

O'REILLY®

2-е издание

Машинное обучение с использованием Python

Сборник рецептов

Практические решения
от предобработки
до глубокого обучения



elist

Кайл Галлатин,
Крис Элбон

Кайл Галлатин,
Крис Элбон

Машинное обучение с использованием Python

Сборник рецептов

Практические решения
от предобработки
до глубокого обучения

Астана
«АЛИСТ»
2024

УДК 004.4'236
ББК 32.988.02-018
Г16

Галлатин К.

Г16 Машинное обучение с использованием Python. Сборник рецептов:
Пер. с англ. / К. Галлатин, К. Элбон. — 2-е изд., перераб. и доп. — Астана:
АЛИСТ, 2024. — 448 с.: ил.

ISBN 978-601-08-4119-2

Книга содержит около 200 задач машинного обучения, таких как загрузка и обработка текстовых или числовых данных, отбор модели и многие другие. Рассмотрена работа с языком Python, библиотеками pandas и scikit-learn. Коды примеров можно вставлять, объединять и адаптировать, создавая собственное приложение. Приведены рецепты решений с использованием: векторов, матриц и массивов; данных из CSV, JSON, SQL, баз данных, облачных хранилищ и других источников; 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.

Во втором издании все примеры обновлены, рассмотрены задачи и фреймворки глубокого обучения, расширены разделы с тензорами, нейронными сетями и библиотекой глубокого обучения PyTorch.

Для разработчиков систем машинного обучения

УДК 004.4'236
ББК 32.988.02-018

© 2024 ALIST LLP

Authorized Russian translation of the English edition of *Machine Learning with Python Cookbook*,
2nd edition ISBN 9781098135720 © 2023 Kyle Gallatin.

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 с английского языка на русский издания *Machine Learning with Python Cookbook*,
2nd edition ISBN 9781098135720 © 2023 Kyle Gallatin.

Перевод опубликован и продается с разрешения компании-правообладателя O'Reilly Media, Inc.

ISBN 978-1-098-13572-0 (англ.)
ISBN 978-601-08-4119-2 (каз.)

© Kyle Gallatin, 2023
© Издание на русском языке. ТОО "АЛИСТ", 2024

Содержание

Введение.....	13
Условные обозначения	13
Использование примеров кода.....	14
Благодарности	14
 ГЛАВА 1. Работа с векторами, матрицами и массивами в NumPy.....	17
Введение	17
1.1. Создание вектора	17
1.2. Создание матрицы.....	18
1.3. Создание разреженной матрицы.....	19
1.4. Предварительное распределение массивов в NumPy	21
1.5. Обращение к элементу массива	22
1.6. Описание матрицы	24
1.7. Использование функций для каждого элемента	24
1.8. Поиск наибольших и наименьших значений.....	26
1.9. Вычисление среднего значения, дисперсии и стандартного отклонения	27
1.10. Изменение формы массива.....	28
1.11. Транспонирование вектора или матрицы	29
1.12. Преобразование матрицы в одномерный вектор	30
1.13. Определение ранга матрицы	31
1.14. Вывод диагонали матрицы	32
1.15. Вычисление суммы по диагонали матрицы	33
1.16. Вычисление скалярного произведения	34
1.17. Сложение и вычитание матриц.....	35
1.18. Умножение матриц	36
1.19. Создание обратной матрицы.....	37
1.20. Генерация псевдослучайных значений	38
 ГЛАВА 2. Загрузка данных.....	40
Введение	40
2.1. Загрузка готового набора данных.....	40

2.2. Создание искусственного набора данных.....	42
2.3. Загрузка CSV-файла.....	45
2.4. Загрузка файла Excel.....	47
2.5. Загрузка файла JSON	48
2.6. Загрузка файла Parquet	49
2.7. Загрузка файла Avro	50
2.8. Запрос к базе данных SQLite.....	51
2.9. Удаленное подключение к базе данных SQL	52
2.10. Загрузка данных из Google Sheet	53
2.11. Загрузка данных из S3 bucket.....	54
2.12. Загрузка неструктурированных данных	55

