения систем машинного обучения	39
Предварительная обработка – приведение данных в приемлемую форму	40
Обучение и выбор прогнозирующей модели	41
Оценка моделей и прогнозирование на не встречавшихся ранее образцах данных	42
Использование Python для машинного обучения	42
Установка Python и необходимых пакетов	43
Использование дистрибутива Anaconda и диспетчера пакетов	43
Пакеты для научных вычислений, науки о данных и машинного обучения	44
Резюме	44
<b>Глава 2. Обучение простых алгоритмов МО для классификации</b>	47
Искусственные нейроны – беглое знакомство с ранней историей машинного обучения	48
Формальное определение искусственного нейрона	49
Правило обучения перцептрона	51
Реализация алгоритма обучения перцептрона на Python	54
Объектно-ориентированный API-интерфейс перцептрона	54
Обучение модели перцептрона на наборе данных Iris	58
Адаптивные линейные нейроны и сходимость обучения	64
Минимизация функций издержек с помощью градиентного спуска	65
Реализация Adaline на Python	69
Улучшение градиентного спуска посредством масштабирования признаков	73
Крупномасштабное машинное обучение и стохастический градиентный спуск	75
Резюме	81

<b>Глава 3. Обзор классификаторов на основе машинного обучения с использованием scikit-learn</b>	83
Выбор алгоритма классификации	84
Первые шаги в освоении scikit-learn – обучение персептрона	85
Моделирование вероятностей классов посредством логистической регрессии	91
Понятие логистической регрессии и условные вероятности	92
Выяснение весов функции издержек для логистической регрессии	96
Преобразование реализации Adaline в алгоритм для логистической регрессии	99
Обучение логистической регрессионной модели с помощью scikit-learn	103
Решение проблемы переобучения с помощью регуляризации	106
Классификация с максимальным зазором с помощью методов опорных векторов	109
Понятие максимального зазора	110
Обработка нелинейно сепарабельного случая с использованием фиктивных переменных	111
Альтернативные реализации в scikit-learn	114
Решение нелинейных задач с применением ядерного метода опорных векторов	115
Ядерные методы для линейно сепарабельных данных	115
Использование ядерного трюка для нахождения разделяющих гиперплоскостей в пространстве высокой размерности	117
Обучение на основе деревьев принятия решений	121
Доведение до максимума прироста информации – получение наибольшей отдачи	122
Построение дерева принятия решений	128
Объединение множества деревьев принятия решений с помощью случайных лесов	131
Метод $k$ ближайших соседей – алгоритм ленивого обучения	135
Резюме	138
<b>Глава 4. Построение хороших обучающих наборов с помощью предварительной обработки данных</b>	141
Решение проблемы с недостающими данными	142
Идентификация недостающих значений в табличных данных	142
Исключение образцов или признаков с недостающими значениями	144
Условный расчет недостающих значений	145
Понятие API-интерфейса оценщиков scikit-learn	146
Обработка категориальных данных	147
Именные и порядковые признаки	147
Отображение порядковых признаков	148
Кодирование меток классов	149
Выполнение унитарного кодирования на именных признаках	151
Разбиение набора данных на отдельные обучающий и испытательный наборы	153

