

2-е издание

Python для сложных задач

Наука о данных

Джейк Вандер Плас

SPRINT
book

2025

Краткое содержание

Предисловие	19
-------------------	----

ЧАСТЬ I. JUPYTER: ЗА ПРЕДЕЛАМИ ОБЫЧНОГО PYTHON

Глава 1. Знакомство с IPython и Jupyter	27
Глава 2. Расширенные интерактивные возможности	38
Глава 3. Отладка и профилирование	48

ЧАСТЬ II. ВВЕДЕНИЕ В NUMPY

Глава 4. Типы данных в Python	64
Глава 5. Введение в массивы NumPy	72
Глава 6. Вычисления с массивами NumPy: универсальные функции	81
Глава 7. Агрегирование: минимум, максимум и все, что посередине	91
Глава 8. Операции над массивами. Транслирование	97
Глава 9. Сравнения, маски и булева логика	104
Глава 10. «Прихотливая» индексация	114
Глава 11. Сортировка массивов	123
Глава 12. Структурированные данные: структурированные массивы NumPy	130

ЧАСТЬ III. МАНИПУЛЯЦИИ НАД ДАННЫМИ С ПОМОЩЬЮ ПАКЕТА PANDAS

Глава 13. Знакомство с объектами библиотеки Pandas	138
Глава 14. Индексация и выборка данных	147
Глава 15. Операции над данными в библиотеке Pandas	155
Глава 16. Обработка отсутствующих данных	161
Глава 17. Иерархическая индексация	171
Глава 18. Объединение наборов данных: конкатенация и добавление в конец	185
Глава 19. Объединение наборов данных: слияние и соединение	191
Глава 20. Агрегирование и группировка	205
Глава 21. Сводные таблицы	218
Глава 22. Векторизованные операции над строками	227
Глава 23. Работа с временными рядами	237
Глава 24. Увеличение производительности библиотеки Pandas: eval() и query()	259

ЧАСТЬ IV. ВИЗУАЛИЗАЦИЯ С ПОМОЩЬЮ БИБЛИОТЕКИ MATPLOTLIB

Глава 25. Общие советы по библиотеке Matplotlib	269
Глава 26. Простые линейные графики	276
Глава 27. Простые диаграммы рассеяния	287
Глава 28. Графики плотности и контурные графики.....	297
Глава 29. Настройка легенд на графиках	308
Глава 30. Настройка цветовых шкал	315
Глава 31. Множественные субграфики	324
Глава 32. Текст и поясняющие надписи	332
Глава 33. Настройка делений на осях координат	340
Глава 34. Настройка Matplotlib: конфигурации и таблицы стилей	349
Глава 35. Построение трехмерных графиков в библиотеке Matplotlib	358
Глава 36. Визуализация с помощью библиотеки Seaborn	368

ЧАСТЬ V. МАШИННОЕ ОБУЧЕНИЕ

Глава 37. Что такое машинное обучение	389
Глава 38. Знакомство с библиотекой Scikit-Learn	401
Глава 39. Гиперпараметры и проверка модели	419
Глава 40. Проектирование признаков	437
Глава 41. Заглянем глубже: наивная байесовская классификация.....	445
Глава 42. Заглянем глубже: линейная регрессия.....	455
Глава 43. Заглянем глубже: метод опорных векторов	472
Глава 44. Заглянем глубже: деревья решений и случайные леса	489
Глава 45. Заглянем глубже: метод главных компонент.....	501
Глава 46. Заглянем глубже: обучение на базе многообразий	515
Глава 47. Заглянем глубже: кластеризация методом k средних.....	534
Глава 48. Заглянем глубже: смеси гауссовых распределений.....	549
Глава 49. Заглянем глубже: ядерная оценка плотности распределения	565
Глава 50. Прикладная задача: конвейер распознавания лиц.....	578
Об авторе	589
Иллюстрация на обложке	590

