

УДК 004.438Python:004.6

ББК 32.973.22

К76

**Коэльо, Луис.**

К76 Построение систем машинного обучения на языке Python / Л. П. Коэльо, В. Ричарт ; пер. с англ. А. А. Слинкина. — 3-е изд., эл. — 1 файл pdf : 304 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-331-8

Применение машинного обучения для лучшего понимания природы данных — умение, необходимое любому современному разработчику программ или аналитику. Python — замечательный язык для создания приложений машинного обучения. Благодаря своей динамичности он позволяет быстро производить разведочный анализ данных и экспериментировать с ними. Обладая первоклассным набором библиотек машинного обучения с открытым исходным кодом, Python дает возможность сосредоточиться на решаемой задаче и в то же время опробовать различные идеи.

Книга начинается с краткого введения в предмет машинного обучения и знакомства с библиотеками NumPy, SciPy, scikit-learn. Но довольно быстро авторы переходят к более серьезным проектам с реальными наборами данных, в частности, тематическому моделированию, анализу корзины покупок, облачным вычислениям и др.

Издание рассчитано на программистов, пишущих на Python и желающих узнать о построении систем машинного обучения и научиться извлекать из данных ценную информацию, необходимую для решения различных задач.

УДК 004.438Python:004.6

ББК 32.973.22

**Электронное издание на основе печатного издания:** Построение систем машинного обучения на языке Python / Л. П. Коэльо, В. Ричарт ; пер. с англ. А. А. Слинкина. — 2-е изд. — Москва : ДМК Пресс, 2016. — 304 с. — ISBN 978-5-97060-330-7. — Текст : непосредственный.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-331-8

© 2015 Packt Publishing

© Оформление, перевод на русский язык,  
ДМК Пресс, 2016



# ОГЛАВЛЕНИЕ

<b>Об авторах .....</b>	<b>11</b>
<b>О рецензентах.....</b>	<b>13</b>
<b>Предисловие .....</b>	<b>15</b>
О содержании книги .....	15
Что необходимо для чтения этой книги.....	17
На кого рассчитана эта книга .....	17
Графические выделения.....	17
Отзывы.....	18
Поддержка клиентов .....	18
Загрузка кода примеров .....	19
Опечатки.....	19
Нарушение авторских прав .....	19
Вопросы .....	20
<b>Глава 1. Введение в машинное обучение на языке Python .....</b>	<b>21</b>
Машинное обучение и Python – команда мечты .....	22
Что вы узнаете (и чего не узнаете) из этой книги .....	23
Что делать, если вы застряли .....	25
Приступая к работе .....	26
Введение в NumPy, SciPy и matplotlib .....	26
Установка Python .....	27
NumPy как средство эффективной и SciPy как средство интеллектуальной обработки данных .....	27
Изучаем NumPy.....	27
Изучаем SciPy .....	32
Наше первое (простенькое) приложение машинного обучения .....	33
Чтение данных .....	33
Предварительная обработка и очистка данных .....	35
Выбор подходящей модели и обучающего алгоритма .....	36
Резюме .....	46

## Глава 2. Классификация в реальной жизни..... 47

Набор данных Iris.....	48
Визуализация – первый шаг к цели .....	48
Построение первой модели классификации .....	50
Оценка качества – резервирование данных и перекрестная проверка.....	53
Построение более сложных классификаторов.....	57
Более сложный набор данных и более сложный классификатор ...	58
Набор данных Seeds .....	58
Признаки и подготовка признаков.....	59
Классификация по ближайшему соседу .....	60
Классификация с помощью scikit-learn .....	61
Решающие границы .....	62
Бинарная и многоклассовая классификация .....	65
Резюме .....	66

## Глава 3. Кластеризация – поиск взаимосвязанных сообщений ..... 68

Измерение сходства сообщений .....	69
Как не надо делать .....	69
Как надо делать .....	70
Предварительная обработка – количество общих слов как мера сходства .....	71
Преобразование простого текста в набор слов .....	71
Развитие концепции стоп-слов .....	80
Чего мы достигли и к чему стремимся .....	81
Кластеризация .....	82
Метод К средних.....	83
Тестовые данные для проверки наших идей .....	85
Кластеризация сообщений .....	87
Решение исходной задачи.....	88
Другой взгляд на шум .....	90
Настройка параметров .....	92
Резюме .....	92

## Глава 4. Тематическое моделирование..... 93

Латентное размещение Дирихле .....	93
Построение тематической модели .....	95
Сравнение документов по темам.....	100
Моделирование всей википедии.....	103
Выбор числа тем .....	106

