

Содержание

Об авторе	17
Посвящение Луки	18
Посвящение Джона	18
Благодарности Луки	19
Благодарности Джона	19
Введение	20
О книге	20
Соглашения, принятые в книге	21
Глупые предположения	22
Источники дополнительной информации	23
Что дальше	24
Ждем ваших отзывов!	26
Часть 1. Приступая к работе с наукой о данных и языком Python	27
Глава 1. Взаимосвязь науки о данных с языком Python	29
Популярная профессия	32
Появление науки о данных	32
Основные компетенции аналитика данных	33
Связь между наукой о данных, большими данными и искусственным интеллектом	34
Роль программирования	34
Создание конвейера науки о данных	35
Подготовка данных	36
Предварительный анализ данных	36
Изучение данных	36
Визуализация	36
Осознание сути и смысла данных	37
Роль языка Python в науке о данных	37
Смещение профиля аналитика данных	37
Работа с многоцелевым, простым и эффективным языком	38
Быстро учимся использовать Python	39

Загрузка данных	40
Обучение модели	41
Просмотр результата	41
Глава 2. Возможности и чудеса языка Python	43
Почему Python?	44
Базовая философия языка Python	45
Вклад в науку о данных	46
Настоящие и будущие цели развития	47
Работа с языком Python	47
Знакомство с языком	48
Необходимость отступа	48
Работа в командной строке или в IDE	49
Создание новых сеансов с помощью командной строки Anaconda	50
Вход в среду IPython	52
Вход в среду Jupyter QtConsole	53
Редактирование сценариев с использованием Spyder	53
Быстрое создание прототипа и эксперименты	54
Скорость выполнения	56
Сила визуализации	58
Использование экосистемы Python для науки о данных	60
Доступ к научным инструментам с помощью SciPy	60
Фундаментальные научные вычисления с использованием NumPy	60
Анализа данных с использованием библиотеки pandas	61
Реализация машинного обучения с использованием Scikit-learn	61
Глубокое обучение с использованием Keras и TensorFlow	61
Графический вывод данных с использованием matplotlib	62
Создание графиков с помощью NetworkX	62
Анализ документов HTML с использованием BeautifulSoup	63
Глава 3. Конфигурация Python для науки о данных	65
Готовые кросс-платформенные научные дистрибутивы	66
Получение пакета Anaconda от Continuum Analytics	67
Получение продукта Enthought Canopy Express	68
Получение WinPython	68
Установка Anaconda на Windows	69
Установка Anaconda на Linux	74
Установка Anaconda на Mac OS X	75
Загрузка наборов данных и примеров кода	76

Использование Jupyter Notebook	76
Определение хранилища кода	78
Наборы данных, используемые в этой книге	85
Глава 4. Работа с Google Colab	87
Определение Google Colab	88
Что делает Google Colab	88
Особенности сетевого программирования	90
Поддержка локальной среды выполнения	91
Получение учетной записи Google	92
Создание учетной записи	92
Вход в систему	93
Работа с блокнотами	94
Создание нового блокнота	95
Открытие существующих блокнотов	95
Сохранение блокнотов	98
Загрузка блокнотов	101
Выполнение общих задач	101
Создание ячеек кода	102
Создание текстовых ячеек	104
Создание специальных ячеек	104
Редактирование ячеек	106
Движущиеся ячейки	106
Использование аппаратного ускорения	106
Выполнение кода	107
Просмотр блокнота	108
Отображение оглавления	108
Получение информации о блокноте	108
Проверка выполнения кода	109
Совместное использование блокнота	110
Получение помощи	112
Часть 2. Данные	115
Глава 5. Инструменты	117
Использование консоли Jupyter	118
Взаимодействие с текстом на экране	118
Изменение внешнего вида окна	121
Получение справки по Python	122

Получение справки по IPython	124
Использование магических функций	125
Работа с магическими функциями	125
Обнаружение объектов	126
Использование Jupyter Notebook	128
Работа со стилями	128
Перезапуск ядра	130
Восстановление контрольной точки	131
Интеграция мультимедиа и графики	131
Встраивание графиков и других изображений	132
Загрузка примеров с сайтов в Интернете	132
Получение сетевой графики и мультимедиа	132
Глава 6. Работа с реальными данными	135
Загрузка, потоковая передача и выборка данных	137
Загрузка небольших объемов данных в память	137
Загрузка в память большого количества данных	138
Генерация вариаций в данных изображения	139
Выборка данных разными способами	141
Доступ к данным в форме структурированного плоского файла	142
Чтение из текстового файла	143
Чтение формата CSV с разделителями	144
Чтение файлов Excel и других файлов Microsoft Office	146
Передача данных в форме неструктурированного файла	148
Работа с данными из реляционных баз данных	151
Взаимодействие с данными из баз NoSQL	153
Доступ к данным из Интернета	153
Глава 7. Подготовка данных	159
Баланс между NumPy и pandas	160
Когда использовать NumPy	160
Когда использовать pandas	161
Проверка данных	162
Выяснение содержимого данных	162
Удаление дубликатов	164
Создание карты и плана данных	165
Манипулирование категориальными переменными	167
Создание категориальных переменных	168