ГЛАВА 3. Первичная обработка данных 57

Введение	57
3.1. Создание DataFrame.....	58
3.2. Получение информации о данных.....	59
3.3. Срез DataFrame	61
3.4. Выбор строк на основе условных конструкций	64
3.5. Сортировка значений.....	65
3.6. Замена значений.....	66
3.7. Переименование столбцов	67
3.8. Нахождение наименьшего, наибольшего и среднего значения, суммы и количества значений	69
3.9. Поиск уникальных значений.....	70
3.10. Отбор отсутствующих значений	71
3.11. Удаление столбца.....	73
3.12. Удаление строки.....	75
3.13. Удаление дубликатов.....	76
3.14. Группировка строк по значениям.....	78
3.15. Группировка строк по времени.....	80
3.16. Агрегатные операции и статистика	82
3.17. Обход столбца в цикле	84
3.18. Применение функции ко всем элементам столбца	85
3.19. Применение функции к группам	86
3.20. Конкатенация структур данных DataFrame	87
3.21. Слияние структур данных DataFrame	88

ГЛАВА 4. Управление количественными данными 92

Введение	92
4.1. Масштабирование признака.....	92
4.2. Стандартизация признака.....	94
4.3. Нормализация наблюдаемых данных.....	96

4.4. Генерирование полиномиальных и взаимодействующих признаков	98
4.5. Преобразование признаков	100
4.6. Обнаружение выбросов	101
4.7. Управление выбросами	103
4.8. Дискретизация признаков	105
4.9. Группирование наблюдений при помощи кластеризации	107
4.10. Удаление элементов с отсутствующим значением	109
4.11. Восстановление пропущенных значений	111

ГЛАВА 5. Управление категориальными данными 114

Введение	114
5.1. Преобразование признаков с номинальными категориями	115
5.2. Преобразование признаков с порядковыми категориями	118
5.3. Преобразование словарей признаков	120
5.4. Восстановление классов пропущенных значений	122
5.5. Управление несбалансированными классами	124

ГЛАВА 6. Обработка текста 129

Введение	129
6.1. Очистка текста.....	129
6.2. Синтаксический анализ и очистка HTML-документов	132
6.3. Удаление знаков препинания.....	133
6.4. Токенизация текста	134
6.5. Удаление стоп-слов.....	135
6.6. Стемминг слов.....	136
6.7. Разметка частей речи	137
6.8. Распознавание именованных сущностей	139
6.9. Представление текста в виде мешка слов	141
6.10. Оценка значимости слов.....	143
6.11. Использование векторизации текстов для измерения соответствия текста поисковому запросу	145
6.12. Использование анализа тональности при классификации	147

ГЛАВА 7. Обработка данных даты и времени 149

Введение	149
7.1. Преобразование строк в даты	149
7.2. Работа с часовыми поясами	151
7.3. Выбор даты и времени.....	152
7.4. Разложение временных данных на отдельные признаки	153
7.5. Вычисление разницы между датами	154

7.6. Вычисление дней недели.....	155
7.7. Создание признаков с запаздыванием по времени	156
7.8. Использование метода скользящего окна.....	157
7.9. Обработка пропущенных значений во временных рядах	159

ГЛАВА 8. Работа с изображениями 162

Введение	162
8.1. Загрузка изображений.....	162
8.2. Сохранение изображений.....	165
8.3. Измерение размера изображений	166
8.4. Кадрирование изображений	167
8.5. Размытие изображений.....	168
8.6. Настройка резкости изображений	170
8.7. Увеличение контраста	171
8.8. Распознавание цветов	173
8.9. Бинаризация изображений	175
8.10. Удаление фона.....	177
8.11. Определение границ	180
8.12. Обнаружение углов.....	181
8.13. Создание признаков для машинного обучения.....	185
8.14. Создание гистограмм цветов в виде признаков	187
8.15. Использование предобученных эмбеддингов в качестве признаков	190
8.16. Распознавание объектов при помощи OpenCV	193
8.17. Классификация изображений при помощи Pytorch.....	194