---

Приведение признаков к тому же самому масштабу	156
Выбор значимых признаков	159
Регуляризация L1 и L2 как штрафы за сложность модели	160
Геометрическая интерпретация регуляризации L2	160
Разреженные решения с регуляризацией L1	163
Алгоритмы последовательного выбора признаков	167
Оценка важности признаков с помощью случайных лесов	174
Резюме	177
<b>Глава 5. Сжатие данных с помощью понижения размерности</b>	179
Понижение размерности без учителя с помощью анализа главных компонент	180
Основные шаги при анализе главных компонент	180
Выделение главных компонент шаг за шагом	182
Полная и объясненная дисперсия	185
Трансформация признаков	187
Анализ главных компонент в scikit-learn	190
Сжатие данных с учителем посредством линейного дискриминантного анализа	194
Анализ главных компонент в сравнении с линейным дискриминантным анализом	194
Внутреннее устройство линейного дискриминантного анализа	196
Вычисление матриц рассеяния	197
Выбор линейных дискриминантов для нового подпространства признаков	199
Проецирование образцов в новое подпространство признаков	202
Реализация LDA в scikit-learn	203
Использование ядерного анализа главных компонент для нелинейных отображающих функций	205
Ядерные функции и ядерный трюк	206
Реализация ядерного анализа главных компонент на Python	211
Проецирование новых точек данных	218
Ядерный анализ главных компонент в scikit-learn	222
Резюме	224
<b>Глава 6. Изучение практического опыта оценки моделей и настройки гиперпараметров</b>	225
Модернизация рабочих потоков с помощью конвейеров	226
Загрузка набора данных Breast Cancer Wisconsin	226
Объединение преобразователей и оценщиков в конвейер	228
Использование перекрестной проверки по $k$ блокам для оценки эффективности модели	230
Метод перекрестной проверки с удержанием	231
Перекрестная проверка по $k$ блокам	232

---

Отладка алгоритмов с помощью кривых обучения и проверки	238
Диагностирование проблем со смещением и дисперсией с помощью кривых обучения	238
Решение проблем недообучения и переобучения с помощью кривых проверки	242
Точная настройка моделей машинного обучения с помощью решетчатого поиска	244
Настройка гиперпараметров с помощью решетчатого поиска	244
Отбор алгоритма с помощью вложенной перекрестной проверки	246
Использование других метрик оценки эффективности	249
Чтение матрицы неточностей	249
Оптимизация точности и полноты классификационной модели	251
Построение кривой рабочей характеристики приемника	253
Метрики подсчета для многоклассовой классификации	256
Решение проблемы с дисбалансом классов	257
Резюме	260
<b>Глава 7. Объединение разных моделей для ансамблевого обучения</b>	<b>263</b>
Обучение с помощью ансамблей	263
Объединение классификаторов с помощью мажоритарного голосования	268
Реализация простого классификатора с мажоритарным голосованием	269
Использование принципа мажоритарного голосования для выработки прогнозов	275
Оценка и настройка ансамблевого классификатора	278
Бэггинг – построение ансамбля классификаторов из бутстрэп-образцов	284
Коротко о бэггинге	285
Применение бэггинга для классификации образцов в наборе данных Wine	287
Использование в своих интересах слабых учеников посредством адаптивного бустинга	291
Как работает бустинг	292
Применение алгоритма AdaBoost с помощью scikit-learn	297
Резюме	300
<b>Глава 8. Применение машинного обучения для смыслового анализа</b>	<b>301</b>
Подготовка данных с рецензиями на фильмы IMDb для обработки текста	302
Получение набора данных с рецензиями на фильмы	302
Предварительная обработка набора данных с целью приведения в более удобный формат	303
Модель суммирования слов	305
Трансформирование слов в векторы признаков	305
Оценка важности слов с помощью приема tf-idf	308
Очистка текстовых данных	310
Переработка документов в лексемы	312
Обучение логистической регрессионной модели для классификации документов	315

