

<b>Предисловие</b> .....	11
<b>Об авторе</b> .....	12
<b>О рецензентах</b> .....	13
<b>Введение</b> .....	15
<b>Глава 1. Наделение компьютеров способностью обучаться на данных</b> .....	25
Построение интеллектуальных машин для преобразования данных в знания .....	25
Три типа машинного обучения.....	26
Выполнение прогнозов о будущем на основе обучения с учителем.....	26
Задача классификации – распознавание меток классов .....	27
Задача регрессии – предсказание непрерывных результатов.....	28
Решение интерактивных задач на основе обучения с подкреплением .....	29
Обнаружение скрытых структур при помощи обучения без учителя .....	30
Выявление подгрупп при помощи кластеризации .....	30
Снижение размерности для сжатия данных .....	31
Введение в основополагающую терминологию и систему обозначений .....	32
Дорожная карта для построения систем машинного обучения.....	33
Предобработка – приведение данных в приемлемый вид.....	34
Тренировка и отбор прогнозной модели.....	35
Оценка моделей и прогнозирование на ранее не встречавшихся экземплярах данных.....	36
Использование Python для машинного обучения.....	36
Установка библиотек Python .....	37
Записные книжки Jupyter/IPython.....	38
Резюме .....	40
<b>Глава 2. Тренировка алгоритмов машинного обучения для задачи классификации</b> .....	42
Искусственные нейроны – краткий обзор ранней истории машинного обучения.....	42
Реализация алгоритма обучения персептрона на Python.....	48
Тренировка персептронной модели на наборе данных цветков ириса .....	50
Адаптивные линейные нейроны и сходимость обучения.....	54
Минимизация функций стоимости методом градиентного спуска .....	55
Реализация адаптивного линейного нейрона на Python .....	57
Крупномасштабное машинное обучение и стохастический градиентный спуск .....	62
Резюме .....	67
<b>Глава 3. Обзор классификаторов с использованием библиотеки scikit-learn</b> .....	68
Выбор алгоритма классификации.....	68

Первые шаги в работе с scikit-learn.....	69
Тренировка перцептрона в scikit-learn.....	69
Моделирование вероятностей классов логистической регрессии.....	73
Интуитивное понимание логистической регрессии и условные вероятности.....	74
Извлечение весов логистической функции стоимости.....	77
Тренировка логистической регрессионной модели в scikit-learn .....	79
Решение проблемы переподргонки при помощи регуляризации.....	81
Классификация с максимальной маржой на основе машин опорных векторов.....	84
Интуитивное понимание максимальной маржи.....	85
Обработка нелинейно разделимого случая при помощи ослабленных переменных .....	86
Альтернативные реализации в scikit-learn .....	88
Решение нелинейных задач ядерным методом SVM .....	88
Использование ядерного трюка для нахождения разделяющих гиперплоскостей в пространстве более высокой размерности .....	90
Обучение на основе деревьев решений .....	93
Максимизация прироста информации – получение наибольшей отдачи .....	94
Построение дерева решений .....	98
Объединение слабых учеников для создания сильного при помощи случайных лесов.....	100
$k$ ближайших соседей – алгоритм ленивого обучения.....	103
Резюме .....	106

## **Глава 4. Создание хороших тренировочных наборов – предобработка данных .....**

Решение проблемы пропущенных данных .....	107
Устранение образцов либо признаков с пропущенными значениями .....	109
Импутация пропущенных значений.....	110
Концепция взаимодействия с эстиматорами в библиотеке scikit-learn .....	110
Обработка категориальных данных .....	112
Преобразование порядковых признаков.....	112
Кодирование меток классов.....	113
Прямое кодирование на номинальных признаках.....	114
Разбивка набора данных на тренировочное и тестовое подмножества .....	116
Приведение признаков к одинаковой шкале.....	117
Отбор содержательных признаков .....	119
Разреженные решения при помощи L1-регуляризации .....	119
Алгоритмы последовательного отбора признаков.....	125
Определение важности признаков при помощи случайных лесов .....	130
Резюме .....	132

## **Глава 5. Сжатие данных путем снижения размерности.....**

Снижение размерности без учителя на основе анализа главных компонент .....	133
Общая и объясненная дисперсия.....	135
Преобразование признаков .....	138
Анализ главных компонент в scikit-learn .....	140
Сжатие данных с учителем путем линейного дискриминантного анализа.....	143

