

**Крис Элбон**

**Машинное обучение  
с использованием Python.  
Сборник рецептов**

Санкт-Петербург  
«БХВ-Петербург»  
2022

УДК 004.8+004.438Python  
ББК 32.973.26-018.1  
Э45

### Элбон Крис

Э45 Машинное обучение с использованием Python. Сборник рецептов:  
Пер. с англ. — СПб.: БХВ-Петербург, 2022. — 384 с.: ил.

ISBN 978-5-9775-4056-8

Книга содержит около 200 рецептов решения практических задач машинного обучения, таких как загрузка и обработка текстовых или числовых данных, отбор модели, уменьшение размерности и многие другие. Рассмотрена работа с языком Python и его библиотеками, в том числе pandas и scikit-learn. Решения всех задач сопровождаются подробными объяснениями. Каждый рецепт содержит работающий программный код, который можно вставлять, объединять и адаптировать, создавая собственное приложение.

Приведены рецепты решений с использованием: векторов, матриц и массивов; обработки данных, текста, изображений, дат и времени; уменьшения размерности и методов выделения или отбора признаков; оценивания и отбора моделей; линейной и логистической регрессии, деревьев, лесов и  $k$  ближайших соседей; опорно-векторных машин (SVM), наивных байесовых классификаторов, кластеризации и нейронных сетей; сохранения и загрузки натренированных моделей.

*Для разработчиков систем машинного обучения*

УДК 004.8+004.438Python  
ББК 32.973.26-018.1

#### Группа подготовки издания:

Руководитель проекта	<i>Евгений Рыбаков</i>
Зав. редакцией	<i>Екатерина Сависте</i>
Перевод с английского	<i>Андрея Логунова</i>
Компьютерная верстка	<i>Ольги Сергиенко</i>
Оформление обложки	<i>Карины Соловьевой</i>

© 2019 BHV

Authorized translation of the English edition of *Machine Learning with Python Cookbook*

ISBN 978-1-491-98938-8 © 2018 Chris Albon.

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 *Machine Learning with Python Cookbook*

ISBN 978-1-491-98938-8 © 2018 Chris Albon.

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

Подписано в печать 10.06.22.

Формат 70×100<sup>1/8</sup>. Печать офсетная. Усл. печ. л. 30,96.

Доп. тираж 600 экз. Заказ № 4235.

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

Отпечатано с готового оригинал-макета

ООО "Принт-М", 142300, М.О., г. Чехов, ул. Полиграфистов, д. 1

ISBN 978-1-491-98938-8 (англ.)  
ISBN 978-5-9775-4056-8 (рус.)

© 2018 Chris Albon

© Перевод на русский язык, оформление. ООО "БХВ-Петербург",  
ООО "БХВ", 2019, 2022

---

# Оглавление

<b>Об авторе.....</b>	<b>1</b>
<b>Предисловие .....</b>	<b>3</b>
Для кого предназначена книга .....	4
Для кого не предназначена книга.....	4
Терминология, используемая в книге .....	4
Признательности .....	5
Комментарии переводчика .....	5
Исходный код.....	7
Протокол установки библиотек.....	7
Установка библиотек Python из whl-файлов .....	8
Блокноты Jupyter.....	9
<b>Глава 1. Векторы, матрицы, массивы.....</b>	<b>11</b>
Введение.....	11
1.1. Создание вектора.....	11
1.2. Создание матрицы .....	12
1.3. Создание разреженной матрицы .....	13
1.4. Выбор элементов .....	14
1.5. Описание матрицы .....	16
1.6. Применение операций к элементам .....	17
1.7. Нахождение максимального и минимального значений.....	18
1.8. Вычисление среднего значения, дисперсии и стандартного отклонения.....	19
1.9. Реформирование массивов.....	20
1.10. Транспонирование вектора в матрицу.....	21
1.11. Сглаживание матрицы.....	22
1.12. Нахождение ранга матрицы.....	22
1.13. Вычисление определителя матрицы .....	23
1.14. Получение диагонали матрицы .....	24
1.15. Вычисление следа матрицы.....	24
1.16. Нахождение собственных значений и собственных векторов .....	25
1.17. Вычисление скалярных произведений .....	27
1.18. Сложение и вычитание матриц .....	28
1.19. Умножение матриц.....	29

