



C o m m u n i t y E x p e r i e n c e D i s t i l l e d

Apache Mahout Essentials

Implement top-notch machine learning algorithms for classification, clustering, and recommendations with Apache Mahout



Jayani Withanawasam

[PACKT] open source*
PUBLISHING community experience distilled

Apache Mahout Essentials

Implement top-notch machine learning algorithms for classification, clustering, and recommendations with Apache Mahout

Jayani Withanawasam



BIRMINGHAM - MUMBAI

Apache Mahout Essentials

Copyright © 2015 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: June 2015

Production reference: 1120615

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78355-499-7

www.packtpub.com

Credits

Author

Jayani Withanawasam

Project Coordinator

Vijay Kushlani

Reviewers

Guillaume Agis

Saleem A. Ansari

Sahil Kharb

Pavan Kumar Narayanan

Proofreader

Safis Editing

Indexer

Tejal Soni

Commissioning Editor

Akram Hussain

Graphics

Sheetal Aute

Jason Monteiro

Acquisition Editor

Shaon Basu

Production Coordinator

Melwyn D'sa

Content Development Editor

Nikhil Potdukhe

Cover Work

Melwyn D'sa

Technical Editor

Tanmayee Patil

Copy Editor

Dipti Kapadia

About the Author

Jayani Withanawasam is R&D engineer and a senior software engineer at Zaizi Asia, where she focuses on applying machine learning techniques to provide smart content management solutions.

She is currently pursuing an MSc degree in artificial intelligence at the University of Moratuwa, Sri Lanka, and has completed her BE in software engineering (with first class honors) from the University of Westminster, UK.

She has more than 6 years of industry experience, and she has worked in areas such as machine learning, natural language processing, and semantic web technologies during her tenure.

She is passionate about working with semantic technologies and big data.

First of all, I would like to thank the Apache Mahout contributors for the invaluable effort that they have put in the project, crafting it as a popular scalable machine learning library in the industry.

Also, I would like to thank Rafa Haro for leading me toward the exciting world of machine learning and natural language processing.

I am sincerely grateful to Shaon Basu, an acquisition editor at Packt Publishing, and Nikhil Potdukhe, a content development editor at Packt Publishing, for their remarkable guidance and encouragement as I wrote this book amid my other commitments.

Furthermore, my heartfelt gratitude goes to Abinia Sachithanantham and Dedunu Dhananjaya for motivating me throughout the journey of writing the book.

Last but not least, I am eternally thankful to my parents for staying by my side throughout all my pursuits and being pillars of strength.

About the Reviewers

Guillaume Agis is a French 25 year old with a master's degree in computer science from Epitech, where he studied for 4 years in France and 1 year in Finland.

Open-minded and interested in a lot of domains, such as healthcare, innovation, high-tech, and science, he is always open to new adventures and experiments. Currently, he works as a software engineer in London at a company called Touch Surgery, where he is developing an application. The application is a surgery simulator that allows you to practice and rehearse operations even before setting foot in the operating room.

His previous jobs were, for the most part, in R&D, where he worked with very innovative technologies, such as Mahout, to implement collaborative filtering into artificial intelligence.

He always does his best to bring his team to the top and tries to make a difference.

He's also helping while42, a worldwide alumni network of French engineers, to grow as well as manage the London chapter.

I would like to thank all the people who have brought me to the top and helped me become what I am now.

Saleem A. Ansari is a full stack Java/Scala/Ruby developer with over 7 years of industry experience and a special interest in machine learning and information retrieval. Having implemented data ingestion and processing pipeline in Core Java and Ruby separately, he knows the challenges faced by huge datasets in such systems. He has worked for companies such as Red Hat, Impetus Technologies, Belzabar Software Design, and Exzeo Software Pvt Ltd. He is also a passionate member of the Free and Open Source Software (FOSS) Community. He started his journey with FOSS in the year 2004. In 2005, he formed JMILUG - Linux User's Group at Jamia Millia Islamia University, New Delhi. Since then, he has been contributing to FOSS by organizing community activities and also by contributing code to various projects (<http://github.com/tuxdna>). He also mentors students on FOSS and its benefits. He is currently enrolled at Georgia Institute of Technology, USA, on the MSCS program. He can be reached at tuxdna@fedoraproject.org.

Apart from reviewing this book, he maintains a blog at <http://tuxdna.in/>.

First of all, I would like to thank the vibrant, talented, and generous Apache Mahout community that created such a wonderful machine learning library. I would like to thank Packt Publishing and its staff for giving me this wonderful opportunity. I would like to thank the author for his hard work in simplifying and elaborating on the latest information in Apache Mahout.

