

O'REILLY®

ПОЛНОЦВЕТНОЕ
ИЗДАНИЕ



ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ С ПОМОЩЬЮ PYTHON

РУКОВОДСТВО ДЛЯ СПЕЦИАЛИСТОВ
ПО РАБОТЕ С ДАННЫМИ

Андреас Мюллер, Сара Гвидо

Введение в машинное обучение с помощью Python

РУКОВОДСТВО ДЛЯ СПЕЦИАЛИСТОВ
ПО РАБОТЕ С ДАННЫМИ

Андреас Мюллер, Сара Гвидо



Москва · Санкт-Петербург · Киев
2017

32.973.26-018.2.75
М98
7ДК 681.3.07

Компьютерное издательство “Диалектика”
Главный редактор С.Н. Тригуб
Зав. редакцией А.В. Слепцов
Перевод с английского и редакция А.В. Груздева

По общим вопросам обращайтесь в издательство “Диалектика” по адресу:
info@dialektika.com, http://www.dialektika.com

Мюллер, Андреас, Гвидо, Сара.

498 Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. : Пер. с англ. — СПб. : ООО “Альфа-книга”, 2017. — 480 с. : ил. — Парал. тит. англ.

ISBN 978-5-9908910-8-1 (рус.)

ББК 32.973.26-018.2.75

Все названия программных продуктов являются зарегистрированными торговыми марками соответствующих фирм.

Никакая часть настоящего издания ни в каких целях не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, будь то электронные или механические, включая фотокопирование и запись на магнитный носитель, если на это нет письменного разрешения издательства O'Reilly & Associates.

Authorized Russian translation of the English edition of Introduction to Machine Learning with Python. A Guide for Data Scientists (ISBN 978-1-449-36941-5) © 2017 Sarah Guido, Andreas Muller.

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.

All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage or retrieval system, without the prior written permission of the copyright owner and the Publisher.

Научно-популярное издание
Андреас Мюллер, Сара Гвидо
Введение в машинное обучение с помощью Python
Руководство для специалистов по работе с данными

Литературный редактор *Л.Н. Красножон*
Верстка *О.В. Мишутина*
Художественный редактор *В.Г. Павлютин*
Корректор *Л.А. Гордиенко*

Подписано в печать 03.05.2017. Формат 70x100/16
Гарнитура Times. Печать офсетная
Усл. печ. л. 38,7. Уч.-изд. л. 27,5
Тираж 1000 экз. Заказ № 3029

Отпечатано в АО “Первая Образцовая типография”
Филиал “Чеховский Печатный Двор”
142300, Московская область, г. Чехов, ул. Полиграфистов, д. 1

ООО “Альфа-книга”, 195027, Санкт-Петербург, Магнитогорская ул., д. 30

ISBN 978-5-9908910-8-1 (рус.)

© 2017, Компьютерное изд-во “Диалектика”,
перевод, оформление, макетирование

ISBN 978-1-449-36941-5 (англ.)

© 2017, Sarah Guido, Andreas Muller

Оглавление

Предисловие	15
Глава 1. Введение	21
Глава 2. Методы машинного обучения с учителем	53
Глава 3. Методы машинного обучения без учителя и предварительная обработка данных	177
Глава 4. Типы данных и конструирование признаков	269
Глава 5. Оценка и улучшение качества модели	319
Глава 6. Объединение алгоритмов в цепочки и конвейеры	385
Глава 7. Работа с текстовыми данными	407
Глава 8. Подведение итогов	451
Предметный указатель	465

Содержание

Об авторах	11
Об изображении на обложке	11
Благодарности	13
От Андреаса	13
От Сары	14
От редакции	14
Предисловие	15
Кому стоит прочитать эту книгу	15
Почему мы написали эту книгу	16
Структура книги	16
Онлайн-ресурсы	18
Условные обозначения, принятые в этой книге	18
Пиктограммы, используемые в этой книге	19
Использование примеров программного кода	19
От издательства “Диалектика”	20
Глава 1. Введение	21
Зачем нужно использовать машинное обучение	22
Задачи, которые можно решить с помощью машинного обучения	23
Постановка задач и знакомство с данными	26
Почему нужно использовать Python	27
Библиотека scikit-learn	28
Установка библиотеки scikit-learn	28
Основные библиотеки и инструменты	29
Пакет Jupyter Notebook	30
Пакет NumPy	30
Пакет SciPy	31
Пакет matplotlib	32
Библиотека pandas	33
Библиотека mglearn	35