Оглавление

Отзывы ко второму изданию	18
Предисловие	19
Что такое наука о данных	19
Для кого предназначена эта книга	20
Почему Python	21
Общая структура книги	21
Вопросы установки	22
Условные обозначения	23
Использование примеров кода	24
Полноцветные иллюстрации	24
 ЧАСТЬ I. JUPYTER: ЗА ПРЕДЕЛАМИ ОБЫЧНОГО PYTHON	
Глава 1. Знакомство с IPython и Jupyter	27
Запуск командной оболочки IPython	27
Запуск Jupyter Notebook	28
Справка и документация в IPython	28
Доступ к документации с помощью символа ?	29
Доступ к исходному коду с помощью символов ??	31
Исследование содержимого модулей с помощью функции автодополнения	32
Горячие клавиши в командной оболочке IPython	34
Навигационные горячие клавиши	35
Горячие клавиши ввода текста	35
Горячие клавиши для истории команд	36
Прочие горячие клавиши	37
Глава 2. Расширенные интерактивные возможности	38
Магические команды IPython	38
Выполнение внешнего кода: %run	38
Измерение продолжительности выполнения кода: %timeit	39
Справка по «магическим» функциям: ?, %magic и %lsmagic	40
История ввода и вывода	40
Объекты In и Out оболочки IPython	40
Быстрый доступ к предыдущим выводам с помощью знака подчеркивания	42
Подавление вывода	42
Соответствующие «магические» команды	43
IPython и использование системного командного процессора	43
Краткое введение в использование командного процессора	44
Инструкции командного процессора в оболочке IPython	45
Передача значений в командный процессор и из него	45
«Магические» команды для командного процессора	46

Глава 3. Отладка и профилирование	48
Ошибки и отладка	48
Управление исключениями: %xmode	48
Отладка: что делать, если информации в трассировке недостаточно	50
Профилирование и хронометраж выполнения кода	53
Хронометраж выполнения фрагментов кода: %timeit и %time	53
Профилирование сценариев целиком: %prun	55
Пошаговое профилирование с помощью %lprun	56
Профилирование потребления памяти: %memit и %mprun	57
Дополнительные источники информации об оболочке IPython	59
Веб-ресурсы	59
Книги	59

ЧАСТЬ II. ВВЕДЕНИЕ В NUMPY

Глава 4. Типы данных в Python	64
Целое число в Python — больше, чем просто целое число	65
Список в Python — больше, чем просто список	66
Массивы фиксированного типа в Python	68
Создание массивов из списков	68
Создание массивов с нуля	69
Стандартные типы данных NumPy	70
Глава 5. Введение в массивы NumPy	72
Атрибуты массивов NumPy	72
Индексация массива: доступ к отдельным элементам	73
Срезы массивов: доступ к подмассивам	74
Одномерные подмассивы	75
Многомерные подмассивы	75
Подмассивы как представления	76
Создание копий массивов	77
Изменение формы массивов	77
Слияние и разбиение массивов	78
Слияние массивов	78
Разбиение массивов	80
Глава 6. Вычисления с массивами NumPy: универсальные функции	81
Медлительность циклов	81
Введение в универсальные функции	83
Обзор универсальных функций в библиотеке NumPy	84
Арифметические операции над массивами	84
Абсолютное значение	85
Тригонометрические функции	86
Показательные функции и логарифмы	86
Специализированные универсальные функции	87
Продвинутое возможности универсальных функций	88
Сохранение результатов в массиве	88

