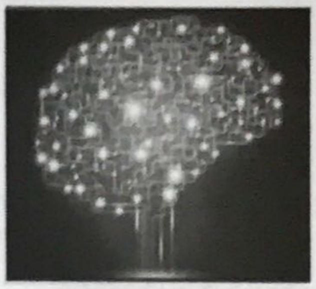
Machine Learning

HANDS-ON FOR DEVELOPERS AND TECHNICAL PROFESSIONALS

Jason Bell

WILEY



Machine Learning

Hands-On for Developers and Technical Professionals

Jason Bell

WILEY

Machine Learning: Hands-On for Developers and Technical Professionals

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

ISBN: 978-1-118-88906-0
ISBN: 978-1-118-88939-8 (ebk)
ISBN: 978-1-118-88949-7 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2014946682

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

To Wendy and Clarissa.

Credits

Executive Editor
Carol Long

Project Editor
Charlotte Kughen

Technical Editor
Mitchell Wyle

Production Editor
Christine Mugnolo

Copy Editor
Katherine Burt

Production Manager
Kathleen Wisor

Manager of Content Development and Assembly
Mary Beth Wakefield

Director of Community Marketing
David Mayhew

Marketing Manager
Carrie Sherrill

Business Manager
Amy Knies

Professional Technology & Strategy Director
Barry Pruett

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Patrick Redmond

Proofreader
Nancy Carrasco

Indexer
Johnna Dinse

Cover Designer
Wiley

Cover Image
© iStock.com/VLADGRIN

About the Author

Jason Bell has been working with point-of-sale and customer-loyalty data since 2002, and he has been involved in software development for more than 25 years. He is founder of Datasentiment, a UK business that helps companies worldwide with data acquisition, processing, and insight.

Acknowledgments

During the autumn of 2013, I was presented with some interesting options: either do a research-based PhD or co-author a book on machine learning. One would take six years and the other would take seven to eight months. Because of the speed the data industry was, and still is, progressing, the idea of the book was more appealing because I would be able to get something out while it was still fresh and relevant, and that was more important to me.

I say "co-author" because the original plan was to write a machine learning book with Aidan Rogers. Due to circumstances beyond his control he had to pull out. With Aidan's blessing, I continued under my own steam, and for that opportunity I can't thank him enough for his grace, encouragement, and support in that decision.

Many thanks goes to Wiley, especially Executive Editor, Carol Long, for letting me tweak things here and there with the original concept and bring it to a more practical level than a theoretical one; Project Editor, Charlotte Kughen, who kept me on the straight and narrow when there were times I didn't make sense; and Mitchell Wyle for reviewing the technical side of things. Also big thanks to the Wiley family as a whole for looking after me with this project.

Over the years I've met and worked with some incredible people, so in no particular order here goes: Garrett Murphy, Clare Conway, Colin Mitchell, David Crozier, Edd Dumbill, Matt Biddulph, Jim Weber, Tara Simpson, Marty Neill, John Girvin, Greg O'Hanlon, Clare Rowland, Tim Spear, Ronan Cunningham, Tom Grey, Stevie Morrow, Steve Orr, Kevin Parker, John Reid, James Blundell, Mary McKenna, Mark Nagurski, Alan Hook, Jon Brookes, Conal Loughrey, Paul Graham, Frankie Colclough, and countless others (whom I will be kicking myself that I've forgotten) for all the meetings, the chats, the ideas, and the collaborations.

Thanks to Tim Brundle, Matt Johnson, and Alan Thorburn for their support and for introducing me to the people who would inspire thoughts that would spur me on to bigger challenges with data. An enormous thank you to Thomas Spinks for having faith in me, without him there wouldn't have been a career in computing.

In relation to the challenge of writing a book I have to thank Ben Hammersley, Alistair Croll, Alasdair Allan, and John Foreman for their advice and support throughout the whole process.

I also must thank my dear friend, Colin McHale, who, on one late evening while waiting for the soccer data to refresh, taught me Perl on the back of a KitKat wrapper, thus kick-starting a journey of software development.