Работа с более крупными данными – динамические алгоритмы и внешнее обучение	318
Тематическое моделирование с помощью латентного размещения Дирихле	322
Разбиение текстовых документов с помощью LDA	323
Реализация LDA в scikit-learn	324
Резюме	328
<b>Глава 9. Встраивание модели машинного обучения в веб-приложение</b>	331
Сериализация подогнанных оценщиков scikit-learn	332
Настройка базы данных SQLite для хранилища данных	335
Разработка веб-приложения с помощью Flask	338
Первое веб-приложение Flask	339
Проверка достоверности и визуализация форм	341
Превращение классификатора рецензий на фильмы в веб-приложение	346
Файлы и подкаталоги – дерево каталогов	348
Реализация главного приложения как <code>app.py</code>	348
Настройка формы для рецензии	351
Создание шаблона страницы результатов	352
Развертывание веб-приложения на публичном сервере	355
Создание учетной записи PythonAnywhere	355
Загрузка файлов для приложения классификации рецензий на фильмы	356
Обновление классификатора рецензий на фильмы	357
Резюме	359
<b>Глава 10. Прогнозирование значений непрерывных целевых переменных с помощью регрессионного анализа</b>	361
Ведение в линейную регрессию	362
Простая линейная регрессия	362
Множественная линейная регрессия	363
Исследование набора данных Housing	364
Загрузка набора данных Housing в объект DataFrame	365
Визуализация важных характеристик набора данных	367
Просмотр взаимосвязей с использованием корреляционной матрицы	369
Реализация линейной регрессионной модели с использованием обычного метода наименьших квадратов	372
Использование градиентного спуска для выяснения параметров регрессии	373
Оценка коэффициентов регрессионной модели с помощью scikit-learn	377
Подгонка надежной регрессионной модели с использованием RANSAC	379
Оценка эффективности линейных регрессионных моделей	382
Использование регуляризованных методов для регрессии	386
Превращение линейной регрессионной модели в криволинейную – полиномиальная регрессия	388

---

Добавление полиномиальных членов с использованием scikit-learn	388
Моделирование нелинейных связей в наборе данных Housing	390
Обработка нелинейных связей с использованием случайных лесов	393
Регрессия на основе дерева принятия решений	394
Регрессия на основе случайного леса	396
Резюме	400
<b>Глава 11. Работа с непомеченными данными – кластерный анализ</b>	<b>401</b>
Группирование объектов по подобию с применением метода k-средних	402
Кластеризация методом k-средних с использованием scikit-learn	402
Более интеллектуальный способ размещения начальных центроидов кластеров с использованием метода k-средних++	408
Жесткая или мягкая кластеризация	409
Использование метода локтя для нахождения оптимального количества кластеров	412
Количественная оценка качества кластеризации через графики силуэтов	413
Организация кластеров в виде иерархического дерева	418
Группирование кластеров в восходящей манере	419
Выполнение иерархической кластеризации на матрице расстояний	421
Прикрепление дендрограмм к тепловой карте	425
Применение агломеративной иерархической кластеризации с помощью scikit-learn	427
Нахождение областей высокой плотности с помощью DBSCAN	428
Резюме	434
<b>Глава 12. Реализация многослойной искусственной нейронной сети с нуля</b>	<b>437</b>
Моделирование сложных функций с помощью искусственных нейронных сетей	438
Краткое повторение однослойных нейронных сетей	440
Введение в архитектуру многослойных нейронных сетей	442
Активация нейронной сети посредством прямого распространения	445
Классификация рукописных цифр	448
Получение набора данных MNIST	449
Реализация многослойного персептрона	456
Обучение искусственной нейронной сети	467
Вычисление логистической функции издержек	467
Выработка общего понимания обратного распространения	470
Обучение нейронных сетей с помощью обратного распространения	472
О сходимости в нейронных сетях	476
Несколько заключительных замечаний о реализации нейронных сетей	477
Резюме	478

