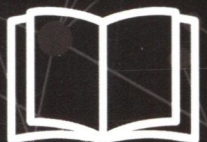
Atul Tripathi

Practical Machine Learning Cookbook

Resolving and offering solutions to your machine learning problems with R



Packt>

Practical Machine Learning Cookbook

Machine learning has become the new black. The challenge in today's world is the explosion of data from existing legacy data and incoming new structured and unstructured data. The complexity of discovering, understanding, performing analysis, and predicting outcomes on the data using machine learning algorithms is a challenge. This cookbook will help solve the everyday challenges you will face as a data scientist. The application of various data science techniques and on multiple datasets based on real-world challenges you face will help you appreciate a variety of techniques used in various situations.

The first half of the book provides recipes on fairly complex machine-learning systems, where you'll learn to explore new areas of applications of machine learning and improve its efficiency. That includes recipes on classifications, neural networks, unsupervised and supervised learning, deep learning, reinforcement learning, and more.

The second half of the book focuses on three different machine learning case studies, all based on real-world data, and offers solutions and solves specific machine-learning issues in each one.

Things you will learn:

- Get equipped with a deeper understanding of how to apply machine learning techniques
- Implement each of the advanced machine learning techniques
- Solve real-life problems that are encountered in order to make your applications produce improved results
- Gain hands-on experience in problem solving for your machine learning systems
- Understand the methods of collecting data, preparing data for usage, training the model, evaluating the model's performance, and improving the model's performance

Packt>
www.packtpub.com

$ **59.99** US
£ **49.99** UK

Prices do not include local sales
Tax or VAT where applicable

ISBN 978-1-78528-051-1
55999
9 781785 280511

Practical Machine Learning Cookbook

Atul Tripathi

Practical Machine Learning Cookbook

Resolving and offering solutions to your machine learning problems with R

Atul Tripathi

BIRMINGHAM - MUMBAI

Practical Machine Learning Cookbook

First published: April 2017

Production reference: 1070417

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.
ISBN 978-1-78528-051-1

www.packtpub.com

Credits

Author
Atul Tripathi

Reviewer
Ryota Kamoshida

Commissioning Editor
Akram Hussain

Acquisition Editor
Tushar Gupta

Content Development Editor
Aishwarya Pandere

Technical Editor
Prasad Ramesh

Copy Editor
Safis Editing

Project Coordinator
Nidhi Joshi

Proofreader
Safis Editing

Indexer
Tejal Daruwale Soni

Graphics
Tania Dutta

Production Coordinator
Shantanu Zagade

About the Author

Atul Tripathi has spent more than 11 years in the fields of machine learning and quantitative finance. He has a total of 14 years of experience in software development and research. He has worked on advanced machine learning techniques, such as neural networks and Markov models. While working on these techniques, he has solved problems related to image processing, telecommunications, human speech recognition, and natural language processing. He has also developed tools for text mining using neural networks. In the field of quantitative finance, he has developed models for Value at Risk, Extreme Value Theorem, Option Pricing, and Energy Derivatives using Monte Carlo simulation techniques.

About the Reviewer

Ryota Kamoshida is the developer of the Python library **MALSS** (**MAchine Learning Support System**), (`https://github.com/canard0328/malss`) and now works as a senior researcher in the field of computer science at Hitachi, Ltd.

www.PacktPub.com

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://www.packtpub.com/mapt`

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

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

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/dp/1785280511`.

If you'd like to join our team of regular reviewers, you can e-mail us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface 1

Chapter 1: Introduction to Machine Learning 7

What is machine learning? 7
An overview of classification 9
An overview of clustering 10
An overview of supervised learning 10
An overview of unsupervised learning 12
An overview of reinforcement learning 12
An overview of structured prediction 14
An overview of neural networks 15
An overview of deep learning 16

