

УДК 004.438Python:004.6
ББК 32.973.22
С33

Шарден Б., Массарон Л., Боскетти А.

Ш25 Крупномасштабное машинное обучение вместе с Python / пер. с англ.
А. В. Логунова. – М.: ДМК Пресс, 2018. – 358 с.: ил.

ISBN978-5-97060-506-6

Главная задача настоящей книги состоит в том, чтобы предоставить способы применения мощных методов машинного обучения с открытым исходным кодом в крупномасштабных проектах без привлечения дорогостоящих корпоративных решений или больших вычислительных кластеров. Описаны масштабируемое обучение в Scikit-learn, нейронные сети и глубокое обучение с использованием Theano, H2O и TensorFlow. Рассмотрены классификационные и регрессионные деревья, а также обучение без учителя. Охвачены эффективные методы машинного обучения в вычислительной среде MapReduce на платформах Hadoop и Spark на языке Python.

УДК 004.438Python:004.6
ББК 32.973.22

First published in the English language under the title 'Large Scale Machine Learning with Python – (9781785887215).

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.

ISBN 978-1-78588-721-5 (англ.)
ISBN 978-5-97060-506-6 (рус.)

Copyright © 2016 Packt Publishing
© Оформление, издание, перевод, ДМК Пресс, 2018

Содержание

Об авторах	9
О рецензентах	10
Предисловие	11
Глава 1. Первые шаги к масштабируемости	20
Подробное объяснение термина масштабируемости	21
Приведение крупномасштабных примеров	23
Введение в язык Python	24
Вертикальное масштабирование средствами Python.....	25
Горизонтальное масштабирование средствами Python	26
Python для крупномасштабного машинного обучения	27
Выбор между Python 2 и Python 3	27
Инсталляция среды Python.....	28
Пошаговая установка	28
Установка библиотек.....	29
Способы обновления библиотек	31
Научные дистрибутивы	32
Введение в Jupyter	33
Библиотеки Python	37
NumPy.....	37
SciPy.....	37
Pandas.....	37
Scikit-learn.....	38
Резюме.....	44
Глава 2. Масштабируемое обучение в Scikit-learn	46
Внеядерное обучение	47
Подвыборка как приемлемый вариант	48
Оптимизация по одному прецеденту за раз	48
Создание системы внеядерного обучения	50
Потоковая передача данных из источников.....	51
Наборы данных для реальных дел	51
Первый пример – потоковая передача набора данных Bike-sharing.....	54
Использование инструментов ввода-вывода библиотеки pandas.....	56
Работа с базами данных.....	57
Особое внимание упорядочению прецедентов	61
Стохастическое обучение.....	63
Пакетный градиентный спуск	64
Стохастический градиентный спуск.....	67
Реализация алгоритма SGD в библиотеке Scikit-learn	68
Определение параметров обучения алгоритма SGD	70
Управление признаками на потоках данных	72
Описание целевой переменной	76

Хэширование признаков	79
Другие элементарные преобразования	82
Тестирование и перекрестная проверка в потоке	83
Применение алгоритма SGD в деле	84
Резюме	88
Глава 3. Быстрообучающиеся реализации машин SVM	89
Наборы данных для самостоятельного экспериментирования	90
Набор данных Bike-sharing	90
Набор данных Covertypes	91
Машины опорных векторов	91
Кусочно-линейная функция потерь и ее варианты	97
Объяснение реализации алгоритма SVM в Scikit-learn	98
Поиск нелинейных SVM с привлечением подвыборки	101
Реализация SVM в крупном масштабе на основе SGD	104
Отбор признаков посредством регуляризации	112
Добавление нелинейности в алгоритм SGD	114
Испытание явных высокоразмерных отображений	115
Доводка гиперпараметров	117
Другие альтернативы быстро обучающихся реализаций SVM	121
Резюме	133
Глава 4. Искусственные нейронные сети и глубокое обучение	134
Архитектура нейронной сети	135
Чему и как нейронные сети обучаются	144
Выбор правильной архитектуры	148
Нейронные сети в действии	149
Параллелизация для библиотеки sknn	150
Нейронные сети и регуляризация	151
Нейронные сети и гиперпараметрическая оптимизация	153
Нейронные сети и границы решения	154
Глубокое обучение в крупном масштабе с H2O	157
Крупномасштабное глубокое обучение с H2O	158
Сеточный поиск в H2O	161
Глубокое обучение и предтренировка без учителя	162
Глубокое обучение с theano	162
Автокодировщики и обучение без учителя	164
Автокодировщик	164
Резюме	168
Глава 5. Глубокое обучение с библиотекой TensorFlow	170
Инсталляция TensorFlow	172
Операции TensorFlow	172
Машинное обучение в TensorFlow посредством SkFlow	177
Глубокое обучение с большими файлами – инкрементное обучение	183
Инсталляция библиотеки Keras и платформа TensorFlow	186

