

O'REILLY®



Python

ДЛЯ СЛОЖНЫХ
ЗАДАЧ

наука о данных
и машинное обучение

powered by



Дж. Вандер Плас

Дж. Вандер Плас

Python

для сложных задач
наука о данных:
и машинное обучение



Санкт-Петербург · Москва · Екатеринбург · Воронеж
Нижний Новгород · Ростов-на-Дону · Самара · Минск

2018

ББК 32.973.2-018.1
УДК 004.43
ПЗ7

Плас Дж. Вандер

ПЗ7 Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил. — (Серия «Бестселлеры O'Reilly»).
ISBN 978-5-496-03068-7

Книга «Python для сложных задач: наука о данных и машинное обучение» — это подробное руководство по самым разным вычислительным и статистическим методам, без которых немислима любая интенсивная обработка данных, научные исследования и передовые разработки. Читатели, уже имеющие опыт программирования и желающие эффективно использовать Python в сфере Data Science, найдут в этой книге ответы на всевозможные вопросы, например: как считать этот формат данных в скрипт? как преобразовать, очистить эти данные и манипулировать ими? как визуализировать данные такого типа? как при помощи этих данных разобраться в ситуации, получить ответы на вопросы, построить статистические модели или реализовать машинное обучение?

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.973.2-018.1
УДК 004.43

Права на издание получены по соглашению с O'Reilly. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-1491912058 англ.

ISBN 978-5-496-03068-7

Authorized Russian translation of the English edition of Python Data Science Handbook, ISBN 9781491912058 © 2017 Jake VanderPlas

© Перевод на русский язык ООО Издательство «Питер», 2018

© Издание на русском языке, оформление ООО Издательство «Питер», 2018

© Серия «Бестселлеры O'Reilly», 2018

Оглавление

Предисловие	16
Что такое наука о данных.....	16
Для кого предназначена эта книга	17
Почему Python.....	18
Общая структура книги.....	19
Использование примеров кода	19
Вопросы установки	20
Условные обозначения	21
Глава 1. IPython: за пределами обычного Python	22
Командная строка или блокнот?	23
Запуск командной оболочки IPython.....	23
Запуск блокнота Jupiter	23
Справка и документация в оболочке IPython	24
Доступ к документации с помощью символа ?	25
Доступ к исходному коду с помощью символов ??	27
Просмотр содержимого модулей с помощью Tab-автодополнения.....	28
Сочетания горячих клавиш в командной оболочке IPython	30
Навигационные горячие клавиши.....	31
Горячие клавиши ввода текста.....	31
Горячие клавиши для истории команд.....	32
Прочие горячие клавиши	33

6 Оглавление

«Магические» команды IPython	33
Вставка блоков кода: %paste и %cpaste.....	34
Выполнение внешнего кода: %run	35
Длительность выполнения кода: %timeit.....	36
Справка по «магическим» функциям: ?, %magic и %lsmagic	36
История ввода и вывода	37
Объекты In и Out оболочки IPython.....	37
Быстрый доступ к предыдущим выводам с помощью знака подчеркивания	38
Подавление вывода.....	39
Соответствующие «магические» команды	39
Оболочка IPython и использование системного командного процессора	40
Краткое введение в использование командного процессора.....	40
Инструкции командного процессора в оболочке IPython.....	42
Передача значений в командный процессор и из него.....	42
«Магические» команды для командного процессора.....	43
Ошибки и отладка	44
Управление исключениями: %xmode	44
Отладка: что делать, если чтения трассировок недостаточно	47
Профилирование и мониторинг скорости выполнения кода.....	49
Оценка времени выполнения фрагментов кода: %timeit и %time	50
Профилирование сценариев целиком: %prun.....	52
Пошаговое профилирование с помощью %lprun.....	53
Профилирование использования памяти: %memit и %mprun.....	54
Дополнительные источники информации об оболочке IPython	56
Веб-ресурсы	56
Книги	56
Глава 2. Введение в библиотеку NumPy	58
Работа с типами данных в языке Python	59
Целое число в языке Python — больше, чем просто целое число.....	60
Список в языке Python — больше, чем просто список.....	62
Массивы с фиксированным типом в языке Python.....	63