Chapter 2: Classification 17

Introduction 17
Discriminant function analysis - geological measurements on brines from wells 19
Getting ready 19
Step 1 - collecting and describing data 19
How to do it... 20
Step 2 - exploring data 20
Step 3 - transforming data 21
Step 4 - training the model 22
Step 5 - classifying the data 24
Step 6 - evaluating the model 27
Multinomial logistic regression - understanding program choices made by students 29
Getting ready 29
Step 1 - collecting data 30
How to do it... 30
Step 2 - exploring data 31
Step 3 - training the model 32
Step 4 - testing the results of the model 33
Step 5 - model improvement performance 33
Tobit regression - measuring the students' academic aptitude 36
Getting ready 36
Step 1 - collecting data 37
How to do it... 37

Step 2 - exploring data 37
Step 3 - plotting data 38
Step 4 - exploring relationships 40
Step 5 - training the model 41
Step 6 - testing the model 43
Poisson regression - understanding species present in Galapagos Islands 46
Getting ready 46
Step 1 - collecting and describing the data 46
How to do it... 46
Step 2 - exploring the data 47
Step 3 - plotting data and testing empirical data 49
Step 4 - rectifying discretization of the Poisson model 50
Step 5 - training and evaluating the model using the link function 52
Step 6 - revaluating using the Poisson model 53
Step 7 - revaluating using the linear model 54
Chapter 3: Clustering 67
Introduction 67
Hierarchical clustering - World Bank sample dataset 68
Getting ready 69
Step 1 - collecting and describing data 69
How to do it... 69
Step 2 - exploring data 70
Step 3 - transforming data 70
Step 4 - training and evaluating the model performance 72
Step 5 - plotting the model 73
Hierarchical clustering - Amazon rainforest burned between 1999-2010 76
Getting ready 77
Step 1 - collecting and describing data 77
How to do it... 78
Step 2 - exploring data 78
Step 3 - transforming data 79
Step 4 - training and evaluating model performance 81
Step 5 - plotting the model 82
Step 6 - improving model performance 84
Hierarchical clustering - gene clustering 91
Getting ready 92
Step 1 - collecting and describing data 92
How to do it... 94
Step 2 - exploring data 94
Step 3 - transforming data 95
Step 4 - training the model 97
Step 5 - plotting the model 103
Binary clustering - math test 111

Getting ready 112
Step 1 - collecting and describing data 112
How to do it... 112
Step 2 - exploring data 112
Step 3 - training and evaluating model performance 112
Step 4 - plotting the model 115
Step 5 - K-medoids clustering 117
K-means clustering - European countries protein consumption 120
Getting ready 120
Step 1 - collecting and describing data 120
How to do it... 121
Step 2 - exploring data 121
Step 3 - clustering 122
Step 4 - improving the model 124
K-means clustering - foodstuff 129
Getting ready 129
Step 1 - collecting and describing data 130
How to do it... 130
Step 2 - exploring data 130
Step 3 - transforming data 131
Step 4 - clustering 133
Step 5 - visualizing the clusters 135
Chapter 4: Model Selection and Regularization 139
Introduction 139
Shrinkage methods - calories burned per day 141
Getting ready 141
Step 1 - collecting and describing data 141
How to do it... 142
Step 2 - exploring data 142
Step 3 - building the model 145
Step 4 - improving the model 148
Step 5 - comparing the model 156
Dimension reduction methods - Delta's Aircraft Fleet 159
Getting ready 160
Step 1 - collecting and describing data 160
How to do it... 160
Step 2 - exploring data 160
Step 3 - applying principal components analysis 162
Step 4 - scaling the data 166
Step 5 - visualizing in 3D plot 170
Principal component analysis - understanding world cuisine 174
Getting ready 174
Step 1 - collecting and describing data 174