Сверточные нейронные сети в TensorFlow посредством Keras	190
Сверточный слой	192
Объединяющий слой	193
Полносвязный слой	194
CNN-сети с подходом на основе инкрементной тренировки	195
Вычисления на GPU	196
Резюме	199

Глава 6. Классификационные и регрессионные деревья

в крупном масштабе	200
Агрегация бутстрапированных выборок	203
Случайный лес и экстремально рандомизированный лес	204
Быстрая параметрическая оптимизация посредством рандомизированного поиска	208
Экстремально рандомизированные деревья и большие наборы данных	210
Алгоритм CART и бустинг	214
Машины градиентного бустинга	214
Алгоритм XGBoost	221
Регрессия на основе XGBoost	224
Потоковая передача больших наборов данных посредством XGBoost	227
Персистентность модели XGBoost	228
Внеядерный алгоритм CART в среде H2O	229
Случайный лес и сеточный поиск в H2O	229
Стохастический градиентный бустинг и сеточный поиск в H2O	231
Резюме	233

Глава 7. Обучение без учителя в крупном масштабе

Методы машинного обучения без учителя	235
Разложение признаков – PCA	236
Алгоритм PCA в среде H2O	246
Кластеризация – алгоритм К-средних	247
Методы инициализации	250
Допущения алгоритма К-средних	251
Подбор оптимальной величины К	253
Масштабирование алгоритма К-средних – мини-пакет	257
Алгоритм К-средних в среде H2O	261
Алгоритм LDA	263
Масштабирование алгоритма LDA – оперативная память, CPU и машины	271
Резюме	272

Глава 8. Распределенные среды – Hadoop и Spark

От автономной машины к набору узлов	273
Зачем нужна распределенная платформа?	275
Настройка виртуальной машины	276
Виртуализатор VirtualBox	277
Конфигуратор Vagrant	279

8 ❖ Содержание

Использование виртуальной машины.....	279
Экосистема Hadoop.....	281
Архитектура.....	281
Распределенная файловая система HDFS.....	282
Вычислительная парадигма MapReduce.....	289
Менеджер ресурсов YARN.....	298
Платформа Spark.....	299
Библиотека pySpark.....	299
Резюме.....	309
Глава 9. Практическое машинное обучение в среде Spark.....	310
Настройка виртуальной машины для данной главы.....	310
Распространение переменных по всем узлам кластера.....	311
Широковещательные переменные только для чтения.....	311
Аккумуляторные переменные только для записи.....	313
Широковещательные и аккумуляторные переменные – пример.....	314
Предобработка данных в среде Spark.....	316
Файлы JSON и объекты DataFrame платформы Spark.....	317
Работа с пропущенными данными.....	319
Группирование и создание таблиц в оперативной памяти.....	320
Запись предобработанного объекта DataFrame или RDD-набора на диск.....	322
Работа с объектами DataFrame.....	323
Машинное обучение с платформой Spark.....	326
Платформа Spark на наборе данных KDD99.....	326
Чтение набора данных.....	327
Конструирование признаков.....	329
Тренировка ученика.....	334
Оценка результативности ученика.....	335
Возможности конвейера машинного обучения.....	336
Ручная доводка.....	338
Перекрестная проверка.....	340
Заключительная очистка.....	342
Резюме.....	342
Приложение. Введение в графические процессоры и платформа Theano.....	344
Вычисления на GPU.....	344
Платформа Theano – параллельные вычисления на GPU.....	346
Установка платформы Theano.....	347
Предметный указатель.....	350