Переименование уровней	170
Объединение уровней	170
Работа с датами в данных	172
Форматирование значений даты и времени	172
Правильное преобразование времени	173
Борьба с отсутствием данных	174
Нахождение недостающих данных	174
Отсутствие в коде	175
Добавление недостающих данных	176
Разделение и дробление: фильтрация и выбор данных	177
Разделение строк	178
Разделение столбцов	178
Дробление	179
Конкатенация и преобразование	180
Добавление новых переменных и случаев	180
Удаление данных	182
Сортировка и перетасовка	183
Агрегирование данных на любом уровне	184
Глава 8. Формирование данных	187
Работа со страницами HTML	188
Анализ XML и HTML	188
Использование XPath для извлечения данных	189
Работа с необработанным текстом	191
Работа с Unicode	191
Морфологический поиск и удаление стоп-слов	192
Знакомство с регулярными выражениями	194
Использование модели наборов слов	197
Понятие модели “набор слов”	198
Работа с n-граммами	200
Реализация преобразований TF-IDF	201
Работа с данными графов	204
Понятие матрицы смежности	204
Использование основ NetworkX	205
Глава 9. Применение знаний на практике	207
Помещение в контекст задач и данных	208
Оценка задачи науки о данных	209
Исследовательские решения	212
Формулировка гипотезы	213

Подготовка данных	213
Искусство создания признаков	214
Определение создания признака	214
Объединение переменных	215
Понятие группирования и дискретизации	216
Использование индикаторных переменных	216
Преобразование распределений	217
Операции над массивами	218
Использование векторизации	218
Простые арифметические действия с векторами и матрицами	219
Матричное векторное умножение	219
Умножение матриц	220
Часть 3. Визуализация информации	221
Глава 10. Ускоренный курс по Matplotlib	223
Начнем с графика	224
Определение сюжета графика	225
Рисование нескольких линий и графиков	225
Сохранение работы на диске	226
Настройка осей, отметок, сеток	227
Получение осей	228
Форматирование осей	228
Добавление сетки	230
Определение внешнего вида линии	230
Работа со стилями линий	232
Использование цвета	232
Добавление маркеров	234
Использование меток, аннотаций и легенд	236
Добавление меток	237
Аннотирование диаграммы	237
Создание легенды	238
Глава 11. Визуализация данных	241
Выбор правильного графика	242
Демонстрация части целого на круговой диаграмме	242
Сравнение на гистограмме	243
Отображение распределений с использованием гистограмм	245
Обозначение групп с использованием диаграмм размаха	246

Просмотр шаблонов данных с использованием диаграмм рассеяния	249
Создание расширенных диаграмм рассеяния	249
Отображение групп	250
Отображение корреляций	251
Построение временных рядов	252
Представление времени по осям	253
Отображение трендов с течением времени	254
Отображение географических данных	257
Использование среды Notebook	258
Получение набора инструментов Basemap	259
Решение проблем устаревания библиотек	260
Использование Basemap для вывода географических данных	261
Визуализация графов	262
Разработка ненаправленных графов	264
Разработка направленных графов	266

Часть 4. Манипулирование данными 269

Глава 12. Расширение возможностей Python 271

Пакет Scikit-learn	272
Понятие классов в Scikit-learn	272
Определение приложений для науки о данных	274
Трюк хеширования	277
Использование хеш-функций	277
Демонстрация трюка хеширования	278
Детерминированный отбор	281
Учет сроков и производительности	283
Сравнительный анализ с использованием timeit	283
Работа с профилировщиком памяти	287
Параллельная работа на нескольких ядрах	289
Реализация многоядерного параллелизма	290
Демонстрация многопроцессорности	291

Глава 13. Разведочный анализ данных 293

Подход EDA	294
Определение описательной статистики для числовых данных	295
Измерение центральной тенденции	297
Измерение дисперсии и диапазона	297
Работа с процентиями	298

Определение мер нормальности	299
Подсчет для категориальных данных	301
Понятие частот	302
Создание таблиц сопряженности	303
Создание прикладной визуализации для EDA	304
Исследование диаграмм размаха	304
Поиск t-критериев после диаграмм размаха	306
Наблюдение параллельных координат	307
Графики распределения	307
Построение диаграмм рассеяния	308
Понятие корреляции	310
Использование ковариации и корреляции	310
Использование непараметрической корреляции	313
Учет критерия хи-квадрат для таблиц	313
Изменение распределения данных	314
Использование разных статистических распределений	315
Создание стандартизации z-оценки	315
Преобразование других известных распределений	316
Глава 14. Уменьшение размерности	317
Понятие SVD	318
В поисках уменьшения размерности	319
Использование SVD для измерения невидимого	321
Выполнение факторного анализа и PCA	322
Психометрическая модель	323
В поисках скрытых факторов	323
Использование компонентов, а не факторов	324
Уменьшение размерности	325
Сжатие информации с использованием t-SNE	326
Понимание некоторых приложений	328
Распознавание лиц с помощью PCA	328
Извлечение тем с использованием NMF	331
Рекомендация фильмов	334
Глава 15. Кластеризация	337
Кластеризация методом k-средних	339
Понятие алгоритмов на основе центроидов	340
Пример с данными изображения	342
Поиск оптимального решения	343