ГЛАВА 9. Снижение размерности через выделение признаков 197

Введение	197
9.1. Сокращение признаков при помощи главных компонент	198
9.2. Сокращение признаков у линейно неразделимых данных.....	200
9.3. Сокращение признаков при помощи повышения разделимости классов.....	202
9.4. Сокращение признаков при помощи разложения матрицы	205
9.5. Сокращение признаков разреженных данных.....	207

ГЛАВА 10. Снижение размерности через отбор признаков..... 210

Введение	210
10.1. Пороговый отбор числовых признаков дисперсии.....	211
10.2. Пороговый отбор дисперсии бинарных признаков	212
10.3. Обработка признаков с высоким коэффициентом корреляции	214
10.4. Удаление нерелевантных для классификации признаков	215
10.5. Рекурсивное исключение признаков.....	218

ГЛАВА 11. Оценка модели	221
Введение	221
11.1. Модели на основе кросс-валидации	221
11.2. Создание базовой регрессионной модели.....	225
11.3. Создание базовой модели классификации.....	227
11.4. Оценка предсказаний бинарного классификатора.....	229
11.5. Оценка порогов бинарного классификатора	232
11.6. Оценка прогнозов многоклассового классификатора	236
11.7. Визуализация показателей качества классификации.....	237
11.8. Оценка регрессионных моделей	240
11.9. Оценки моделей кластеризации.....	242
11.10. Создание собственной оценочной метрики.....	244
11.11. Визуализация зависимости показателей от размера обучающего набора	245
11.12. Создание текстового отчета с метриками качества	248
11.13. Визуализация влияния значений гиперпараметров	249
 ГЛАВА 12. Выбор модели.....	 253
Введение	253
12.1. Поиск лучших моделей методом полного перебора.....	254
Обсуждение	255
12.2. Случайный поиск лучших моделей.....	256
12.3. Поиск лучшей модели среди множества алгоритмов обучения	258
12.4. Поиск лучшей модели в процессе первичной обработки.....	260
12.5. Ускорение поиска модели через распараллеливание	263
12.6. Ускорение поиска модели с помощью алгоритмических методов	264
12.7. Оценка качества после выбора модели	266
 ГЛАВА 13. Линейная регрессия	 269
Введение	269
13.1. Настройка линейной связи	269
13.2. Управление взаимосвязанными коэффициентами.....	271
13.3. Настройка нелинейной взаимосвязи	273
13.4. Уменьшение дисперсии через регуляризацию.....	276
13.5. Уменьшение дисперсии через лассо-регрессию	278
 ГЛАВА 14. Деревья и леса.....	 280
Введение	280
14.1. Обучение классификатора на основе деревьев решений	280
14.2. Обучение регрессора с помощью деревьев решений	282
14.3. Визуализация модели деревьев решений.....	284

14.4. Обучение классификатора на основе случайного леса.....	286
14.5. Обучение регрессора методом случайного леса	288
14.6. Оценка моделей случайного леса методом out-of-bag-ошибки	289
14.7. Определение ключевых признаков в моделях случайного леса.....	290
14.8. Выбор ключевых признаков в моделях случайного леса.....	292
14.9. Управление несбалансированными классами	294
14.10. Контроль размера дерева.....	295
14.11. Улучшение качества через бустинг	297
14.12. Обучение модели на основе алгоритма XGBoost	298
14.13. Улучшение производительности при помощи LightGBM	300