Создание массивов из списков языка Python	64
Создание массивов с нуля	65
Стандартные типы данных библиотеки NumPy	66
Введение в массивы библиотеки NumPy	67
Атрибуты массивов библиотеки NumPy	68
Индексация массива: доступ к отдельным элементам	69
Срезы массивов: доступ к подмассивам	70
Изменение формы массивов	74
Слияние и разбиение массивов	75
Выполнение вычислений над массивами библиотеки NumPy: универсальные функции	77
Медлительность циклов	77
Введение в универсальные функции	79
Обзор универсальных функций библиотеки NumPy	80
Продвинутые возможности универсальных функций	84
Универсальные функции: дальнейшая информация	86
Агрегирование: минимум, максимум и все, что посередине	86
Суммирование значений из массива	87
Минимум и максимум	87
Пример: чему равен средний рост президентов США	90
Операции над массивами. Транслирование	91
Введение в транслирование	92
Правила транслирования	94
Транслирование на практике	97
Сравнения, маски и булева логика	98
Пример: подсчет количества дождливых дней	98
Операторы сравнения как универсальные функции	100
Работа с булевыми массивами	102
Булевы массивы как маски	104
«Прихотливая» индексация	108
Исследуем возможности «прихотливой» индексации	108

Комбинированная индексация.....	109
Пример: выборка случайных точек.....	110
Изменение значений с помощью прихотливой индексации.....	112
Пример: разбиение данных на интервалы.....	113
Сортировка массивов.....	116
Быстрая сортировка в библиотеке NumPy: функции <code>np.sort</code> и <code>np.argsort</code>	117
Частичные сортировки: секционирование.....	118
Пример: К ближайших соседей.....	119
Структурированные данные: структурированные массивы библиотеки NumPy.....	123
Создание структурированных массивов.....	125
Более продвинутые типы данных.....	126
Массивы записей: структурированные массивы с дополнительными возможностями.....	127
Вперед, к Pandas.....	128

Глава 3. Манипуляции над данными с помощью пакета Pandas.....	129
Установка и использование библиотеки Pandas.....	130
Знакомство с объектами библиотеки Pandas.....	131
Объект Series библиотеки Pandas.....	131
Объект DataFrame библиотеки Pandas.....	135
Объект Index библиотеки Pandas.....	138
Индексация и выборка данных.....	140
Выборка данных из объекта Series.....	140
Выборка данных из объекта DataFrame.....	144
Операции над данными в библиотеке Pandas.....	149
Универсальные функции: сохранение индекса.....	149
Универсальные функции: выравнивание индексов.....	150
Универсальные функции: выполнение операции между объектами DataFrame и Series.....	153
Обработка отсутствующих данных.....	154
Компромиссы при обозначении отсутствующих данных.....	155

Отсутствующие данные в библиотеке Pandas	155
Операции над пустыми значениями.....	159
Иерархическая индексация.....	164
Мультииндексированный объект Series.....	164
Методы создания мультииндексов.....	168
Индексация и срезы по мультииндексу.....	171
Перегруппировка мультииндексов.....	174
Агрегирование по мультииндексам.....	177
Объединение наборов данных: конкатенация и добавление в конец	178
Напоминание: конкатенация массивов NumPy	179
Простая конкатенация с помощью метода pd.concat.....	180
Объединение наборов данных: слияние и соединение.....	184
Реляционная алгебра	184
Виды соединений	185
Задание ключа слияния.....	187
Задание операций над множествами для соединений.....	191
Пересекающиеся названия столбцов: ключевое слово suffixes	192
Пример: данные по штатам США	193
Агрегирование и группировка.....	197
Данные о планетах.....	198
Простое агрегирование в библиотеке Pandas	198
GroupBy: разбиение, применение, объединение	200
Сводные таблицы	210
Данные для примеров работы со сводными таблицами	210
Сводные таблицы «вручную»	211
Синтаксис сводных таблиц	212
Пример: данные о рождаемости	214
Векторизованные операции над строками	219
Знакомство со строковыми операциями библиотеки Pandas	219
Таблицы методов работы со строками библиотеки Pandas.....	221
Пример: база данных рецептов	226
Работа с временными рядами	230
Дата и время в языке Python	231

