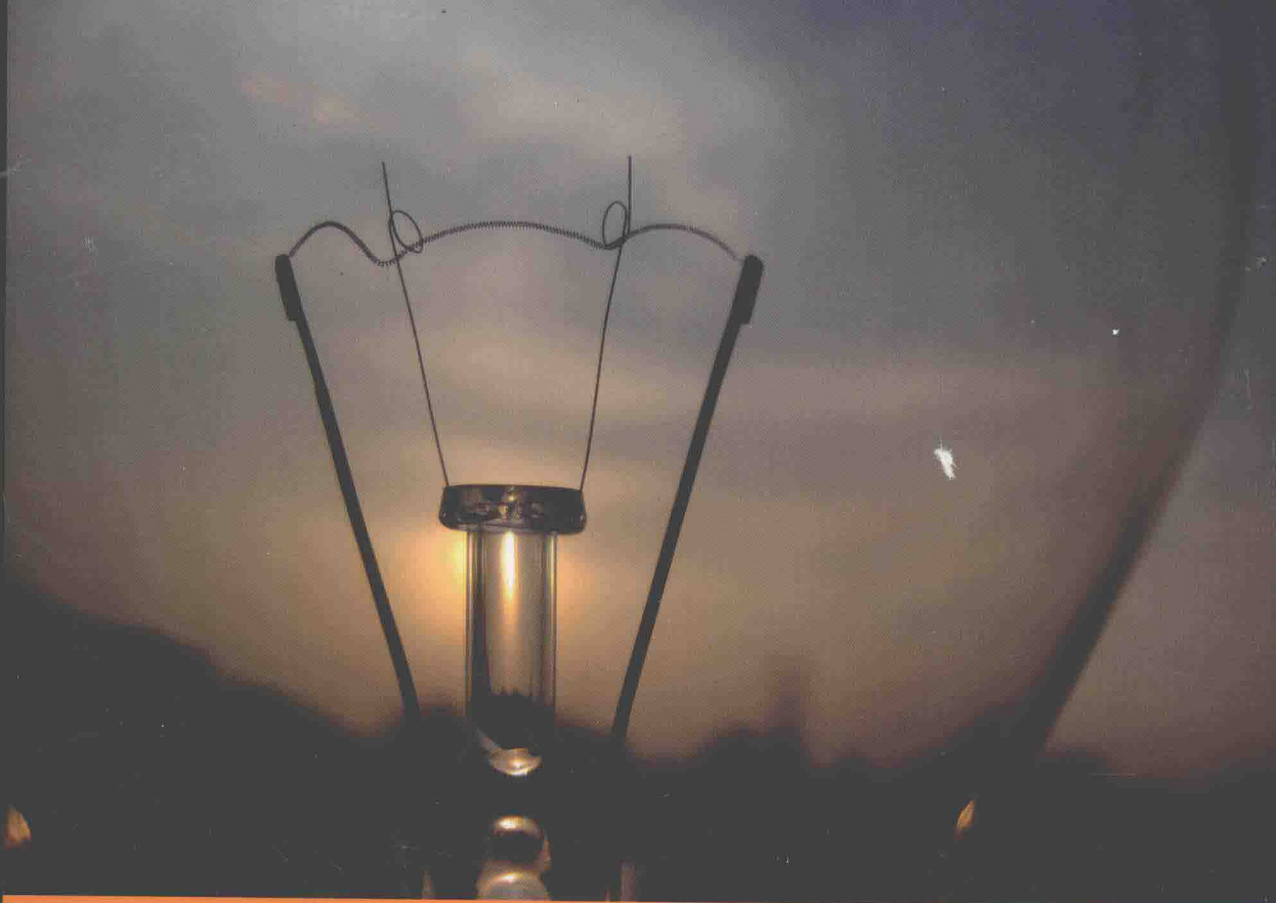

Spark机器学习（影印版）

Machine Learning with Spark

ick Pentreath 著

[PACKT] PUBLISHING

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Spark 机器学习(影印版)