---

<b>Глава 13. Распараллеливание процесса обучения нейронных сетей с помощью TensorFlow</b>	479
TensorFlow и производительность обучения	480
Что такое TensorFlow?	482
Как мы будем изучать TensorFlow	483
Первые шаги с библиотекой TensorFlow	483
Работа с массивами	486
Разработка простой модели с помощью низкоуровневого API-интерфейса TensorFlow	488
Эффективное обучение нейронных сетей с помощью высокоуровневых API-интерфейсов TensorFlow	492
Построение многослойных нейронных сетей с использованием API-интерфейса Layers библиотеки TensorFlow	493
Разработка многослойной нейронной сети с помощью Keras	497
Выбор функций активации для многослойных сетей	503
Резюме по логистической функции	504
Оценка вероятностей классов в многоклассовой классификации через многопеременную логистическую функцию	506
Расширение выходного спектра с использованием гиперболического тангенса	507
Активация на основе выпрямленного линейного элемента	509
Резюме	511
<b>Глава 14. Погружаемся глубже – механика TensorFlow</b>	513
Ключевые средства TensorFlow	514
Ранги и тензоры TensorFlow	514
Получение ранга и формы тензора	515
Понятие вычислительных графов TensorFlow	516
Заполнители в TensorFlow	519
Определение заполнителей	519
Подача данных заполнителям	520
Определение заполнителей для массивов данных с варьирующимися размерами мини-пакетов	521
Переменные в TensorFlow	522
Определение переменных	522
Инициализация переменных	525
Область видимости переменной	526
Повторное использование переменных	528
Построение регрессионной модели	531
Выполнение объектов в графе TensorFlow с использованием их имен	535
Сохранение и восстановление модели в TensorFlow	536
Преобразование тензоров как многомерных массивов данных	540

---

Использование механики управления потоком при построении графов	543
Визуализация графа с помощью TensorBoard	547
Расширение навыков работы с TensorBoard	550
Резюме	551
<b>Глава 15. Классификация изображений с помощью глубоких сверточных нейронных сетей</b>	553
Строительные блоки сверточных нейронных сетей	554
Понятие сетей CNN и выявление иерархий признаков	554
Выполнение дискретных сверток	556
Подвыборка	566
Группирование всего вместе для построения сверточной нейронной сети	568
Работа с множественными входными или цветовыми каналами	568
Регуляризация нейронной сети с помощью отключения	572
Реализация глубокой сверточной нейронной сети с применением TensorFlow	574
Архитектура многослойной сверточной нейронной сети	575
Загрузка и предварительная обработка данных	576
Реализация сверточной нейронной сети с помощью низкоуровневого API-интерфейса TensorFlow	577
Реализация сверточной нейронной сети с помощью API-интерфейса Layers из TensorFlow	590
Резюме	596
<b>Глава 16. Моделирование последовательных данных с использованием рекуррентных нейронных сетей</b>	597
Понятие последовательных данных	598
Моделирование последовательных данных – вопросы порядка	598
Представление последовательностей	599
Категории моделирования последовательностей	600
Рекуррентные нейронные сети для моделирования последовательностей	601
Структура и поток данных рекуррентной нейронной сети	601
Вычисление активаций в сети RNN	604
Сложности изучения долгосрочных взаимодействий	607
Элементы LSTM	608
Реализация многослойной рекуррентной нейронной сети для моделирования последовательностей в TensorFlow	611
Проект 1 – выполнение смыслового анализа рецензий на фильмы IMDb с использованием многослойной рекуррентной нейронной сети	611
Подготовка данных	612
Векторное представление	616
Построение модели на основе рекуррентной нейронной сети	619
Конструктор класса <code>SentimentRNN</code>	619



Метод <code>build</code>	620
Метод <code>train</code>	624
Метод <code>predict</code>	625
Создание объекта класса <code>SentimentRNN</code>	626
Обучение и оптимизация модели на основе рекуррентной нейронной сети для смыслового анализа	627
Проект два – реализация рекуррентной нейронной сети для моделирования языка на уровне символов в TensorFlow	628
Подготовка данных	629
Построение символьной модели языка на основе рекуррентной нейронной сети	633
Конструктор класса <code>CharRNN</code>	633
Метод <code>build</code>	635
Метод <code>train</code>	637
Метод <code>sample</code>	638
Создание и обучение модели <code>CharRNN</code>	640
Модель <code>CharRNN</code> в режиме выборки	640
Резюме по главе и по книг	641
<b>Предметный указатель</b>	644