Вычисление матриц разброса.....	145
Отбор линейных дискриминантов для нового подпространства признаков.....	147
Проецирование образцов на новое пространство признаков.....	149
Метод LDA в scikit-learn .....	150
Использование ядерного метода анализа главных компонент для нелинейных отображений.....	151
Ядерные функции и ядерный трюк.....	152
Реализация ядерного метода анализа главных компонент на Python .....	156
Пример 1. Разделение фигур в форме полумесяца.....	157
Пример 2. Разделение концентрических кругов .....	159
Проецирование новых точек данных.....	162
Ядерный метод анализа главных компонент в scikit-learn.....	165
Резюме .....	166

## **Глава 6. Изучение наиболее успешных методов оценки моделей и тонкой настройки гиперпараметров .....**

<b>167</b>	<b>167</b>
Оптимизация потоков операций при помощи конвейеров.....	167
Загрузка набора данных Breast Cancer Wisconsin.....	167
Совмещение преобразователей и эстиматоров в конвейере.....	169
Использование $k$ -блочной перекрестной проверки для оценки работоспособности модели.....	170
Метод проверки с откладыванием данных .....	171
$k$ -блочная перекрестная проверка.....	172
Отладка алгоритмов при помощи кривой обучения и проверочной кривой.....	176
Диагностирование проблем со смещением и дисперсией при помощи кривых обучения.....	176
Решение проблемы перепогонки и недопогонки при помощи проверочных кривых .....	179
Тонкая настройка машинно-обучаемых моделей методом сеточного поиска .....	181
Настройка гиперпараметров методом поиска по сетке параметров .....	181
Отбор алгоритмов методом вложенной перекрестной проверки .....	183
Обзор других метрик оценки работоспособности.....	184
Прочтение матрицы несоответствий .....	185
Оптимизация точности и полноты классификационной модели .....	186
Построение графика характеристической кривой.....	188
Оценочные метрики для многоклассовой классификации .....	191
Резюме .....	192

## **Глава 7. Объединение моделей для методов ансамблевого обучения .....**

<b>193</b>	<b>193</b>
Обучение при помощи ансамблей.....	193
Реализация простого классификатора с мажоритарным голосованием .....	197
Объединение разных алгоритмов классификации методом мажоритарного голосования.....	202
Оценка и тонкая настройка ансамблевого классификатора .....	205
Бэггинг – сборка ансамбля классификаторов из бутстрап-выборок.....	210
Усиление слабых учеников методом адаптивного бустинга .....	214
Резюме .....	221

<b>Глава 8. Применение алгоритмов машинного обучения в анализе мнений</b> .....	222
Получение набора данных киноотзывов IMDb .....	222
Концепция модели мешка слов.....	224
Преобразование слов в векторы признаков .....	225
Оценка релевантности слова методом tf-idf.....	226
Очистка текстовых данных .....	228
Переработка документов в лексемы.....	229
Тренировка логистической регрессионной модели для задачи классификации документов .....	232
Работа с более крупными данными – динамические алгоритмы и обучение вне ядра.....	234
Резюме .....	237
<b>Глава 9. Встраивание алгоритма машинного обучения в веб-приложение</b> .....	239
Сериализация подогнанных эстиматоров библиотеки scikit-learn .....	239
Настройка базы данных SQLite для хранения данных .....	242
Разработка веб-приложения в веб-платформе Flask.....	244
Наше первое веб-приложение Flask.....	245
Валидация и отображение формы .....	246
Превращение классификатора кинофильмов в веб-приложение.....	249
Развертывание веб-приложения на публичном сервере.....	256
Обновление классификатора киноотзывов.....	258
Резюме .....	259
<b>Глава 10. Прогнозирование непрерывных целевых величин на основе регрессионного анализа</b> .....	260
Введение в простую линейную регрессионную модель .....	260
Разведочный анализ набора данных Housing .....	261
Визуализация важных характеристик набора данных .....	263
Реализация линейной регрессионной модели обычным методом наименьших квадратов .....	266
Решение уравнения регрессии для параметров регрессии методом градиентного спуска .....	267
Оценивание коэффициента регрессионной модели в scikit-learn .....	270
Подгонка стабильной регрессионной модели алгоритмом RANSAC.....	272
Оценивание работоспособности линейных регрессионных моделей .....	274
Применение регуляризованных методов для регрессии.....	277
Превращение линейной регрессионной модели в криволинейную – полиномиальная регрессия .....	278
Моделирование нелинейных связей в наборе данных Housing.....	280
Обработка нелинейных связей при помощи случайных лесов .....	283
Регрессия на основе дерева решений.....	283
Регрессия на основе случайного леса.....	285
Резюме .....	287

