

O'REILLY®



Data Science

Наука о данных с нуля

bhv®

Джоэл Грас

Джоэл Грас

Data Science

Наука о данных с нуля

Санкт-Петербург

«БХВ-Петербург»

2018

УДК 004.6

ББК 32.81

Г77

Грас Дж.

Г77 Data Science. Наука о данных с нуля: Пер. с англ. — СПб.: БХВ-Петербург, 2018. — 336 с.: ил.

ISBN 978-5-9775-3758-2

Книга позволяет изучить науку о данных (Data Science) и применить полученные знания на практике. Она написана так, что способствует погружению в Data Science аналитика, фактически не обладающего глубокими знаниями в этой прикладной дисциплине.

В объемах, достаточных для начала работы в области Data Science, книга содержит интенсивный курс языка Python, элементы линейной алгебры, математической статистики, теории вероятностей, методов сбора, очистки, нормализации и обработки данных. Даны основы машинного обучения. Описаны различные математические модели и их реализация по методу k ближайших соседей, наивной байесовской классификации, линейной и логистической регрессии, а также модели на основе деревьев принятия решений, нейронных сетей и кластеризации. Рассказано о работе с рекомендательными системами, описаны приемы обработки естественного языка, методы анализа социальных сетей, основы баз данных, SQL и MapReduce.

Для аналитиков данных

УДК 004.6
ББК 32.81

Группа подготовки издания:

Главный редактор	Екатерина Кондукова
Зам. главного редактора	Евгений Рыбаков
Зав. редакцией	Екатерина Капалыгина
Перевод с английского	Андрея Логунова
Редактор	Анна Кузьмина
Компьютерная верстка	Ольги Сергиенко
Корректор	Зинаида Дмитриева
Оформление обложки	Марины Дамбиевой

Authorized translation of the English edition of Data Science from Scratch (978-1-491-90142-7) © 2015 Joel Grus.
This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Авторизованный перевод английской редакции книги Data Science from Scratch (ISBN: 978-1-491-90142-7)
© 2015 Joel Grus.

Перевод опубликован и продается с разрешения O'Reilly Media, Inc., собственника всех прав на публикацию и продажу издания.

Подписано в печать 31.08.17.

Формат 70×100 $\frac{1}{16}$. Печать офсетная. Усл. печ. л. 27,09.

Доп. тираж 1500 экз. Заказ №5076.

"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

ООО "Печатное дело",
142300, МО, г. Чехов, ул. Полиграфистов, д. 1

ISBN 978-1-491-90142-7 (англ.)
ISBN 978-5-9775-3758-2 (рус.)

© 2015 Joel Grus
© Перевод на русский язык "БХВ-Петербург", 2017, 2018

Оглавление

Предисловие	11
Наука о данных	11
С чистого листа	12
Условные обозначения, принятые в книге	13
Использование примеров кода	14
Благодарности	15
Комментарий переводчика	16
Python 2 и Python 3	16
Установка и удаление дистрибутива Anaconda	17
Настройка дистрибутива Anaconda	18
Установка инструментальной среды Spyder	18
Настройка инструментальной среды Spyder	19
Настройка среды Spyder с Python 3 для работы с Python 2	19
Факультативно	20
Запуск сервера записных книжек Jupyter	20
Установка библиотек Python из whl-файла	20
Подготовка среды Python 3 в ОС Ubuntu Linux	21
Управление пакетами .deb в Ubuntu Linux	21
Об авторе	23
Глава 1. Введение	25
Господство данных	25
Что такое наука о данных?	25
Оправдание для выдумки: DataScienteer	27
Поиск ключевых звеньев	27
Аналитики, которых вы должны знать	30
Зарплаты и опыт работы	33
Оплата премиум-аккаунтов	35
Популярные темы	36
Вперед	38
Глава 2. Интенсивный курс языка Python	39
Основы	39
Установка	39
Дзен языка Python	40
Пробельные символы	40
Модули	41
Арифметические операции	42

Функции.....	43
Строки.....	44
Исключения.....	44
Списки	45
Кортежи	46
Словари.....	47
Словарь <i>defaultdict</i>	48
Словарь <i>Counter</i>	49
Множества.....	50
Управляющие конструкции	50
Истинность	51
Не совсем основы	52
Сортировка	52
Генераторы последовательностей	53
Функции-генераторы и генераторные выражения	54
Случайные числа.....	55
Регулярные выражения	56
Объектно-ориентированное программирование.....	56
Инструменты функционального программирования	58
Функция <i>enumerate</i>	59
Функция <i>zip</i> и распаковка аргументов	60
Переменные <i>args</i> и <i>kwargs</i>	60
Добро пожаловать в DataScienteester!	62
Для дальнейшего изучения	62
Глава 3. Визуализация данных.....	63
Библиотека <i>matplotlib</i>	63
Столбчатые диаграммы	65
Линейные графики.....	68
Точечные диаграммы	70
Для дальнейшего изучения	72
Глава 4. Линейная алгебра.....	73
Векторы	73
Матрицы	77
Для дальнейшего изучения	80
Глава 5. Статистика	81
Описание одиночного набора данных	81
Показатели центра распределения	83
Показатели вариации.....	85
Корреляция.....	87
Парадокс Симпсона	90
Некоторые другие ловушки корреляции	91
Корреляция и причинная зависимость.....	91
Для дальнейшего изучения	92
Глава 6. Теория вероятностей	93
Зависимость и независимость.....	93
Условная вероятность	94