Временные ряды библиотеки Pandas: индексация по времени.....	235
Структуры данных для временных рядов библиотеки Pandas	235
Периодичность и смещения дат.....	238
Где найти дополнительную информацию	246
Пример: визуализация количества велосипедов в Сиэтле	246
Увеличение производительности библиотеки Pandas: eval() и query()	252
Основания для использования функций query() и eval(): составные выражения	254
Использование функции pandas.eval() для эффективных операций.....	255
Использование метода DataFrame.eval() для выполнения операций по столбцам	257
Метод DataFrame.query().....	259
Производительность: когда следует использовать эти функции	259
Дополнительные источники информации	260

Глава 4. Визуализация с помощью библиотеки Matplotlib

Общие советы по библиотеке Matplotlib.....	263
Импорт matplotlib	263
Настройка стилей.....	263
Использовать show() или не использовать? Как отображать свои графики	264
Сохранение рисунков в файл	266
Два интерфейса по цене одного.....	267
Интерфейс в стиле MATLAB	267
Объектно-ориентированный интерфейс	268
Простые линейные графики	269
Настройка графика: цвета и стили линий.....	271
Настройка графика: пределы осей координат	273
Метки на графиках.....	276
Простые диаграммы рассеяния	278
Построение диаграмм рассеяния с помощью функции plt.plot.....	279
Построение диаграмм рассеяния с помощью функции plt.scatter	281

Сравнение функций plot и scatter: примечание относительно производительности.....	283
Визуализация погрешностей.....	283
Простые планки погрешностей.....	283
Непрерывные погрешности	285
Графики плотности и контурные графики.....	286
Гистограммы, разбиения по интервалам и плотность	290
Двумерные гистограммы и разбиения по интервалам.....	292
Ядерная оценка плотности распределения.....	294
Пользовательские настройки легенд на графиках	295
Выбор элементов для легенды	297
Задание легенды для точек различного размера	298
Отображение нескольких легенд.....	300
Пользовательские настройки шкал цветов.....	301
Выбор карты цветов.....	302
Ограничения и расширенные возможности по использованию цветов.....	305
Дискретные шкалы цветов	306
Пример: рукописные цифры.....	306
Множественные субграфики	308
plt.axes: создание субграфиков вручную	309
plt.subplot: простые сетки субграфиков	310
Функция plt.subplots: создание всей сетки за один раз	311
Функция plt.GridSpec: более сложные конфигурации.....	313
Текст и поясняющие надписи	314
Пример: влияние выходных дней на рождение детей в США.....	315
Преобразования и координаты текста	317
Стрелки и поясняющие надписи	319
Пользовательские настройки делений на осях координат.....	321
Основные и промежуточные деления осей координат	322
Прячем деления и/или метки	323
Уменьшение или увеличение количества делений.....	324
Более экзотические форматы делений	325