Finally, to my wife, Wendy, and my daughter, Clarissa, for absolutely everything and encouraging me to do this book to the best of my nerdy ability. I couldn't have done it without you both. And to the Bell family—George, Maggie and my sister Fern—who have encouraged my computing journey from a very early age.

During the course of writing this book, musical enlightenment was brought to me by St. Vincent, Trey Gunn, Suzanne Vega, Tackhead, Peter Gabriel, Doug Wimbish, King Crimson, and Level 42.

Contents

Introduction xix

Chapter 1 What Is Machine Learning? 1

History of Machine Learning 1

- Alan Turing 1
- Arthur Samuel 2
- Tom M. Mitchell 2
- Summary Definition 2

Algorithm Types for Machine Learning 3

- Supervised Learning 3
- Unsupervised Learning 3

The Human Touch 4

Uses for Machine Learning 4

- Software 4
- Stock Trading 5
- Robotics 6
- Medicine and Healthcare 6
- Advertising 6
- Retail and E-Commerce 7
- Gaming Analytics 8
- The Internet of Things 9

Languages for Machine Learning 10

- Python 10
- R 10
- Matlab 10
- Scala 10
- Clojure 11
- Ruby 11

Software Used in This Book 11
- Checking the Java Version 11
- Weka Toolkit 12
- Mahout 12
- SpringXD 13
- Hadoop 13
- Using an IDE 14

Data Repositories 14
- UC Irvine Machine Learning Repository 14
- Infochimps 14
- Kaggle 15

Summary 15

Chapter 2 Planning for Machine Learning 17

The Machine Learning Cycle 17

It All Starts with a Question 18

I Don't Have Data! 19
- Starting Local 19
- Competitions 19

One Solution Fits All? 20

Defining the Process 20
- Planning 20
- Developing 21
- Testing 21
- Reporting 21
- Refining 22
- Production 22

Building a Data Team 22
- Mathematics and Statistics 22
- Programming 23
- Graphic Design 23
- Domain Knowledge 23

Data Processing 23
- Using Your Computer 24
- A Cluster of Machines 24
- Cloud-Based Services 24

Data Storage 25
- Physical Discs 25
- Cloud-Based Storage 25

Data Privacy 25
- Cultural Norms 25
- Generational Expectations 26
- The Anonymity of User Data 26
- Don't Cross "The Creepy Line" 27

Data Quality and Cleaning 28
- Presence Checks 28

	Type Checks	29
	Length Checks	29
	Range Checks	30
	Format Checks	30
	The Britney Dilemma	30
	What's in a Country Name?	33
	Dates and Times	35
	Final Thoughts on Data Cleaning	35
	Thinking about Input Data	36
	Raw Text	36
	Comma Separated Variables	36
	JSON	37
	YAML	39
	XML	39
	Spreadsheets	40
	Databases	41
	Thinking about Output Data	42
	Don't Be Afraid to Experiment	42
	Summary	43
Chapter 3	**Working with Decision Trees**	**45**
	The Basics of Decision Trees	45
	Uses for Decision Trees	45
	Advantages of Decision Trees	46
	Limitations of Decision Trees	46
	Different Algorithm Types	47
	How Decision Trees Work	48
	Decision Trees in Weka	53
	The Requirement	53
	Training Data	53
	Using Weka to Create a Decision Tree	55
	Creating Java Code from the Classification	60
	Testing the Classifier Code	64
	Thinking about Future Iterations	66
	Summary	67
Chapter 4	**Bayesian Networks**	**69**
	Pilots to Paperclips	69
	A Little Graph Theory	70
	A Little Probability Theory	72
	Coin Flips	72
	Conditional Probability	72
	Winning the Lottery	73
	Bayes' Theorem	73
	How Bayesian Networks Work	75
	Assigning Probabilities	76
	Calculating Results	77

Node Counts 78
Using Domain Experts 78
A Bayesian Network Walkthrough 79
Java APIs for Bayesian Networks 79
Planning the Network 79
Coding Up the Network 81
Summary 90