Кластеризация больших данных	346
Иерархическая кластеризация	348
Использование иерархического кластерного решения	349
Использование двухфазного кластерного решения	351
Обнаружение новых групп с DBScan	353
Глава 16. Поиск выбросов в данных	357
Обнаружение выбросов	358
Что еще может пойти не так	359
Понятие аномалий и новых данных	360
Изучение простого одномерного метода	362
Опора на гауссово распределение	364
Предположения и проверка	365
Выработка многомерного подхода	367
Использование анализа основных компонентов	367
Использование кластерного анализа для определения выбросов	368
Автоматическое обнаружение с помощью изоляционного леса	370
Часть 5. Обучение на данных	373
Глава 17. Четыре простых, но эффективных алгоритма	375
Угадай число: линейная регрессия	376
Определение семейства линейных моделей	376
Использование большего количества переменных	378
Ограничения и проблемы	380
Переход к логистической регрессии	381
Применение логистической регрессии	381
Учет нескольких классов	382
Просто, как наивный байесовский классификатор	384
Наивный Байес не такой уж и наивный	386
Прогнозирование текстовых классификаций	387
Ленивое обучение с ближайшими соседями	389
Прогнозирование после наблюдения соседей	390
Осмысленный выбор параметра k	392
Глава 18. Перекрестная проверка, отбор и оптимизация	395
Размышляя над проблемой подбора модели	396
Понятие смещения и дисперсии	398
Определение стратегии выбора моделей	398

Различие между учебными и тестовыми наборами	402
Перекрестная проверка	405
Перекрестная проверка k-блоков	406
Выборка стратификации для сложных данных	407
Профессиональный выбор переменных	409
Выбор по одномерным критериям	409
Использование жадного поиска	411
Гиперпараметры	412
Реализация сеточного поиска	413
Попытка случайного поиска	418
Глава 19. Увеличение сложности с помощью линейных и нелинейных трюков	419
Использование нелинейных преобразований	420
Преобразования переменных	421
Создание взаимодействий между переменными	423
Регуляризация линейных моделей	428
Использование регрессии Ридж (L2)	429
Использование регрессии Лассо (L1)	430
Использование регуляризации	430
Объединение L1 и L2: ElasticNet	431
Как справиться с большими данными фрагмент за фрагментом	432
Определение, когда данных слишком много	432
Реализация стохастического градиентного спуска	432
Понятие метода опорных векторов	436
Вычислительный метод	437
Исправление многих новых параметров	440
Классификация с использованием SVC	442
Переходить на нелинейность легко	448
Выполнение регрессии с помощью SVR	450
Создание стохастического решения с помощью SVM	452
Играя с нейронными сетями	456
Понятие нейронных сетей	457
Классификация и регрессия с нейронами	458
Глава 20. Сила единения	461
Простое дерево решений	462
Понятие дерева решений	462
Создание деревьев для разных целей	466

Как сделать доступным машинное обучение	469
Работа с классификатором Random Forest	471
Работа с регрессором Random Forest	472
Оптимизация Random Forest	473
Бустинг прогнозов	475
Зная, что победят многие слабые предикторы	475
Установка классификатора градиентного бустинга	476
Запуск регрессора градиентного бустинга	477
Использование гиперпараметров GBM	478
Часть 6. Великолепные десятки	481
Глава 21. Десять основных источников данных	483
Поиск новостей в Subreddit	484
Хорошее начало с KDnuggets	484
Поиск бесплатных учебных материалов с помощью Quora	485
Получение знаний на блоге Oracle Data Science	485
Доступ к огромному списку ресурсов на Data Science Central	486
Изучение новых трюков на Aspirational Data Scientist	486
Наиболее авторитетные источники на Udacity	487
Получение справки о передовых темах в Conductrics	487
Получение фактов науки о данных с открытым исходным кодом от мастеров	488
Как сосредоточиться на ресурсах для разработчиков с Джонатаном Бауэром	489
Глава 22. Десять задач, которые вы должны решить	491
Знакомство с конкурсом Data Science London + Scikit-Learn	492
Прогнозирование выживания на “Титанике”	493
Как находить конкурсы Kaggle, соответствующие вашим потребностям	493
Как оттачивать свои стратегии	494
Пробираясь через набор данных MovieLens	495
Избавление от спама	496
Работа с рукописной информацией	496
Работа с изображениями	498
Анализ обзоров Amazon.com	499
Взаимодействие с огромным графом	499
Предметный указатель	501