Nick Pentreath 著

南京　东南大学出版社

图书在版编目(CIP)数据

Spark 机器学习:英文/(英)彭特里思(Pentreath,N.)著.—影印本.—南京:东南大学出版社,2016.1

书名原文:Machine Learning with Spark

ISBN 978-7-5641-6091-3

Ⅰ.①S… Ⅱ.①彭… Ⅲ.①数据处理软件—机器学习—英文 Ⅳ.①TP274②TP181

中国版本图书馆 CIP 数据核字(2015)第 256607 号

© 2015 by PACKT Publishing Ltd

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2016. Authorized reprint of the original English edition, 2015 PACKT Publishing Ltd, the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2015。

英文影印版由东南大学出版社出版 2016。此影印版的出版和销售得到出版权和销售权的所有者—— PACKT Publishing Ltd 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

Spark 机器学习(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼 2 号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787 毫米×980 毫米　　16 开本
印　　张:21
字　　数:411 千字
版　　次:2016 年 1 月第 1 版
印　　次:2016 年 1 月第 1 次印刷
书　　号:ISBN 978-7-5641-6091-3
定　　价:68.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Credits

Author
Nick Pentreath

Reviewers
Andrea Mostosi
Hao Ren
Krishna Sankar

Commissioning Editor
Rebecca Youé

Acquisition Editor
Rebecca Youé

Content Development Editor
Susmita Sabat

Technical Editors
Vivek Arora
Pankaj Kadam

Copy Editor
Karuna Narayanan

Project Coordinator
Milton Dsouza

Proofreaders
Simran Bhogal
Maria Gould
Ameesha Green
Paul Hindle

Indexer
Priya Sane

Graphics
Sheetal Aute
Abhinash Sahu

Production Coordinator
Nitesh Thakur

Cover Work
Nitesh Thakur

About the Author

Nick Pentreath has a background in financial markets, machine learning, and software development. He has worked at Goldman Sachs Group, Inc.; as a research scientist at the online ad targeting start-up Cognitive Match Limited, London; and led the Data Science and Analytics team at Mxit, Africa's largest social network.

He is a cofounder of Graphflow, a big data and machine learning company focused on user-centric recommendations and customer intelligence. He is passionate about combining commercial focus with machine learning and cutting-edge technology to build intelligent systems that learn from data to add value to the bottom line.

Nick is a member of the Apache Spark Project Management Committee.

此为试读，需要完整PDF请访问：www.ertongbook.com

Acknowledgments

Writing this book has been quite a rollercoaster ride over the past year, with many ups and downs, late nights, and working weekends. It has also been extremely rewarding to combine my passion for machine learning with my love of the Apache Spark project, and I hope to bring some of this out in this book.

I would like to thank the Packt Publishing team for all their assistance throughout the writing and editing process: Rebecca, Susmita, Sudhir, Amey, Neil, Vivek, Pankaj, and everyone who worked on the book.

Thanks also go to Debora Donato at StumbleUpon for assistance with data- and legal-related queries.

Writing a book like this can be a somewhat lonely process, so it is incredibly helpful to get the feedback of reviewers to understand whether one is headed in the right direction (and what course adjustments need to be made). I'm deeply grateful to Andrea Mostosi, Hao Ren, and Krishna Sankar for taking the time to provide such detailed and critical feedback.

I could not have gotten through this project without the unwavering support of all my family and friends, especially my wonderful wife, Tammy, who will be glad to have me back in the evenings and on weekends once again. Thank you all!

Finally, thanks to all of you reading this; I hope you find it useful!

About the Reviewers

Andrea Mostosi is a technology enthusiast. An innovation lover since he was a child, he started a professional job in 2003 and worked on several projects, playing almost every role in the computer science environment. He is currently the CTO at The Fool, a company that tries to make sense of web and social data. During his free time, he likes traveling, running, cooking, biking, and coding.

> I would like to thank my geek friends: Simone M, Daniele V, Luca T, Luigi P, Michele N, Luca O, Luca B, Diego C, and Fabio B. They are the smartest people I know, and comparing myself with them has always pushed me to be better.