<b>Глава 11. Работа с немаркированными данными – кластерный анализ</b> .....	289
Группирование объектов по подобию методом $k$ средних .....	289
Алгоритм $k$ -средних++ .....	292
Жесткая кластеризация в сопоставлении с мягкой .....	294
Использование метода локтя для нахождения оптимального числа кластеров .....	296
Количественная оценка качества кластеризации методом силуэтных графиков .....	298
Организация кластеров в виде иерархического дерева .....	302
Выполнение иерархической кластеризации на матрице расстояний .....	303
Прикрепление дендограмм к теплокарте .....	307
Применение агломеративной кластеризации в scikit-learn .....	308
Локализация областей высокой плотности алгоритмом DBSCAN .....	309
Резюме .....	313
<b>Глава 12. Тренировка искусственных нейронных сетей для распознавания изображений</b> .....	315
Моделирование сложных функций искусственными нейронными сетями .....	315
Краткое резюме однослойных нейронных сетей .....	317
Введение в многослойную нейросетевую архитектуру .....	318
Активация нейронной сети методом прямого распространения сигналов .....	320
Классификация рукописных цифр .....	322
Получение набора данных MNIST .....	323
Реализация многослойного персептрона .....	328
Тренировка искусственной нейронной сети .....	339
Вычисление логистической функции стоимости .....	339
Тренировка нейронных сетей методом обратного распространения ошибки .....	341
Развитие интуитивного понимания алгоритма обратного распространения ошибки .....	344
Отладка нейронных сетей процедурой проверки градиента .....	345
Сходимость в нейронных сетях .....	350
Другие нейросетевые архитектуры .....	351
Сверточные нейронные сети .....	352
Рекуррентные нейронные сети .....	354
Несколько последних замечаний по реализации нейронной сети .....	355
Резюме .....	355
<b>Глава 13. Распараллеливание тренировки нейронных сетей при помощи Theano</b> .....	356
Сборка, компиляция и выполнение выражений в Theano .....	356
Что такое Theano? .....	358
Первые шаги с библиотекой Theano .....	359
Конфигурирование библиотеки Theano .....	360
Работа с матричными структурами .....	362
Завершающий пример – линейная регрессия .....	364

Выбор функций активации для нейронных сетей с прямым распространением сигналов.....	367
Краткое резюме логистической функции .....	368
Оценивание вероятностей в многоклассовой классификации функцией softmax .....	370
Расширение выходного спектра при помощи гиперболического тангенса.....	371
Эффективная тренировка нейронных сетей при помощи библиотеки Keras .....	373
Резюме .....	378
<b>Приложение А</b> .....	<b>380</b>
Оценка моделей.....	380
Что такое перепогонка?.....	380
Как оценивать модель? .....	381
Сценарий 1. Элементарно обучить простую модель.....	381
Сценарий 2. Натренировать модель и выполнить тонкую настройку (оптимизировать гиперпараметры) .....	382
Сценарий 3. Построить разные модели и сравнить разные алгоритмы (например, SVM против логистической регрессии против случайных лесов и т. д.) .....	383
Перекрестная проверка. Оценка работоспособности эстиматора.....	384
Перекрестная проверка с исключением по одному .....	386
Пример стратифицированной $k$ -блочной перекрестной проверки.....	387
Расширенный пример вложенной перекрестной проверки.....	387
А. Вложенная кросс-валидация: быстрая версия.....	388
Б. Вложенная кросс-валидация: ручной подход с распечаткой модельных параметров .....	388
В. Регулярная $k$ -блочная кросс-валидация для оптимизации модели на полном наборе тренировочных данных .....	389
График проверочной (валидационной) кривой.....	389
Настройка типового конвейера и сеточного поиска.....	391
Машинное обучение .....	393
В чем разница между классификатором и моделью?.....	393
В чем разница между функцией стоимости и функцией потерь? .....	394
Обеспечение персистентности моделей scikit-learn на основе JSON.....	395
<b>Глоссарий основных терминов и сокращений</b> .....	<b>400</b>
Термины .....	400
Сокращения .....	406
<b>Предметный указатель</b> .....	<b>408</b>