1.20. Обращение матрицы.....	30
1.21. Генерирование случайных значений .....	31
<b>Глава 2. Загрузка данных .....</b>	<b>33</b>
Введение.....	33
2.1. Загрузка образца набора данных.....	33
2.2. Создание симулированного набора данных.....	35
2.3. Загрузка файла CSV .....	38
2.4. Загрузка файла Excel .....	39
2.5. Загрузка файла JSON.....	40
2.6. Опрашивание базы данных SQL .....	41
<b>Глава 3. Упорядочение данных .....</b>	<b>42</b>
Введение.....	42
3.1. Создание фрейма данных.....	43
3.2. Описание данных.....	44
3.3. Навигация по фреймам данных .....	46
3.4. Выбор строк на основе условных конструкций.....	48
3.5. Замена значений .....	49
3.6. Переименование столбцов .....	51
3.7. Нахождение минимума, максимума, суммы, среднего арифметического и количества.....	52
3.8. Нахождение уникальных значений.....	53
3.9. Отбор пропущенных значений.....	55
3.10. Удаление столбца .....	56
3.11. Удаление строки .....	58
3.12. Удаление повторяющихся строк .....	59
3.13. Группирование строк по значениям .....	61
3.14. Группирование строк по времени .....	62
3.15. Обход столбца в цикле .....	65
3.16. Применение функции ко всем элементам в столбце .....	66
3.17. Применение функции к группам.....	66
3.18. Конкатенация фреймов данных.....	67
3.19. Слияние фреймов данных .....	69
<b>Глава 4. Работа с числовыми данными .....</b>	<b>73</b>
Введение.....	73
4.1. Шкалирование признака .....	73
4.2. Стандартизация признака .....	75
4.3. Нормализация наблюдений .....	76
4.4. Генерирование полиномиальных и взаимодействующих признаков .....	78
4.5. Преобразование признаков.....	80
4.6. Обнаружение выбросов.....	81

4.7. Обработка выбросов .....	83
4.8. Дискретизация признаков .....	86
4.9. Группирование наблюдений с помощью кластеризации .....	87
4.10. Удаление наблюдений с пропущенными значениями .....	89
4.11. Импутация пропущенных значений .....	91
<b>Глава 5. Работа с категориальными данными .....</b>	<b>94</b>
Введение .....	94
5.1. Кодирование номинальных категориальных признаков .....	95
5.2. Кодирование порядковых категориальных признаков .....	98
5.3. Кодирование словарей признаков .....	100
5.4. Импутация пропущенных значений классов .....	102
5.5. Работа с несбалансированными классами .....	104
<b>Глава 6. Работа с текстом .....</b>	<b>109</b>
Введение .....	109
6.1. Очистка текста .....	109
6.2. Разбор и очистка разметки HTML .....	111
6.3. Удаление знаков препинания .....	112
6.4. Лексемизация текста .....	113
6.5. Удаление стоп-слов .....	114
6.6. Выделение основ слов .....	115
6.7. Лемматизация слов .....	116
6.8. Разметка слов на части речи .....	117
6.9. Кодирование текста в качестве мешка слов .....	120
6.10. Взвешивание важности слов .....	123
<b>Глава 7. Работа с датами и временем .....</b>	<b>126</b>
Введение .....	126
7.1. Конвертирование строковых значений в даты .....	126
7.2. Обработка часовых поясов .....	128
7.3. Выбор дат и времени .....	129
7.4. Разбиение данных даты на несколько признаков .....	130
7.5. Вычисление разницы между датами .....	131
7.6. Кодирование дней недели .....	132
7.7. Создание запаздывающего признака .....	133
7.8. Использование скользящих временных окон .....	134
7.9. Обработка пропущенных дат во временном ряду .....	136
<b>Глава 8. Работа с изображениями .....</b>	<b>139</b>
Введение .....	139
8.1. Загрузка изображений .....	140
8.2. Сохранение изображений .....	142