Sahil Kharb has recently graduated from the Indian Institute of Technology, Jodhpur (India), and is working at Rockon Technologies. In the past, he has worked on Mahout and Hadoop for the last two years. His area of interest is data mining on a large scale. Nowadays, he works on Apache Spark and Apache Storm, doing real-time data analytics and batch processing with the help of Apache Mahout.

He has also reviewed *Learning Apache Mahout*, Packt Publishing.

I would like to thank my family, for their unconditional love and support, and God Almighty, for giving me strength and endurance. Also, I am thankful to my friend Chandni, who helped me in testing the code.

Pavan Kumar Narayanan is an applied mathematician with over 3 years of experience in mathematical programming, data science, and analytics. Currently based in New York, he has worked to build a marketing analytics product for a startup using Apache Mahout and has published and presented papers in algorithmic research at Transportation Research Board, Washington DC, and SUNY Research Conference, Albany, New York. He also runs a blog, DataScience Hacks (<https://datasciencehacks.wordpress.com/>). His interests are exploring new problem solving techniques and software, from industrial mathematics to machine learning writing book reviews.

Pavan can be contacted at pavan.narayanan@gmail.com.

I would like to thank my family, for their unconditional love and support, and God Almighty, for giving me strength and endurance.

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.



<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	vii
Chapter 1: Introducing Apache Mahout	1
Machine learning in a nutshell	1
Features	2
Supervised learning versus unsupervised learning	2
Machine learning applications	3
Information retrieval	3
Business	5
Market segmentation (clustering)	5
Stock market predictions (regression)	5
Health care	5
Using a mammogram for cancer tissue detection	6
Machine learning libraries	6
Open source or commercial	6
Scalability	7
Languages used	7
Algorithm support	7
Batch processing versus stream processing	7
The story so far	8
Apache Mahout	9
Setting up Apache Mahout	10
How Apache Mahout works?	11
The high-level design	11
The distribution	12
From Hadoop MapReduce to Spark	12
Problems with Hadoop MapReduce	12
In-memory data processing with Spark and H2O	13
Why is Mahout shifting from Hadoop MapReduce to Spark?	13

When is it appropriate to use Apache Mahout?	14
Summary	14
Chapter 2: Clustering	15
Unsupervised learning and clustering	15
Applications of clustering	16
Computer vision and image processing	16
Types of clustering	17
Hard clustering versus soft clustering	17
Flat clustering versus hierarchical clustering	18
Model-based clustering	18
K-Means clustering	18
Getting your hands dirty!	20
Running K-Means using Java programming	20
Data preparation	20
Understanding important parameters	21
Cluster visualization	24
Distance measure	25
Writing a custom distance measure	28
K-Means clustering with MapReduce	28
MapReduce in Apache Mahout	29
The map function	31
The reduce function	31
Additional clustering algorithms	31
Canopy clustering	31
Fuzzy K-Means	33
Streaming K-Means	35
The streaming step	36
The ball K-Means step	37
Spectral clustering	38
Dirichlet clustering	38
Text clustering	39
The vector space model and TF-IDF	39
N-grams and collocations	40
Preprocessing text with Lucene	40
Text clustering with the K-Means algorithm	41
Topic modeling	44
Optimizing clustering performance	44
Selecting the right features	44
Selecting the right algorithms	45
Selecting the right distance measure	45

Evaluating clusters	45
The initialization of centroids and the number of clusters	45
Tuning up parameters	45
The decision on infrastructure	46
Summary	46
Chapter 3: Regression and Classification	47
Supervised learning	47
Target variables and predictor variables	48
Predictive analytics' techniques	48
Regression-based prediction	48
Model-based prediction	49
Tree-based prediction	49
Classification versus regression	49
Linear regression with Apache Spark	49
How does linear regression work?	50
A real-world example	50
The impact of smoking on mortality and different diseases	50
Linear regression with one variable and multiple variables	51
The integration of Apache Spark	53
Setting up Apache Spark with Apache Mahout	53
An example script	54
Distributed row matrix	55
An explanation of the code	56
Mahout references	58
The bias-variance trade-off	58
How to avoid over-fitting and under-fitting	59
Logistic regression with SGD	60
Logistic functions	60
Minimizing the cost function	61
Multinomial logistic regression versus binary logistic regression	62
A real-world example	63
An example script	64
Testing and evaluation	65
The confusion matrix	65
The area under the curve	66
The Naïve Bayes algorithm	66
The Bayes theorem	66
Text classification	66
Naïve assumption and its pros and cons in text classification	68
Improvements that Apache Mahout has made to the Naïve Bayes classification	68