Резюме .....	107
<b>Глава 5. Классификация – выявление плохих ответов .....</b>	<b>109</b>
План действий.....	109
Учимся классифицировать классные ответы .....	110
Подготовка образца.....	110
Настройка классификатора.....	110
Получение данных .....	111
Сокращение объема данных .....	112
Предварительная выборка и обработка атрибутов .....	112
Что считать хорошим ответом? .....	114
Создание первого классификатора .....	115
Метод k ближайших соседей.....	115
Подготовка признаков .....	116
Обучение классификатора .....	117
Измерение качества классификатора .....	117
Проектирование дополнительных признаков .....	118
Как поправить дело? .....	121
Дилемма смещения-дисперсии .....	122
Устранение высокого смещения .....	122
Устранение высокой дисперсии .....	123
Низкое или высокое смещение? .....	123
Логистическая регрессия.....	125
Немного математики на простом примере.....	126
Применение логистической регрессии к задаче классификации .....	128
Не верностью единой – точность и полнота .....	129
Упрощение классификатора.....	133
К поставке готов!.....	134
Резюме .....	135
<b>Глава 6. Классификация II – анализ эмоциональной окраски.....</b>	<b>136</b>
План действий.....	136
Чтение данных из Твиттера.....	137
Введение в наивный байесовский классификатор.....	137
О теореме Байеса .....	138
Что значит быть наивным .....	139
Использование наивного байесовского алгоритма для классификации.....	140
Учет ранее не встречавшихся слов и другие тонкости .....	143
Борьба с потерей точности при вычислениях .....	144

Создание и настройка классификатора .....	147
Сначала решим простую задачу .....	147
Использование всех классов .....	150
Настройка параметров классификатора .....	153
Очистка твитов .....	157
Учет типов слов .....	159
Определение типов слов .....	159
Удачный обмен с помощью SentiWordNet .....	162
Наш первый оценщик .....	164
Соберем все вместе .....	166
Резюме .....	167
<b>Глава 7. Регрессия .....</b>	<b>168</b>
Прогнозирование стоимости домов с помощью регрессии .....	168
Многомерная регрессия .....	172
Перекрестная проверка для регрессии .....	173
Регрессия со штрафом, или регуляризованная регрессия .....	174
Штрафы L1 и L2 .....	175
Lasso и эластичная сеть в библиотеке scikit-learn .....	176
Визуализация пути в Lasso .....	177
Сценарии Р-больше-N .....	178
Пример, основанный на текстовых документах .....	179
Объективный подход к заданию гиперпараметров .....	181
Резюме .....	185
<b>Глава 8. Рекомендации .....</b>	<b>186</b>
Прогноз и рекомендация оценок .....	186
Разделение данных на обучающие и тестовые .....	188
Нормировка обучающих данных .....	189
Рекомендование на основе ближайших соседей .....	191
Регрессионный подход к рекомендованию .....	195
Комбинирование нескольких методов .....	196
Анализ корзины .....	199
Получение полезных прогнозов .....	200
Анализ корзин покупок в супермаркете .....	201
Поиск ассоциативных правил .....	204
Более сложный анализ корзины .....	206
Резюме .....	207
<b>Глава 9. Классификация по музыкальным жанрам .....</b>	<b>208</b>
План действий .....	208
Получение музыкальных данных .....	209
Преобразование в формат WAV .....	209

Взгляд на музыку .....	210
Разложение на синусоидальные волны .....	211
Применение БПФ для построения первого классификатора .....	213
Повышение гибкости эксперимента.....	213
Обучение классификатора .....	215
Применение матрицы неточностей для измерения верности в многоклассовых задачах .....	215
Альтернативный способ измерения качества классификатора с помощью рабочей характеристики приемника .....	218
Повышение качества классификации с помощью мел-частотных кепстральных коэффициентов.....	220
Резюме .....	225
<b>Глава 10. Машинное зрение .....</b>	<b>227</b>
Введение в обработку изображений.....	227
Загрузка и показ изображения.....	228
Бинаризация.....	230
Гауссово размывание.....	231
Помещение центра в фокус.....	233
Простая классификация изображений .....	235
Вычисление признаков по изображению.....	236
Создание собственных признаков .....	237
Использование признаков для поиска похожих изображений .....	239
Классификация на более трудном наборе данных.....	241
Локальные представления признаков .....	242
Резюме .....	246
<b>Глава 11. Понижение размерности .....</b>	<b>248</b>
План действий.....	249
Отбор признаков .....	249
Выявление избыточных признаков с помощью фильтров .....	250
Применение оберток для задания модели вопросов о признаках.....	257
Другие методы отбора признаков .....	259
Выделение признаков .....	260
Об анализе главных компонент .....	260
Ограничения PCA и чем может помочь LDA .....	263
Многомерное шкалирование.....	264
Резюме .....	267
<b>Глава 12. Когда данных больше .....</b>	<b>269</b>
Что такое большие данные .....	269
Использование <code>jug</code> для построения конвейера задач .....	270
Введение в задачи <code>jug</code> .....	271

Заглянем под капот .....	273
Применение <code>jupyter</code> для анализа данных .....	275
Повторное использование частичных результатов .....	278
Работа с Amazon Web Services .....	279
Создание виртуальной машины .....	281
Установка Python-пакетов на Amazon Linux .....	285
Запуск <code>jupyter</code> на облачной машине .....	286
Автоматизированная генерация кластеров с помощью StarCluster .....	287
Резюме .....	291
<b>Где получить дополнительные сведения</b>	
<b>о машинном обучении .....</b>	<b>293</b>
Онлайновые курсы .....	293
Книги .....	293
Вопросно-ответные сайты .....	294
Блоги .....	294
Источники данных .....	295
Участие в конкурсах .....	295
Что не вошло в книгу .....	295
Резюме .....	296
<b>Предметный указатель .....</b>	<b>297</b>