8.3. Изменение размера изображений.....	143
8.4. Обрезка изображений.....	144
8.5. Размытие изображений.....	146
8.6. Увеличение резкости изображений.....	148
8.7. Усиление контрастности.....	150
8.8. Выделение цвета.....	152
8.9. Бинаризация изображений.....	153
8.10. Удаление фонов.....	155
8.11. Обнаружение краев изображений.....	158
8.12. Обнаружение углов.....	159
8.13. Создание признаков для машинного самообучения.....	163
8.14. Кодирование среднего цвета в качестве признака.....	166
8.15. Кодирование гистограмм цветовых каналов в качестве признаков.....	167
<b>Глава 9. Снижение размерности с помощью выделения признаков.....</b>	<b>171</b>
Введение.....	171
9.1. Снижение признаков с помощью главных компонент.....	171
9.2. Уменьшение количества признаков, когда данные линейно неразделимы.....	174
9.3. Уменьшение количества признаков путем максимизации разделимости классов.....	176
9.4. Уменьшение количества признаков с использованием разложения матрицы.....	179
9.5. Уменьшение количества признаков на разреженных данных.....	180
<b>Глава 10. Снижение размерности с помощью отбора признаков.....</b>	<b>184</b>
Введение.....	184
10.1. Пороговая обработка дисперсии числовых признаков.....	184
10.2. Пороговая обработка дисперсии бинарных признаков.....	186
10.3. Обработка высококоррелированных признаков.....	187
10.4. Удаление нерелевантных признаков для классификации.....	189
10.5. Рекурсивное устранение признаков.....	192
<b>Глава 11. Оценивание моделей.....</b>	<b>195</b>
Введение.....	195
11.1. Перекрестная проверка моделей.....	195
11.2. Создание базовой регрессионной модели.....	199
11.3. Создание базовой классификационной модели.....	201
11.4. Оценивание предсказаний бинарного классификатора.....	203
11.5. Оценивание порогов бинарного классификатора.....	206
11.6. Оценивание предсказаний мультиклассового классификатора.....	210
11.7. Визуализация результативности классификатора.....	211
11.8. Оценивание регрессионных моделей.....	213
11.9. Оценивание кластеризующих моделей.....	215
11.10. Создание собственного оценочного метрического показателя.....	217
11.11. Визуализация эффекта размера тренировочного набора.....	219

11.12. Создание текстового отчета об оценочных метрических показателях.....	221
11.13. Визуализация эффекта значений гиперпараметра.....	222
<b>Глава 12. Отбор модели.....</b>	<b>226</b>
Введение.....	226
12.1. Отбор наилучших моделей с помощью исчерпывающего поиска.....	226
12.2. Отбор наилучших моделей с помощью рандомизированного поиска.....	229
12.3. Отбор наилучших моделей из нескольких обучающихся алгоритмов.....	231
12.4. Отбор наилучших моделей во время предобработки.....	233
12.5. Ускорение отбора модели с помощью распараллеливания.....	235
12.6. Ускорение отбора модели с помощью алгоритмически специализированных методов.....	236
12.7. Оценивание результативности после отбора модели.....	238
<b>Глава 13. Линейная регрессия.....</b>	<b>241</b>
Введение.....	241
13.1. Подгонка прямой.....	241
13.2. Обработка интерактивных эффектов.....	243
13.3. Подгонка нелинейной связи.....	245
13.4. Снижение дисперсии с помощью регуляризации.....	247
13.5. Уменьшение количества признаков с помощью лассо-регрессии.....	250
<b>Глава 14. Деревья и леса.....</b>	<b>252</b>
Введение.....	252
14.1. Тренировка классификационного дерева принятия решений.....	252
14.2. Тренировка регрессионного дерева принятия решений.....	254
14.3. Визуализация модели дерева принятия решений.....	255
14.4. Тренировка классификационного случайного леса.....	258
14.5. Тренировка регрессионного случайного леса.....	260
14.6. Идентификация важных признаков в случайных лесах.....	261
14.7. Отбор важных признаков в случайных лесах.....	263
14.8. Обработка несбалансированных классов.....	264
14.9. Управление размером дерева.....	266
14.10. Улучшение результативности с помощью бустинга.....	267
14.11. Оценивание случайных лесов с помощью ошибок внепакетных наблюдений.....	269
<b>Глава 15. К ближайших соседей.....</b>	<b>271</b>
Введение.....	271
15.1. Отыскание ближайших соседей наблюдения.....	271
15.2. Создание классификационной модели $k$ ближайших соседей.....	274
15.3. Идентификация наилучшего размера окрестности.....	276
15.4. Создание радиусного классификатора ближайших соседей.....	277