Краткая сводка локаторов и форматов	328
Пользовательские настройки Matplotlib: конфигурации и таблицы стилей	328
Выполнение пользовательских настроек графиков вручную	329
Изменяем значения по умолчанию: rcParams	330
Таблицы стилей	332
Построение трехмерных графиков в библиотеке Matplotlib	336
Трехмерные точки и линии	337
Трехмерные контурные графики	338
Каркасы и поверхностные графики	340
Триангуляция поверхностей	341
Отображение географических данных с помощью Basemap	344
Картографические проекции	346
Отрисовка фона карты	351
Нанесение данных на карты	353
Пример: города Калифорнии	354
Пример: данные о температуре на поверхности Земли	355
Визуализация с помощью библиотеки Seaborn	357
Seaborn по сравнению с Matplotlib	358
Анализируем графики библиотеки Seaborn	360
Пример: время прохождения марафона	368
Дополнительные источники информации	377
Источники информации о библиотеке Matplotlib	377
Другие графические библиотеки языка Python	377
Глава 5. Машинное обучение	379
Что такое машинное обучение	380
Категории машинного обучения	380
Качественные примеры прикладных задач машинного обучения	381
Классификация: предсказание дискретных меток	381
Резюме	390
Знакомство с библиотекой Scikit-Learn	391
Представление данных в Scikit-Learn	391
API статистического оценивания библиотеки Scikit-Learn	394

Прикладная задача: анализ рукописных цифр.....	403
Резюме.....	408
Гиперпараметры и проверка модели	408
Соображения относительно проверки модели	409
Выбор оптимальной модели	413
Кривые обучения	420
Проверка на практике: поиск по сетке	425
Резюме.....	426
Проектирование признаков	427
Категориальные признаки	427
Текстовые признаки	429
Признаки для изображений.....	430
Производные признаки	430
Внесение отсутствующих данных	433
Конвейеры признаков	434
Заглянем глубже: наивная байесовская классификация	435
Байесовская классификация.....	435
Гауссов наивный байесовский классификатор	436
Полиномиальный наивный байесовский классификатор	439
Когда имеет смысл использовать наивный байесовский классификатор	442
Заглянем глубже: линейная регрессия	443
Простая линейная регрессия	443
Регрессия по комбинации базисных функций	446
Регуляризация.....	450
Пример: предсказание велосипедного трафика.....	453
Заглянем глубже: метод опорных векторов	459
Основания для использования метода опорных векторов.....	459
Метод опорных векторов: максимизируем отступ.....	461
Пример: распознавание лиц.....	470
Резюме по методу опорных векторов	474
Заглянем глубже: деревья решений и случайные леса	475
Движущая сила случайных лесов: деревья принятия решений.....	475

Ансамбли оценщиков: случайные леса.....	481
Регрессия с помощью случайных лесов	482
Пример: использование случайного леса для классификации цифр	484
Резюме по случайным лесам	486
Заглянем глубже: метод главных компонент	487
Знакомство с методом главных компонент	487
Использование метода PCA для фильтрации шума	495
Пример: метод Eigenfaces.....	497
Резюме метода главных компонент	500
Заглянем глубже: обучение на базе многообразий	500
Обучение на базе многообразий: HELLO.....	501
Многомерное масштабирование (MDS)	502
MDS как обучение на базе многообразий	505
Нелинейные вложения: там, где MDS не работает	507
Нелинейные многообразия: локально линейное вложение	508
Некоторые соображения относительно методов обучения на базе многообразий	510
Пример: использование Isomap для распознавания лиц.....	511
Пример: визуализация структуры цифр.....	515
Заглянем глубже: кластеризация методом k -средних.....	518
Знакомство с методом k -средних	518
Алгоритм k -средних: максимизация математического ожидания.....	520
Примеры	525
Заглянем глубже: смеси Гауссовых распределений.....	532
Причины появления GMM: недостатки метода k -средних.....	532
Обобщение EM-модели: смеси Гауссовых распределений.....	535
GMM как метод оценки плотности распределения	540
Пример: использование метода GMM для генерации новых данных	544
Заглянем глубже: ядерная оценка плотности распределения	547
Обоснование метода KDE: гистограммы	547
Ядерная оценка плотности распределения на практике	552
Пример: KDE на сфере	554
Пример: не столь наивный байес	557

Прикладная задача: конвейер распознавания лиц.....	562
Признаки в методе HOG	563
Метод HOG в действии: простой детектор лиц	564
Предостережения и дальнейшие усовершенствования	569
Дополнительные источники информации по машинному обучению	571
Машинное обучение в языке Python	571
Машинное обучение в целом.....	572
Об авторе	573