A text classification coding example using the 20 newsgroups' example	68
Understand the 20 newsgroups' dataset	68
Text classification using Naïve Bayes – a MapReduce implementation	
with Hadoop	70
Text classification using Naïve Bayes – the Spark implementation	73
The Markov chain	74
Hidden Markov Model	74
A real-world example – developing a POS tagger using HMM	
supervised learning	75
POS tagging	75
HMM for POS tagging	76
HMM implementation in Apache Mahout	77
HMM supervised learning	78
The important parameters	78
Returns	79
The Baum Welch algorithm	79
A code example	80
The important parameters	80
The Viterbi evaluator	80
The Apache Mahout references	81
Summary	81
Chapter 4: Recommendations	83
Collaborative versus content-based filtering	84
Content-based filtering	84
Collaborative filtering	85
Hybrid filtering	86
User-based recommenders	86
A real-world example – movie recommendations	87
Data models	90
The similarity measure	91
The neighborhood	92
Recommenders	93
Evaluation techniques	93
The IR-based method (precision/recall)	94
Addressing the issues with inaccurate recommendation results	95
Item-based recommenders	95
Item-based recommenders with Spark	97
Matrix factorization-based recommenders	97
Alternative least squares	99
Singular value decomposition	99
Algorithm usage tips and tricks	100
Summary	101

Chapter 5: Apache Mahout in Production	103
Introduction	103
Apache Mahout with Hadoop	104
YARN with MapReduce 2.0	105
The resource manager	106
The application manager	106
A node manager	106
The application master	106
Containers	107
Managing storage with HDFS	107
The life cycle of a Hadoop application	108
Setting up Hadoop	109
Setting up Mahout in local mode	110
Prerequisites	110
Setting up Mahout in Hadoop distributed mode	110
Prerequisites	111
The pseudo-distributed mode	112
The fully-distributed mode	114
Monitoring Hadoop	118
Commands/scripts	118
Data nodes	119
Node managers	120
Web UIs	120
Setting up Mahout with Hadoop's fully-distributed mode	121
Troubleshooting Hadoop	121
Optimization tips	122
Summary	123
Chapter 6: Visualization	125
The significance of visualization in machine learning	125
D3.js	126
A visualization example for K-Means clustering	126
Summary	134
Index	135

Preface

Apache Mahout is a scalable machine learning library that provides algorithms for classification, clustering, and recommendations.

This book helps you to use Apache Mahout to implement widely used machine learning algorithms in order to gain better insights about large and complex datasets in a scalable manner.

Starting from fundamental concepts in machine learning and Apache Mahout, real-world applications, a diverse range of popular algorithms and their implementations, code examples, evaluation strategies, and best practices are given for each machine learning technique. Further, this book contains a complete step-by-step guide to set up Apache Mahout in the production environment, using Apache Hadoop to unleash the scalable power of Apache Mahout in a distributed environment. Finally, you are guided toward the data visualization techniques for Apache Mahout, which make your data come alive!

What this book covers

Chapter 1, Introducing Apache Mahout, provides an introduction to machine learning and Apache Mahout.

Chapter 2, Clustering, provides an introduction to unsupervised learning and clustering techniques (K-Means clustering and other algorithms) in Apache Mahout along with performance optimization tips for clustering.

Chapter 3, Regression and Classification, provides an introduction to supervised learning and classification techniques (linear regression, logistic regression, Naïve Bayes, and HMMs) in Apache Mahout.

Chapter 4, Recommendations, provides a comparison between collaborative- and content-based filtering and recommenders in Apache Mahout (user-based, item-based, and matrix-factorization-based).

Chapter 5, Apache Mahout in Production, provides a guide to scaling Apache Mahout in the production environment with Apache Hadoop.

Chapter 6, Visualization, provides a guide to visualizing data using D3.js.

What you need for this book

The following software libraries are needed at various phases of this book:

- Java 1.7 or above
- Apache Mahout
- Apache Hadoop
- Apache Spark
- D3.js

Who this book is for

If you are a Java developer or a data scientist who has not worked with Apache Mahout previously and want to get up to speed on implementing machine learning on big data, then this is a concise and fast-paced guide for you.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows:

"Save the following content in a file named as `KmeansTest.data`."

A block of code is set as follows:

```
<dependency>
  <groupId>org.apache.mahout</groupId>
  <artifactId>mahout-core</artifactId>
  <version>${mahout.version}</version>
</dependency>
```