<b>Глава 16. Логистическая регрессия</b> .....	<b>279</b>
Введение.....	279
16.1. Тренировка бинарного классификатора.....	279
16.2. Тренировка мультиклассового классификатора.....	281
16.3. Снижение дисперсии с помощью регуляризации.....	282
16.4. Тренировка классификатора на очень крупных данных.....	283
16.5. Обработка несбалансированных классов.....	285
<b>Глава 17. Опорно-векторные машины</b> .....	<b>287</b>
Введение.....	287
17.1. Тренировка линейного классификатора.....	287
17.2. Обработка линейно неразделимых классов с помощью ядер.....	290
17.3. Создание предсказанных вероятностей.....	294
17.4. Идентификация опорных векторов.....	295
17.5. Обработка несбалансированных классов.....	297
<b>Глава 18. Наивный Байес</b> .....	<b>299</b>
Введение.....	299
18.1. Тренировка классификатора для непрерывных признаков.....	300
18.2. Тренировка классификатора для дискретных и счетных признаков.....	302
18.3. Тренировка наивного байесова классификатора для бинарных признаков.....	303
18.4. Калибровка предсказанных вероятностей.....	304
<b>Глава 19. Кластеризация</b> .....	<b>307</b>
Введение.....	307
19.1. Кластеризация с помощью $k$ средних.....	307
19.2. Ускорение кластеризации методом $k$ средних.....	310
19.3. Кластеризация методом сдвига к среднему.....	311
19.4. Кластеризация методом DBSCAN.....	313
19.5. Кластеризация методом иерархического слияния.....	314
<b>Глава 20. Нейронные сети</b> .....	<b>317</b>
Введение.....	317
20.1. Предобработка данных для нейронных сетей.....	318
20.2. Проектирование нейронной сети.....	320
20.3. Тренировка бинарного классификатора.....	323
20.4. Тренировка мультиклассового классификатора.....	325
20.5. Тренировка регрессора.....	327
20.6. Выполнение предсказаний.....	329
20.7. Визуализация истории процесса тренировки.....	331
20.8. Снижение перепогонки с помощью регуляризации весов.....	334
20.9. Снижение перепогонки с помощью ранней остановки.....	336
20.10. Снижение перепогонки с помощью отсева.....	338



20.11. Сохранение процесса тренировки модели .....	340
20.12. $k$ -блочная перекрестная проверка нейронных сетей .....	343
20.13. Тонкая настройка нейронных сетей.....	345
20.14. Визуализация нейронных сетей .....	347
20.15. Классификация изображений .....	349
20.16. Улучшение результативности с помощью расширения изображения.....	353
20.17. Классификация текста.....	355
<b>Глава 21. Сохранение и загрузка натренированных моделей .....</b>	<b>359</b>
Введение .....	359
21.1. Сохранение и загрузка модели scikit-learn .....	359
21.2. Сохранение и загрузка модели Keras .....	361
<b>Предметный указатель.....</b>	<b>363</b>