Сводные показатели.....	89
Векторные произведения.....	90
Универсальные функции: дополнительная информация.....	90
Глава 7. Агрегирование: минимум, максимум и все, что посередине.....	91
Суммирование значений в массиве.....	91
Минимум и максимум.....	92
Многомерные сводные показатели.....	93
Другие функции агрегирования.....	93
Пример: чему равен средний рост президентов США.....	94
Глава 8. Операции над массивами. Транслирование.....	97
Введение в транслирование.....	97
Правила транслирования.....	99
Транслирование. Пример 1.....	99
Транслирование. Пример 2.....	100
Транслирование. Пример 3.....	101
Транслирование на практике.....	102
Центрирование массива.....	102
Построение графика двумерной функции.....	103
Глава 9. Сравнения, маски и булева логика.....	104
Пример: подсчет количества дождливых дней.....	104
Операторы сравнения как универсальные функции.....	106
Работа с булевыми массивами.....	107
Подсчет количества элементов.....	108
Булевы операторы.....	109
Булевы массивы как маски.....	110
Ключевые слова and/or и операторы &/.....	111
Глава 10. «Прихотливая» индексация.....	114
Возможности «прихотливой» индексации.....	114
Комбинированная индексация.....	116
Пример: выборка случайных точек.....	116
Изменение значений с помощью «прихотливой» индексации.....	118
Пример: разбиение данных на интервалы.....	120
Глава 11. Сортировка массивов.....	123
Быстрая сортировка в библиотеке NumPy: функции np.sort и np.argsort.....	124
Сортировка по строкам и столбцам.....	124
Частичная сортировка: секционирование.....	125
Пример: k ближайших соседей.....	126
Глава 12. Структурированные данные: структурированные массивы NumPy.....	130
Создание структурированных массивов.....	132
Более продвинутые типы данных.....	133
Массивы записей: структурированные массивы с дополнительными возможностями.....	133
Вперед, к Pandas.....	134

ЧАСТЬ III. МАНИПУЛЯЦИИ НАД ДАННЫМИ С ПОМОЩЬЮ ПАКЕТА PANDAS

Глава 13. Знакомство с объектами библиотеки Pandas	138
Объект Series.....	138
Объект Series как обобщенный массив NumPy.....	139
Объект Series как специализированный словарь.....	140
Создание объектов Series.....	141
Объект DataFrame.....	142
DataFrame как обобщенный массив NumPy.....	142
Объект DataFrame как специализированный словарь.....	143
Создание объектов DataFrame.....	144
Объект Index.....	145
Объект Index как неизменяемый массив.....	146
Index как упорядоченное множество.....	146
Глава 14. Индексация и выборка данных	147
Выборка данных из объекта Series.....	147
Объект Series как словарь.....	147
Объект Series как одномерный массив.....	148
Индексаторы: loc и iloc.....	149
Выборка данных из объекта DataFrame.....	150
Объект DataFrame как словарь.....	150
Объект DataFrame как двумерный массив.....	152
Дополнительный синтаксис для индексации.....	154
Глава 15. Операции над данными в библиотеке Pandas	155
Универсальные функции: сохранение индекса.....	155
Универсальные функции: согласование индексов.....	156
Согласование индексов в объектах Series.....	156
Согласование индексов в объектах DataFrame.....	158
Универсальные функции: операции между объектами DataFrame и Series.....	159
Глава 16. Обработка отсутствующих данных	161
Компромиссы при обозначении отсутствующих данных.....	161
Отсутствующие данные в Pandas.....	162
None как значение-индикатор.....	163
NaN: отсутствующие числовые данные.....	164
Значения NaN и None в библиотеке Pandas.....	165
Типы данных с поддержкой пустых значений в Pandas.....	166
Операции над пустыми значениями.....	167
Выявление пустых значений.....	167
Удаление пустых значений.....	168
Заполнение пустых значений.....	169
Глава 17. Иерархическая индексация	171
Мультииндексированный объект Series.....	171
Плохой способ.....	172