Chapter 5 Artificial Neural Networks 91
What Is a Neural Network? 91
Artificial Neural Network Uses 92
High-Frequency Trading 92
Credit Applications 93
Data Center Management 93
Robotics 93
Medical Monitoring 93
Breaking Down the Artificial Neural Network 94
Perceptrons 94
Activation Functions 95
Multilayer Perceptrons 96
Back Propagation 98
Data Preparation for Artificial Neural Networks 99
Artificial Neural Networks with Weka 100
Generating a Dataset 100
Loading the Data into Weka 102
Configuring the Multilayer Perceptron 103
Training the Network 105
Altering the Network 108
Increasing the Test Data Size 108
Implementing a Neural Network in Java 109
Create the Project 109
The Code 111
Converting from CSV to Arff 114
Running the Neural Network 114
Summary 115

Chapter 6 Association Rules Learning 117
Where Is Association Rules Learning Used? 117
Web Usage Mining 118
Beer and Diapers 118
How Association Rules Learning Works 119
Support 121
Confidence 121
Lift 122
Conviction 122
Defining the Process 122

Algorithms 123
Apriori 123
FP-Growth 124
Mining the Baskets—A Walkthrough 124
Downloading the Raw Data 124
Setting Up the Project in Eclipse 125
Setting Up the Items Data File 126
Setting Up the Data 129
Running Mahout 131
Inspecting the Results 133
Putting It All Together 135
Further Development 136
Summary 137

Chapter 7 Support Vector Machines 139
What Is a Support Vector Machine? 139
Where Are Support Vector Machines Used? 140
The Basic Classification Principles 140
Binary and Multiclass Classification 140
Linear Classifiers 142
Confidence 143
Maximizing and Minimizing to Find the Line 143
How Support Vector Machines Approach Classification 144
Using Linear Classification 144
Using Non-Linear Classification 146
Using Support Vector Machines in Weka 147
Installing LibSVM 147
A Classification Walkthrough 148
Implementing LibSVM with Java 154
Summary 159

Chapter 8 Clustering 161
What Is Clustering? 161
Where Is Clustering Used? 162
The Internet 162
Business and Retail 163
Law Enforcement 163
Computing 163
Clustering Models 164
How the K-Means Works 164
Calculating the Number of Clusters in a Dataset 166
K-Means Clustering with Weka 168
Preparing the Data 168
The Workbench Method 169
The Command-Line Method 174
The Coded Method 178
Summary 186

Chapter 9 Machine Learning in Real Time with Spring XD 187
Capturing the Firehose of Data 187
Considerations of Using Data in Real Time 188
Potential Uses for a Real-Time System 188
Using Spring XD 189
Spring XD Streams 190
Input Sources, Sinks, and Processors 190
Learning from Twitter Data 193
The Development Plan 193
Configuring the Twitter API Developer Application 194
Configuring Spring XD 196
Starting the Spring XD Server 197
Creating Sample Data 198
The Spring XD Shell 198
Streams 101 199
Spring XD and Twitter 202
Setting the Twitter Credentials 202
Creating Your First Twitter Stream 203
Where to Go from Here 205
Introducing Processors 206
How Processors Work within a Stream 206
Creating Your Own Processor 207
Real-Time Sentiment Analysis 215
How the Basic Analysis Works 215
Creating a Sentiment Processor 217
Spring XD Taps 221
Summary 222

Chapter 10 Machine Learning as a Batch Process 223
Is It Big Data? 223
Considerations for Batch Processing Data 224
Volume and Frequency 224
How Much Data? 225
Which Process Method? 225
Practical Examples of Batch Processes 225
Hadoop 225
Sqoop 226
Pig 226
Mahout 226
Cloud-Based Elastic Map Reduce 226
A Note about the Walkthroughs 227
Using the Hadoop Framework 227
The Hadoop Architecture 227
Setting Up a Single-Node Cluster 229