Теорема Байеса	96
Случайные величины.....	97
Непрерывные распределения.....	98
Нормальное распределение	100
Центральная предельная теорема.....	103
Для дальнейшего изучения	105
Глава 7. Гипотеза и вывод.....	106
Проверка статистических гипотез.....	106
Пример: бросание монеты	106
P-значения	110
Доверительные интервалы.....	111
Подгонка <i>p</i> -значения	112
Пример: проведение <i>A/B</i> -тестирования	113
Байесовский статистический вывод.....	115
Для дальнейшего изучения	118
Глава 8. Градиентный спуск.....	119
Идея в основе метода градиентного спуска	119
Вычисление градиента	120
Использование градиента	123
Выбор оптимального размера шага	124
Собираем все вместе	124
Стохастический градиентный спуск	126
Для дальнейшего изучения	127
Глава 9. Сбор данных	129
Объекты <i>stdin</i> и <i>stdout</i>	129
Чтение файлов.....	131
Основы работы с текстовыми файлами	131
Файлы с разделителями.....	132
Извлечение данных из веб-ресурсов	134
Анализ кода HTML	134
Пример: книги об анализе данных издательства O'Reilly	137
Использование программных интерфейсов	141
Формат JSON (и XML)	141
Использование непроверенного API.....	142
Поиск API	144
Пример: использование интерфейсов Twitter API.....	144
Получение учетных данных	145
Использование Twython	146
Для дальнейшего изучения	148
Глава 10. Обработка данных.....	149
Исследование данных.....	149
Исследование одномерных данных	149
Двумерные данные	151
Многомерные данные.....	153
Очистка и форматирование	155
Управление данными	157

Шкалирование	160
Снижение размерности	162
Для дальнейшего изучения	168
Глава 11. Машинное обучение.....	169
Моделирование	169
Что такое машинное обучение?	170
Переобучение и недообучение	171
Правильность модели.....	173
Компромисс между смещением и дисперсией.....	176
Извлечение и отбор признаков	177
Для дальнейшего изучения	178
Глава 12. К ближайших соседей	180
Модель	180
Пример: предпочтительные языки	182
Проблема проклятия размерности	186
Для дальнейшего изучения	190
Глава 13. Наивный Байес	191
Действительно глупый спам-фильтр	191
Более продуманный спам-фильтр	192
Реализация	194
Тестирование модели	196
Для дальнейшего изучения	198
Глава 14. Простая линейная регрессия	199
Модель	199
Применение метода градиентного спуска	202
Метод максимального правдоподобия	203
Для дальнейшего изучения	204
Глава 15. Множественная регрессия.....	205
Модель	205
Другие допущения модели наименьших квадратов	206
Подбор модели	207
Интерпретация модели	208
Качество подбора модели	209
Отступление: бутстрэпирование данных	209
Стандартные ошибки коэффициентов регрессии	211
Регуляризация	213
Для дальнейшего изучения	215
Глава 16. Логистическая регрессия.....	216
Задача	216
Логистическая функция	218
Применение модели	220
Качество подбора модели	221
Метод опорных векторов	223
Для дальнейшего изучения	225

Глава 17. Деревья принятия решений	226
Что такое дерево принятия решений?	226
Энтропия	228
Энтропия разбиения	230
Создание дерева принятия решений	231
Собираем все вместе	233
Случайные леса	236
Для дальнейшего изучения	237
Глава 18. Нейронные сети	238
Перцептроны	238
Нейронные сети прямого распространения	240
Метод обратного распространения ошибки	243
Пример: преодоление капчи	244
Для дальнейшего изучения	249
Глава 19. Кластеризация	250
Идея	250
Модель	251
Пример: встречи для специалистов	252
Выбор числа k	254
Пример: кластеризация цвета	256
Восходящий метод иерархической кластеризации	257
Для дальнейшего изучения	263
Глава 20. Обработка естественного языка.....	264
Облака слов	264
N-граммные модели языка	266
Грамматики	269
Ремарка: метод сэмплирования по Гиббсу	271
Тематическое моделирование	273
Для дальнейшего изучения	278
Глава 21. Анализ социальных сетей.....	279
Центральность по посредничеству	279
Центральность собственного вектора	284
Умножение матриц	284
Центральность	287
Направленные графы и рейтинг PageRank	288
Для дальнейшего изучения	291
Глава 22. Рекомендательные системы.....	292
Неавтоматическое кураторство	293
Рекомендация популярных тем	293
Коллаборативная фильтрация на основе пользователя	294
Коллаборативная фильтрация по схожести предметов	297
Для дальнейшего изучения	300
Глава 23. Базы данных и SQL.....	301
Операторы <i>CREATE TABLE</i> и <i>INSERT</i>	301
Оператор <i>UPDATE</i>	303

Оператор <i>DELETE</i>	304
Оператор <i>SELECT</i>	304
Оператор <i>GROUP BY</i>	306
Оператор <i>ORDER BY</i>	308
Оператор <i>JOIN</i>	309
Подзапросы	311
Индексы	312
Оптимизация запросов	313
Базы данных NoSQL	313
Для дальнейшего изучения	314
Глава 24. Распределенные вычисления MapReduce	315
Пример: подсчет частотности слов	315
Почему MapReduce?	317
MapReduce в более общей реализации	318
Пример: анализ обновлений ленты новостей	319
Пример: умножение матриц	321
Ремарка: сумматоры	322
Для дальнейшего изучения	323
Глава 25. Идите и займитесь аналитикой	324
Интерактивная оболочка IPython	324
Математический аппарат	325
Не с чистого листа	325
Библиотека NumPy	326
Библиотека pandas	326
Библиотека scikit-learn	326
Визуализация	326
Язык программирования R	327
Где найти данные?	327
Занятия анализом данных	328
Новости хакера	328
Пожарные машины	329
Футболки	329
А вы?	330
Предметный указатель	331