Hao Ren is a software developer who is passionate about Scala, distributed systems, machine learning, and Apache Spark. He was an exchange student at EPFL when he learned about Scala in 2012. He is currently working in Paris as a backend and data engineer for ClaraVista—a company that focuses on high-performance marketing. His work responsibility is to build a Spark-based platform for purchase prediction and a new recommender system.

Besides programming, he enjoys running, swimming, and playing basketball and badminton. You can learn more at his blog `http://www.invkrh.me`.

Krishna Sankar is a chief data scientist at BlackArrow, where he is focusing on enhancing user experience via inference, intelligence, and interfaces. Earlier stints include working as a principal architect and data scientist at Tata America International Corporation, director of data science at a bioinformatics start-up company, and as a distinguished engineer at Cisco Systems, Inc. He has spoken at various conferences about data science (`http://goo.gl/9pyJMH`), machine learning (`http://goo.gl/sSem2Y`), and social media analysis (`http://goo.gl/D9YpVQ`). He has also been a guest lecturer at the Naval Postgraduate School. He has written a few books on Java, wireless LAN security, Web 2.0, and now on Spark. His other passion is LEGO robotics. Earlier in April, he was at the St. Louis FLL World Competition as a robots design judge.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

PACKTLIB™

https://www2.packtpub.com/books/subscription/packtlib

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at www.PacktPub.com, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface 1

Chapter 1: Getting Up and Running with Spark 7

Installing and setting up Spark locally 8

Spark clusters 10

The Spark programming model 11

SparkContext and SparkConf 11

The Spark shell 12

Resilient Distributed Datasets 14

Creating RDDs 15

Spark operations 15

Caching RDDs 18

Broadcast variables and accumulators 19

The first step to a Spark program in Scala 21

The first step to a Spark program in Java 24

The first step to a Spark program in Python 28

Getting Spark running on Amazon EC2 30

Launching an EC2 Spark cluster 31

Summary 35

Chapter 2: Designing a Machine Learning System 37

Introducing MovieStream 38

Business use cases for a machine learning system 39

Personalization 40

Targeted marketing and customer segmentation 40

Predictive modeling and analytics 41

Types of machine learning models 41

The components of a data-driven machine learning system 42

Data ingestion and storage 42

Data cleansing and transformation 43

Model training and testing loop 45
Model deployment and integration 45
Model monitoring and feedback 45
Batch versus real time 47
An architecture for a machine learning system **48**
Practical exercise 49
Summary **50**
Chapter 3: Obtaining, Processing, and Preparing Data with Spark **51**
Accessing publicly available datasets **52**
The MovieLens 100k dataset 54
Exploring and visualizing your data **55**
Exploring the user dataset 57
Exploring the movie dataset 62
Exploring the rating dataset 64
Processing and transforming your data **68**
Filling in bad or missing data 69
Extracting useful features from your data **70**
Numerical features 71
Categorical features 71
Derived features 73
Transforming timestamps into categorical features 73
Text features 75
Simple text feature extraction 76
Normalizing features 80
Using MLlib for feature normalization 81
Using packages for feature extraction 82
Summary **82**
Chapter 4: Building a Recommendation Engine with Spark **83**
Types of recommendation models **84**
Content-based filtering 85
Collaborative filtering 85
Matrix factorization 86
Extracting the right features from your data **92**
Extracting features from the MovieLens 100k dataset 92
Training the recommendation model **96**
Training a model on the MovieLens 100k dataset 96
Training a model using implicit feedback data 98
Using the recommendation model **99**
User recommendations 99
Generating movie recommendations from the MovieLens 100k dataset 99