Сравнение Python 2 и Python 3	36
Версии библиотек, используемые в этой книге	37
Пример 1: классификация сортов ириса	38
Загружаем данные	39
Метрики эффективности: обучающий и тестовый наборы	43
Сперва посмотрите на свои данные	45
Построение первой модели: метод k ближайших соседей	47
Получение прогнозов	48
Оценка качества модели	49
Выводы и перспективы	50
Глава 2. Методы машинного обучения с учителем	53
Классификация и регрессия	53
Обобщающая способность, переобучение и недообучение	55
Взаимосвязь между сложностью модели и размером набора данных	58
Алгоритмы машинного обучения с учителем	59
Некоторые наборы данных	59
Метод k ближайших соседей	64
Линейные модели	75
Наивные байесовские классификаторы	103
Деревья решений	105
Ансамбли деревьев решений	123
Ядерный метод опорных векторов	134
Нейронные сети (глубокое обучение)	147
Оценки неопределенности для классификаторов	163
Решающая функция: <code>decision_function</code>	164
Прогнозирование вероятностей: <code>predict_proba</code>	167
Неопределенность в мультиклассовой классификации	169
Выводы и перспективы	174
Глава 3. Методы машинного обучения без учителя и предварительная обработка данных	177
Типы машинного обучения без учителя	177
Проблемы машинного обучения без учителя	178
Предварительная обработка и масштабирование	179
Виды предварительной обработки	180
Применение преобразований данных	181
Масштабирование обучающего и тестового наборов одинаковым образом	184

Влияние предварительной обработки на машинное обучение с учителем	187
Снижение размерности, выделение признаков и множественное обучение	189
Анализ главных компонент (PCA)	189
Факторизация неотрицательных матриц (NMF)	207
Множественное обучение с помощью алгоритма t-SNE	216
Кластеризация	221
Кластеризация по методу k -средних	221
Агломеративная кластеризация	235
Алгоритм DBSCAN	241
Сравнение и оценка качества алгоритмов кластеризации	246
Выводы по методам кластеризации	265
Выводы и перспективы	266
Глава 4. Типы данных и конструирование признаков	269
Категориальные переменные	270
Прямое кодирование (дамми-переменные)	271
Для кодирования категорий можно использовать числа	277
Биннинг, дискретизация, линейные модели и деревья	280
Взаимодействия и полиномы	285
Одномерные нелинейные преобразования	294
Автоматический отбор признаков	298
Одномерные статистики	299
Отбор признаков на основе модели	302
Итеративный отбор признаков	304
Применение экспертных знаний	306
Выводы и перспективы	317
Глава 5. Оценка и улучшение качества модели	319
Перекрестная проверка	320
Перекрестная проверка в библиотеке scikit-learn	321
Преимущества перекрестной проверки	322
Стратифицированная k -блочная перекрестная проверка и другие стратегии	324
Решетчатый поиск	331
Простой решетчатый поиск	332
Опасность переобучения параметров и проверочный набор данных	333
Решетчатый поиск с перекрестной проверкой	335

Метрики качества модели и их вычисление	350
Помните о конечной цели	350
Метрики для бинарной классификации	351
Метрики для мультиклассовой классификации	377
Метрики регрессии	380
Использование метрик оценки для отбора модели	381
Выводы и перспективы	383
Глава 6. Объединение алгоритмов в цепочки и конвейеры	385
Отбор параметров с использованием предварительной обработки	386
Построение конвейеров	388
Использование конвейера, помещенного в объект GridSearchCV	390
Общий интерфейс конвейера	393
Удобный способ построения конвейеров с помощью функции <code>make_pipeline</code>	395
Работа с атрибутами этапов	397
Работа с атрибутами конвейера, помещенного в объект GridSearchCV	397
Находим оптимальные параметры этапов конвейера с помощью решетчатого поиска	399
Выбор оптимальной модели с помощью решетчатого поиска	402
Выводы и перспективы	404
Глава 7. Работа с текстовыми данными	407
Строковые типы данных	407
Пример применения: анализ тональности киноотзывов	410
Представление текстовых данных в виде “мешка слов”	413
Применение модели “мешок слов” к синтетическому набору данных	415
Модель “мешка слов” для киноотзывов	417
Стоп-слова	422
Масштабирование данных с помощью метода <code>tf-idf</code>	424
Исследование коэффициентов модели	427
Модель “мешка слов” для последовательностей из нескольких слов (<i>n</i> -грамм)	429
Продвинутая токенизация, стемминг и лемматизация	436
Моделирование тем и кластеризация документов	440
Латентное размещение Дирихле	441
Выводы и перспективы	449

Глава 8. Подведение итогов	451
Общий подход к решению задач машинного обучения	451
Вмешательство человека в работу модели	453
От прототипа к производству	453
Тестирование производственных систем	455
Создание собственного класса Estimator	455
Куда двигаться дальше	457
Теория	457
Другие фреймворки и пакеты машинного обучения	458
Ранжирование, рекомендательные системы и другие виды обучения	459
Вероятностное моделирование, теория статистического вывода и вероятностное программирование	460
Нейронные сети	461
Масштабирование на больших наборах данных	462
Оттачивание навыков	463
Заключение	464
Предметный указатель	465