Лучший способ: объект MultiIndex	173
Мультииндекс как дополнительное измерение.....	174
Методы создания объектов MultiIndex.....	175
Явные конструкторы MultiIndex	176
Названия уровней мультииндексов.....	177
Мультииндекс для столбцов.....	177
Индексация и срезы по мультииндексу.....	178
Мультииндексация объектов Series	178
Мультииндексация объектов DataFrame	180
Перегруппировка мультииндексов	181
Отсортированные и неотсортированные индексы	181
Выполнение операций stack и unstack над индексами	183
Создание и перестройка индексов	183
Глава 18. Объединение наборов данных: конкатенация и добавление в конец.....	185
Напоминание: конкатенация массивов NumPy	186
Простая конкатенация с помощью метода pd.concat	187
Дублирование индексов	188
Конкатенация с использованием соединений.....	189
Метод append()	190
Глава 19. Объединение наборов данных: слияние и соединение	191
Реляционная алгебра.....	192
Виды соединений.....	192
Соединения «один-к-одному».....	192
Соединения «многие-к-одному».....	193
Соединения «многие-ко-многим»	194
Задание ключа слияния	194
Именованный аргумент on	195
Именованные аргументы left_on и right_on	195
Именованные аргументы left_index и right_index.....	196
Применение операций над множествами для соединений	197
Пересекающиеся имена столбцов: именованный аргумент suffixes	199
Пример: данные по штатам США	200
Глава 20. Агрегирование и группировка	205
Данные о планетах	206
Простое агрегирование в библиотеке Pandas	206
groupby: разбиение, применение, объединение.....	208
Разбиение, применение и объединение	208
Объект GroupBy.....	211
Агрегирование, фильтрация, преобразование, применение	213
Задание ключа разбиения.....	215
Пример группировки.....	217
Глава 21. Сводные таблицы.....	218
Примеры для изучения приемов работы со сводными таблицами	218
Сводные таблицы «вручную».....	219

Синтаксис сводных таблиц	220
Многоуровневые сводные таблицы.....	220
Дополнительные параметры сводных таблиц.....	221
Пример: данные о рождаемости	222
Глава 22. Векторизованные операции над строками	227
Знакомство со строковыми операциями в библиотеке Pandas	227
Таблица строковых методов в библиотеке Pandas.....	228
Методы, аналогичные строковым методам языка Python.....	228
Методы, использующие регулярные выражения.....	230
Прочие методы	231
Пример: база данных рецептов.....	233
Простая рекомендательная система для рецептов.....	235
Дальнейшая работа с рецептами	236
Глава 23. Работа с временными рядами	237
Дата и время в языке Python.....	238
Представление даты и времени в Python: пакеты datetime и dateutil	238
Типизированные массивы значений времени: тип datetime64 библиотеки NumPy.....	239
Даты и время в библиотеке Pandas: лучшее из обоих миров.....	241
Временные ряды библиотеки Pandas: индексация по времени	241
Структуры данных для временных рядов библиотеки Pandas	242
Регулярные последовательности: функция pd.date_range()	243
Периодичность и смещение дат	244
Передискретизация, временные сдвиги и окна.....	246
Передискретизация и изменение периодичности интервалов	248
Временные сдвиги	250
Скользящие окна	251
Пример: визуализация количества велосипедов в Сиэтле	252
Визуализация данных	253
Углубленное изучение данных.....	256
Глава 24. Увеличение производительности библиотеки Pandas: eval() и query().....	259
Основания для использования функций query() и eval(): составные выражения	259
Использование функции pandas.eval() для эффективных операций	260
Использование метода DataFrame.eval() для выполнения операций по столбцам	262
Присваивание в методе DataFrame.eval()	263
Локальные переменные в методе DataFrame.eval().....	264
Метод DataFrame.query()	264
Производительность: когда следует использовать эти функции.....	265
Дополнительные источники информации.....	266

ЧАСТЬ IV. ВИЗУАЛИЗАЦИЯ С ПОМОЩЬЮ БИБЛИОТЕКИ MATPLOTLIB

Глава 25. Общие советы по библиотеке Matplotlib	269
Импортирование matplotlib	269
Настройка стилей	269
Использовать или не использовать show()? Как отображать графики	270
Построение графиков в сценариях	270
Построение графиков из командной оболочки IPython	271
Построение графиков из блокнота Jupyter	271
Сохранение изображений в файлы	272
Два интерфейса по цене одного	273
Глава 26. Простые линейные графики	276
Настройка графика: цвета и стили линий	279
Настройка графика: пределы осей координат	281
Метки на графиках	284
Нюансы использования Matplotlib	285
Глава 27. Простые диаграммы рассеяния	287
Построение диаграмм рассеяния с помощью plt.plot	287
Построение диаграмм рассеяния с помощью plt.scatter	290
plot и scatter: примечание относительно производительности	292
Визуализация погрешностей	293
Простые планки погрешностей	293
Непрерывные погрешности	295
Глава 28. Графики плотности и контурные графики	297
Визуализация трехмерной функции	297
Гистограммы, разбиения по интервалам и плотность	301
Двумерные гистограммы и разбиение по интервалам	304
Функция plt.hist2d: двумерная гистограмма	304
Функция plt.hexbin: гексагональное разбиение по интервалам	305
Ядерная оценка плотности распределения	305
Глава 29. Настройка легенд на графиках	308
Выбор элементов для легенды	310
Задание легенды для точек разного размера	312
Отображение нескольких легенд	313
Глава 30. Настройка цветовых шкал	315
Настройка цветовой шкалы	315
Выбор карты цветов	317
Ограничение и расширение карты цветов	319
Дискретные цветовые шкалы	320
Пример: рукописные цифры	321