How MapReduce Works 233
Mining the Hashtags 234
  Hadoop Support in Spring XD 235
  Objectives for This Walkthrough 235
  What's a Hashtag? 235
  Creating the MapReduce Classes 236
  Performing ETL on Existing Data 247
  Product Recommendation with Mahout 250
Mining Sales Data 256
  Welcome to My Coffee Shop! 257
  Going Small Scale 258
  Writing the Core Methods 258
  Using Hadoop and MapReduce 260
  Using Pig to Mine Sales Data 263
Scheduling Batch Jobs 273
Summary 274

Chapter 11 Apache Spark 275
Spark: A Hadoop Replacement? 275
Java, Scala, or Python? 276
Scala Crash Course 276
  Installing Scala 276
  Packages 277
  Data Types 277
  Classes 278
  Calling Functions 278
  Operators 279
  Control Structures 279
Downloading and Installing Spark 280
A Quick Intro to Spark 280
  Starting the Shell 281
  Data Sources 282
  Testing Spark 282
  Spark Monitor 284
Comparing Hadoop MapReduce to Spark 285
Writing Standalone Programs with Spark 288
  Spark Programs in Scala 288
  Installing SBT 288
  Spark Programs in Java 291
  Spark Program Summary 295
Spark SQL 295
  Basic Concepts 295
  Using SparkSQL with RDDs 296
Spark Streaming 305
  Basic Concepts 305
  Creating Your First Stream with Scala 306
  Creating Your First Stream with Java 309

MLib: The Machine Learning Library 311
Dependencies 311
Decision Trees 312
Clustering 313
Summary 313

Chapter 12 Machine Learning with R 315
Installing R 315
Mac OSX 315
Windows 316
Linux 316
Your First Run 316
Installing R-Studio 317
The R Basics 318
Variables and Vectors 318
Matrices 319
Lists 320
Data Frames 321
Installing Packages 322
Loading in Data 323
Plotting Data 324
Simple Statistics 327
Simple Linear Regression 329
Creating the Data 329
The Initial Graph 329
Regression with the Linear Model 330
Making a Prediction 331
Basic Sentiment Analysis 331
Functions to Load in Word Lists 331
Writing a Function to Score Sentiment 332
Testing the Function 333
Apriori Association Rules 333
Installing the ARules Package 334
The Training Data 334
Importing the Transaction Data 335
Running the Apriori Algorithm 336
Inspecting the Results 336
Accessing R from Java 337
Installing the rJava Package 337
Your First Java Code in R 337
Calling R from Java Programs 338
Setting Up an Eclipse Project 338
Creating the Java/R Class 339
Running the Example 340
Extending Your R Implementations 342
R and Hadoop 342

The RHadoop Project 342
A Sample Map Reduce Job in RHadoop 343
Connecting to Social Media with R 345
Summary 347

Appendix A SpringXD Quick Start 349
Installing Manually 349
Starting SpringXD 349
Creating a Stream 350
Adding a Twitter Application Key 350

Appendix B Hadoop 1.x Quick Start 351
Downloading and Installing Hadoop 351
Formatting the HDFS Filesystem 352
Starting and Stopping Hadoop 353
Process List of a Basic Job 353

Appendix C Useful Unix Commands 355
Using Sample Data 355
Showing the Contents: cat, more, and less 356
Example Command 356
Expected Output 356
Filtering Content: grep 357
Example Command for Finding Text 357
Example Output 357
Sorting Data: sort 358
Example Command for Basic Sorting 358
Example Output 358
Finding Unique Occurrences: uniq 360
Showing the Top of a File: head 361
Counting Words: wc 361
Locating Anything: find 362
Combining Commands and Redirecting Output 363
Picking a Text Editor 363
Colon Frenzy: Vi and Vim 363
Nano 364
Emacs 364

Appendix D Further Reading 367
Machine Learning 367
Statistics 368
Big Data and Data Science 368
Hadoop 368
Visualization 369
Making Decisions 369
Datasets 369
Blogs 370
Useful Websites 370
The Tools of the Trade 370

Index 373