Introduction

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
# Contents

Software Used in This Book
- 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
Contents

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
Contents

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

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

### Chapter 10  Machine Learning as a Batch Process

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is It Big Data?</td>
<td>223</td>
</tr>
<tr>
<td>Considerations for Batch Processing Data</td>
<td>224</td>
</tr>
<tr>
<td>Volume and Frequency</td>
<td>224</td>
</tr>
<tr>
<td>How Much Data?</td>
<td>225</td>
</tr>
<tr>
<td>Which Process Method?</td>
<td>225</td>
</tr>
<tr>
<td>Practical Examples of Batch Processes</td>
<td>225</td>
</tr>
<tr>
<td>Hadoop</td>
<td>225</td>
</tr>
<tr>
<td>Sqoop</td>
<td>226</td>
</tr>
<tr>
<td>Pig</td>
<td>226</td>
</tr>
<tr>
<td>Mahout</td>
<td>226</td>
</tr>
<tr>
<td>Cloud-Based Elastic Map Reduce</td>
<td>226</td>
</tr>
<tr>
<td>A Note about the Walkthroughs</td>
<td>227</td>
</tr>
<tr>
<td>Using the Hadoop Framework</td>
<td>227</td>
</tr>
<tr>
<td>The Hadoop Architecture</td>
<td>227</td>
</tr>
<tr>
<td>Setting Up a Single-Node Cluster</td>
<td>229</td>
</tr>
</tbody>
</table>
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
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