Глава 31. Множественные субграфики	324
plt.axes: создание субграфиков вручную	324
plt.subplot: простые сетки субграфиков	326
plt.subplots: создание всей сетки за один раз	328
plt.GridSpec: более сложные конфигурации	329
Глава 32. Текст и поясняющие надписи	332
Преобразования и координаты текста	334
Стрелки и поясняющие надписи	336
Глава 33. Настройка делений на осях координат	340
Основные и промежуточные деления осей координат	340
Скрытие делений и/или меток	342
Уменьшение или увеличение количества делений	344
Экзотические форматы делений	345
Краткая сводка локаторов и форматеров	348
Глава 34. Настройка Matplotlib: конфигурации и таблицы стилей	349
Настройка графиков вручную	349
Изменение значений по умолчанию: rcParams	351
Таблицы стилей	353
Стиль по умолчанию default	354
Стиль FiveThirtyEight	354
Стиль ggplot	355
Стиль «байесовские методы для хакеров»	355
Стиль с темным фоном	356
Оттенки серого	356
Стиль Seaborn	357
Глава 35. Построение трехмерных графиков в библиотеке Matplotlib	358
Трехмерные точки и линии	359
Трехмерные контурные графики	360
Каркасы и поверхностные графики	362
Триангуляция поверхностей	364
Пример: визуализация ленты Мёбиуса	365
Глава 36. Визуализация с помощью библиотеки Seaborn	368
Анализируем графики Seaborn	369
Гистограммы, KDE и плотности	369
Графики пар	371
Фасетные гистограммы	372
Графики факторов	374
Совместные распределения	375
Столбиковые диаграммы	376
Пример: время прохождения марафона	377
Дополнительные источники информации	385
Другие графические библиотеки для Python	386

ЧАСТЬ V. МАШИННОЕ ОБУЧЕНИЕ

Глава 37. Что такое машинное обучение	389
Категории машинного обучения	390
Качественные примеры прикладных задач машинного обучения	390
Классификация: предсказание дискретных меток.....	391
Регрессия: предсказание непрерывных меток.....	393
Кластеризация: определение меток для немаркированных данных.....	396
Понижение размерности: определение структуры немаркированных данных	398
Резюме	400
Глава 38. Знакомство с библиотекой Scikit-Learn.....	401
Представление данных в Scikit-Learn	401
Матрица признаков	402
Целевой массив	402
API статистического оценивания в Scikit-Learn.....	404
Основы API статистического оценивания	405
Пример обучения с учителем: простая линейная регрессия.....	406
Пример обучения с учителем: классификация набора данных Iris.....	409
Пример обучения без учителя: понижение размерности набора данных Iris	410
Обучение без учителя: кластеризация набора данных Iris	412
Прикладная задача: анализ рукописных цифр	413
Загрузка и визуализация цифр	413
Обучение без учителя: понижение размерности	415
Классификация цифр	416
Резюме	418
Глава 39. Гиперпараметры и проверка модели.....	419
Соображения относительно проверки модели.....	419
Плохой способ проверки модели	420
Хороший способ проверки модели: отложенные данные	420
Перекрестная проверка модели	421
Выбор оптимальной модели.....	424
Компромисс между систематической ошибкой и дисперсией	424
Кривые проверки в библиотеке Scikit-Learn	427
Кривые обучения	430
Проверка на практике: поиск по сетке	435
Резюме	436
Глава 40. Проектирование признаков	437
Категориальные признаки.....	437
Текстовые признаки.....	439
Признаки для изображений.....	440
Производные признаки	440
Подстановка отсутствующих данных.....	443
Конвейеры признаков	444