How to do it... 174
Step 2 - exploring data 174
Step 3 - preparing data 175
Step 4 - applying principal components analysis 178
Chapter 5: Nonlinearity 181
Generalized additive models - measuring the household income of New Zealand 181
Getting ready 181
Step 1 - collecting and describing data 182
How to do it... 182
Step 2 - exploring data 182
Step 3 - setting up the data for the model 184
Step 4 - building the model 185
Smoothing splines - understanding cars and speed 188
How to do it... 188
Step 1 - exploring the data 188
Step 2 - creating the model 189
Step 3 - fitting the smooth curve model 193
Step 4 - plotting the results 197
Local regression - understanding drought warnings and impact 202
Getting ready 203
Step 1 - collecting and describing data 203
How to do it... 203
Step 2 - collecting and exploring data 203
Step 3 - calculating the moving average 205
Step 4 - calculating percentiles 206
Step 5 - plotting results 209
Chapter 6: Supervised Learning 213
Introduction 213
Decision tree learning - Advance Health Directive for patients with chest pain 215
Getting ready 215
Step 1 - collecting and describing the data 215
How to do it... 216
Step 2 - exploring the data 216
Step 3 - preparing the data 218
Step 4 - training the model 221
Step 5- improving the model 224
Decision tree learning - income-based distribution of real estate values 226
Getting ready 227
Step 1 - collecting and describing the data 227
How to do it... 227

Step 2 - exploring the data 227
Step 3 - training the model 229
Step 4 - comparing the predictions 232
Step 5 - improving the model 237
Decision tree learning - predicting the direction of stock movement 242
Getting ready 243
Step 1 - collecting and describing the data 243
How to do it... 243
Step 2 - exploring the data 243
Step 3 - calculating the indicators 244
Step 4 - preparing variables to build datasets 252
Step 5 - building the model 261
Step 6 - improving the model 264
Naive Bayes - predicting the direction of stock movement 266
Getting ready 266
Step 1 - collecting and describing the data 266
How to do it... 266
Step 2 - exploring the data 266
Step 3 - preparing variables to build datasets 268
Step 4 - building the model 273
Step 5 - creating data for a new, improved model 274
Step 6 - improving the model 280
Random forest - currency trading strategy 286
Getting ready 286
Step 1 - collecting and describing the data 286
How to do it... 286
Step 2 - exploring the data 287
Step 3 - preparing variables to build datasets 289
Step 4 - building the model 296
Support vector machine - currency trading strategy 301
Getting ready 301
Step 1 - collecting and describing the data 301
How to do it... 301
Step 2 - exploring the data 302
Step 3 - calculating the indicators 303
Step 4 - preparing variables to build datasets 305
Step 5 - building the model 309
Stochastic gradient descent - adult income 311
Getting ready 312
Step 1 - collecting and describing the data 312
How to do it... 313
Step 2 - exploring the data 313
Step 3 - preparing the data 314
Step 4 - building the model 316
Step 5 - plotting the model 319

Chapter 7: Unsupervised Learning 321

Introduction 321

Self-organizing map - visualizing of heatmaps 322

How to do it... 322

Step 1 - exploring data 323

Step 2 - training the model 325

Step 3 - plotting the model 325

Vector quantization - image clustering 328

Getting ready 328

Step 1 - collecting and describing data 329

How to do it... 329

Step 2 - exploring data 329

Step 3 - data cleaning 329

Step 4 - visualizing cleaned data 330

Step 5 - building the model and visualizing it 331

Chapter 8: Reinforcement Learning 335

Introduction 335

Markov chains - the stocks regime switching model 336

Getting ready 336

Step 1 - collecting and describing the data 337

How to do it... 337

Step 2 - exploring the data 337

Step 3 - preparing the regression model 339

Step 4 - preparing the Markov-switching model 342

Step 5 - plotting the regime probabilities 346

Step 6 - testing the Markov switching model 349

Markov chains - the multi-channel attribution model 355

Getting ready 355

How to do it... 356

Step 1 - preparing the dataset 356

Step 2 - preparing the model 357

Step 3 - plotting the Markov graph 358

Step 4 - simulating the dataset of customer journeys 364

Step 5 - preparing a transition matrix heat map for real data 369

Markov chains - the car rental agency service 370

How to do it... 371

Step 1 - preparing the dataset 371

Step 2 - preparing the model 372

Step 3 - improving the model 374

Continuous Markov chains - vehicle service at a gas station 378

Getting ready 378

How to do it... 378

Step 1 - preparing the dataset 378