此为试读，需要完整PDF请访问：www.ertongbook.com

Item recommendations 102
Generating similar movies for the MovieLens 100k dataset 103
Evaluating the performance of recommendation models 106
Mean Squared Error 107
Mean average precision at K 109
Using MLlib's built-in evaluation functions 113
RMSE and MSE 113
MAP 113
Summary 115
Chapter 5: Building a Classification Model with Spark 117
Types of classification models 120
Linear models 120
Logistic regression 122
Linear support vector machines 123
The naïve Bayes model 124
Decision trees 126
Extracting the right features from your data 128
Extracting features from the Kaggle/StumbleUpon evergreen classification dataset 128
Training classification models 130
Training a classification model on the Kaggle/StumbleUpon evergreen classification dataset 131
Using classification models 133
Generating predictions for the Kaggle/StumbleUpon evergreen classification dataset 133
Evaluating the performance of classification models 134
Accuracy and prediction error 134
Precision and recall 136
ROC curve and AUC 138
Improving model performance and tuning parameters 140
Feature standardization 141
Additional features 144
Using the correct form of data 147
Tuning model parameters 148
Linear models 149
Decision trees 154
The naïve Bayes model 155
Cross-validation 156
Summary 159

Chapter 6: Building a Regression Model with Spark 161
Types of regression models 162
Least squares regression 162
Decision trees for regression 163
Extracting the right features from your data 164
Extracting features from the bike sharing dataset 164
Creating feature vectors for the linear model 168
Creating feature vectors for the decision tree 169
Training and using regression models 170
Training a regression model on the bike sharing dataset 171
Evaluating the performance of regression models 173
Mean Squared Error and Root Mean Squared Error 173
Mean Absolute Error 174
Root Mean Squared Log Error 174
The R-squared coefficient 175
Computing performance metrics on the bike sharing dataset 175
Linear model 175
Decision tree 176
Improving model performance and tuning parameters 177
Transforming the target variable 177
Impact of training on log-transformed targets 180
Tuning model parameters 183
Creating training and testing sets to evaluate parameters 183
The impact of parameter settings for linear models 184
The impact of parameter settings for the decision tree 192
Summary 195
Chapter 7: Building a Clustering Model with Spark 197
Types of clustering models 198
K-means clustering 198
Initialization methods 202
Variants 203
Mixture models 203
Hierarchical clustering 203
Extracting the right features from your data 204
Extracting features from the MovieLens dataset 204
Extracting movie genre labels 205
Training the recommendation model 207
Normalization 207
Training a clustering model 208
Training a clustering model on the MovieLens dataset 208
Making predictions using a clustering model 210
Interpreting cluster predictions on the MovieLens dataset 211
Interpreting the movie clusters 212

Evaluating the performance of clustering models **216**
Internal evaluation metrics 216
External evaluation metrics 216
Computing performance metrics on the MovieLens dataset 217
Tuning parameters for clustering models **217**
Selecting K through cross-validation 217
Summary **219**
Chapter 8: Dimensionality Reduction with Spark **221**
Types of dimensionality reduction **222**
Principal Components Analysis 222
Singular Value Decomposition 223
Relationship with matrix factorization 224
Clustering as dimensionality reduction 224
Extracting the right features from your data **225**
Extracting features from the LFW dataset 225
Exploring the face data 226
Visualizing the face data 228
Extracting facial images as vectors 229
Normalization 233
Training a dimensionality reduction model **234**
Running PCA on the LFW dataset 235
Visualizing the Eigenfaces 236
Interpreting the Eigenfaces 238
Using a dimensionality reduction model **238**
Projecting data using PCA on the LFW dataset 239
The relationship between PCA and SVD 240
Evaluating dimensionality reduction models **242**
Evaluating k for SVD on the LFW dataset 242
Summary **245**
Chapter 9: Advanced Text Processing with Spark **247**
What's so special about text data? **247**
Extracting the right features from your data **248**
Term weighting schemes 248
Feature hashing 249
Extracting the TF-IDF features from the 20 Newsgroups dataset 251
Exploring the 20 Newsgroups data 253
Applying basic tokenization 255
Improving our tokenization 256
Removing stop words 258
Excluding terms based on frequency 261
A note about stemming 264
Training a TF-IDF model 264
Analyzing the TF-IDF weightings 266