Глава 41. Заглянем глубже: наивная байесовская классификация	445
Байесовская классификация	445
Гауссов наивный байесовский классификатор	446
Полиномиальный наивный байесовский классификатор	449
Пример: классификация текста	450
Когда имеет смысл использовать наивный байесовский классификатор	453
Глава 42. Заглянем глубже: линейная регрессия	455
Простая линейная регрессия	455
Регрессия по комбинации базисных функций	458
Полиномиальные базисные функции	458
Гауссовы базисные функции	460
Регуляризация	462
Гребневая регрессия (L_2 -регуляризация)	464
Лассо-регрессия (L_1 -регуляризация)	465
Пример: предсказание велосипедного трафика	466
Глава 43. Заглянем глубже: метод опорных векторов	472
Основания для использования метода опорных векторов	472
Метод опорных векторов: максимизация отступа	474
Аппроксимация методом опорных векторов	475
За пределами линейности: SVM-ядро	478
Настройка SVM: размытие отступов	481
Пример: распознавание лиц	483
Резюме	487
Глава 44. Заглянем глубже: деревья решений и случайные леса	489
Движущая сила случайных лесов: деревья принятия решений	489
Создание дерева принятия решений	490
Деревья принятия решений и переобучение	493
Ансамбли моделей: случайные леса	494
Регрессия с помощью случайных лесов	496
Пример: использование случайного леса для классификации цифр	497
Резюме	500
Глава 45. Заглянем глубже: метод главных компонент	501
Знакомство с методом главных компонент	501
PCA как метод понижения размерности	504
Использование метода PCA для визуализации: рукописные цифры	505
Что означают компоненты?	507
Выбор количества компонент	508
Использование метода PCA для фильтрации шума	509
Пример: метод Eigenfaces	511
Резюме	514
Глава 46. Заглянем глубже: обучение на базе многообразий	515
Обучение на базе многообразий: «HELLO»	516

Многомерное масштабирование (MDS)	517
MDS как обучение на базе многообразий.....	520
Нелинейные вложения: там, где MDS не работает	521
Нелинейные многообразия: локально линейное вложение	523
Некоторые соображения относительно методов обучения на базе многообразий	525
Пример: использование Isomap для распознавания лиц	526
Пример: визуализация структуры цифр	530
Глава 47. Заглянем глубже: кластеризация методом k средних	534
Знакомство с методом k средних	534
Максимизация математического ожидания.....	536
Примеры	542
Пример 1: применение метода k средних для распознавания рукописных цифр	542
Пример 2: использование метода k средних для сжатия цветов	545
Глава 48. Заглянем глубже: смеси гауссовых распределений	549
Причины появления GMM: недостатки метода k средних.....	549
Обобщение EM-модели: смеси гауссовых распределений	553
Выбор типа ковариации	557
GMM как метод оценки плотности распределения	557
Пример: использование метода GMM для генерации новых данных.....	561
Глава 49. Заглянем глубже: ядерная оценка плотности распределения	565
Обоснование метода KDE: гистограммы	565
Ядерная оценка плотности распределения на практике.....	570
Выбор ширины ядра путем перекрестной проверки	571
Пример: не столь наивный байес	572
Внутреннее устройство пользовательской модели	574
Использование пользовательской модели	576
Глава 50. Прикладная задача: конвейер распознавания лиц	578
Признаки HOG	579
Метод HOG в действии: простой детектор лиц	580
1. Получаем набор положительных обучающих образцов	580
2. Получаем набор отрицательных обучающих образцов	581
3. Объединяем наборы и выделяем HOG-признаки.....	582
4. Обучаем метод опорных векторов	583
5. Выполняем поиск лиц в новом изображении.....	583
Предостережения и дальнейшие усовершенствования.....	585
Дополнительные источники информации по машинному обучению	587
Об авторе	589
Иллюстрация на обложке	590