ГЛАВА 15. Метод k ближайших соседей..... 303

Введение	303
15.1. Определение ближайших соседей для наблюдения.....	303
15.2. Создание классификатора на основе k ближайших соседей	306
15.3. Определение оптимального количества ближайших соседей	308
15.4. Создание классификатора методом ближайших соседей в заданном радиусе	309
15.5. Поиск приближенных ближайших соседей.....	310
15.6. Оценка метода приближенных ближайших соседей	314

ГЛАВА 16. Логистическая регрессия..... 316

Введение	316
16.1. Обучение бинарного классификатора.....	316
16.2. Обучение многоклассового классификатора.....	318
16.3. Уменьшение дисперсии за счет регуляризации	319
16.4. Обучение классификатора на очень больших данных	320
16.5. Управление несбалансированными классами	322

ГЛАВА 17. Метод опорных векторов..... 324

Введение	324
17.1. Обучение линейного классификатора.....	324
17.2. Управление линейно неразделимыми классами через ядра свертки	327
17.3. Создание прогнозируемых вероятностей	331
17.4. Поиск опорных векторов.....	332
17.5. Управление несбалансированными классами	334

ГЛАВА 18. Наивный байесовский классификатор 336

Введение	336
18.1. Обучение классификатора для непрерывных признаков	337

18.2. Обучение классификатора для дискретных и количественных признаков	339
18.3. Обучение наивного байесовского классификатора для бинарных признаков	341
18.4. Калибровка прогнозируемых вероятностей	342

ГЛАВА 19. Кластеризация 344

Введение	344
19.1. Кластеризация методом k средних	344
19.2. Ускорение кластеризации методом k средних	347
19.3. Кластеризация со сдвигом среднего.....	348
19.4. Кластеризация DBSCAN	350
19.5. Иерархическая кластеризация.....	351

ГЛАВА 20. Тензоры в PyTorch 354

Введение	354
20.1. Создание тензора.....	354
20.2. Создание тензора из массива NumPy	355
20.3. Создание разреженного тензора	356
20.4. Выбор элементов тензора.....	357
20.5. Описание тензора	359
20.6. Операции над элементами.....	360
20.7. Поиск наибольших и наименьших значений.....	361
20.8. Изменение формы тензоров	362
20.9. Транспонирование тензора	363
20.10. Преобразование в одномерный тензор	364
20.11. Вычисление скалярного произведения	364
20.12. Умножение тензоров	365

ГЛАВА 21. Нейронные сети 367

Введение	367
21.1. Использование autograd в PyTorch	368
21.2. Предобработка данных для нейронных сетей	370
21.3. Разработка нейронной сети	372
21.4. Обучение бинарного классификатора.....	376
21.5. Обучение многоклассового классификатора.....	379
21.6. Обучение регрессора	382
21.7. Создание прогнозов	384
21.8. Визуализация истории обучения	387
21.9. Регуляризация весов для снижения переобучения	390
21.10. Снижение переобучения через раннюю остановку	392
21.11. Уменьшение эффекта переобучения методом исключения	396

21.12. Сохранение прогресса обучения модели	399
21.13. Настройка нейронных сетей	402
21.14. Визуализация нейронной сети	405
ГЛАВА 22. Нейронные сети с неструктурированными данными	409
Введение	409
22.1. Обучение нейронной сети для классификации изображений	410
22.2. Обучение нейронной сети для классификации текстов	413
22.3. Тонкая настройка предобученной модели для классификации изображений	415
22.4. Тонкая настройка предобученной модели для классификации текста	418
ГЛАВА 23. Сохранение, загрузка и развертывание обученных моделей	422
Введение	422
23.1. Сохранение и загрузка модели scikit-learn	422
23.2. Сохранение и загрузка модели TensorFlow	424
23.3. Сохранение и загрузка модели PyTorch	425
23.4. Развертывание моделей scikit-learn	428
23.5. Развертывание моделей TensorFlow	430
23.6. Развертывание моделей PyTorch через Seldon	433
Предметный указатель	438
Об авторах	446
Об изображении на обложке	447