Using a TF-IDF model **268**
Document similarity with the 20 Newsgroups dataset and TF-IDF features 268
Training a text classifier on the 20 Newsgroups dataset using TF-IDF 271
Evaluating the impact of text processing **273**
Comparing raw features with processed TF-IDF features on the 20 Newsgroups dataset 273
Word2Vec models **274**
Word2Vec on the 20 Newsgroups dataset 275
Summary **278**
Chapter 10: Real-time Machine Learning with Spark Streaming **279**
Online learning **279**
Stream processing **281**
An introduction to Spark Streaming 281
Input sources 282
Transformations 282
Actions 284
Window operators 284
Caching and fault tolerance with Spark Streaming 285
Creating a Spark Streaming application **286**
The producer application 287
Creating a basic streaming application 290
Streaming analytics 293
Stateful streaming 296
Online learning with Spark Streaming **298**
Streaming regression 298
A simple streaming regression program 299
Creating a streaming data producer 299
Creating a streaming regression model 302
Streaming K-means 305
Online model evaluation **306**
Comparing model performance with Spark Streaming 306
Summary **310**
Index **311**

Preface

In recent years, the volume of data being collected, stored, and analyzed has exploded, in particular in relation to the activity on the Web and mobile devices, as well as data from the physical world collected via sensor networks. While previously large-scale data storage, processing, analysis, and modeling was the domain of the largest institutions such as Google, Yahoo!, Facebook, and Twitter, increasingly, many organizations are being faced with the challenge of how to handle a massive amount of data.

When faced with this quantity of data and the common requirement to utilize it in real time, human-powered systems quickly become infeasible. This has led to a rise in the so-called big data and machine learning systems that learn from this data to make automated decisions.

In answer to the challenge of dealing with ever larger-scale data without any prohibitive cost, new open source technologies emerged at companies such as Google, Yahoo!, Amazon, and Facebook, which aimed at making it easier to handle massive data volumes by distributing data storage and computation across a cluster of computers.

The most widespread of these is Apache Hadoop, which made it significantly easier and cheaper to both store large amounts of data (via the Hadoop Distributed File System, or HDFS) and run computations on this data (via Hadoop MapReduce, a framework to perform computation tasks in parallel across many nodes in a computer cluster).

However, MapReduce has some important shortcomings, including high overheads to launch each job and reliance on storing intermediate data and results of the computation to disk, both of which make Hadoop relatively ill-suited for use cases of an iterative or low-latency nature. Apache Spark is a new framework for distributed computing that is designed from the ground up to be optimized for low-latency tasks and to store intermediate data and results in memory, thus addressing some of the major drawbacks of the Hadoop framework. Spark provides a clean, functional, and easy-to-understand API to write applications and is fully compatible with the Hadoop ecosystem.

Furthermore, Spark provides native APIs in Scala, Java, and Python. The Scala and Python APIs allow all the benefits of the Scala or Python language, respectively, to be used directly in Spark applications, including using the relevant interpreter for real-time, interactive exploration. Spark itself now provides a toolkit (called MLlib) of distributed machine learning and data mining models that is under heavy development and already contains high-quality, scalable, and efficient algorithms for many common machine learning tasks, some of which we will delve into in this book.

Applying machine learning techniques to massive datasets is challenging, primarily because most well-known machine learning algorithms are not designed for parallel architectures. In many cases, designing such algorithms is not an easy task. The nature of machine learning models is generally iterative, hence the strong appeal of Spark for this use case. While there are many competing frameworks for parallel computing, Spark is one of the few that combines speed, scalability, in-memory processing, and fault tolerance with ease of programming and a flexible, expressive, and powerful API design.

Throughout this book, we will focus on real-world applications of machine learning technology. While we may briefly delve into some theoretical aspects of machine learning algorithms, the book will generally take a practical, applied approach with a focus on using examples and code to illustrate how to effectively use the features of Spark and MLlib, as well as other well-known and freely available packages for machine learning and data analysis, to create a useful machine learning system.

What this book covers

Chapter 1, Getting Up and Running with Spark, shows how to install and set up a local development environment for the Spark framework as well as how to create a Spark cluster in the cloud using Amazon EC2. The Spark programming model and API will be introduced, and a simple Spark application will be created using each of Scala, Java, and Python.

此为试读，需要完整PDF请访问：www.ertongbook.com