

Pro Machine Learning Algorithms

A Hands-On Approach to
Implementing Algorithms in
Python and R

V Kishore Ayyadevara

Apress®

Pro Machine Learning Algorithms

V Kishore Ayyadevara
Hyderabad, Andhra Pradesh, India

ISBN-13 (pbk): 978-1-4842-3563-8
<https://doi.org/10.1007/978-1-4842-3564-5>

ISBN-13 (electronic): 978-1-4842-3564-5

Library of Congress Control Number: 2018947188

Copyright © 2018 by V Kishore Ayyadevara

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Celestine John Suresh
Development Editor: Matthew Moodie
Coordinating Editor: Divya Modi

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit <http://www.apress.com/rights-permissions>.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at <http://www.apress.com/bulk-sales>.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-3563-8. For more detailed information, please visit <http://www.apress.com/source-code>.

Printed on acid-free paper

I would like to dedicate this book to my dear parents, Hema and Subrahmanyeswara Rao, to my lovely wife, Sindhura, and my dearest daughter, Hemanvi. This work would not have been possible without their support and encouragement.

Table of Contents

About the Author	xv
About the Technical Reviewer	xvii
Acknowledgments	xix
Introduction	xxi
Chapter 1: Basics of Machine Learning	1
Regression and Classification	1
Training and Testing Data	2
The Need for Validation Dataset	3
Measures of Accuracy	5
AUC Value and ROC Curve	7
Unsupervised Learning	11
Typical Approach Towards Building a Model	12
Where Is the Data Fetched From?	12
Which Data Needs to Be Fetched?	12
Pre-processing the Data	13
Feature Interaction	14
Feature Generation	14
Building the Models	14
Productionalizing the Models	14
Build, Deploy, Test, and Iterate	15
Summary	15

TABLE OF CONTENTS

- Chapter 2: Linear Regression 17**
- Introducing Linear Regression 17
- Variables: Dependent and Independent..... 18
- Correlation 18
- Causation..... 18
- Simple vs. Multivariate Linear Regression..... 18
- Formalizing Simple Linear Regression 19
- The Bias Term..... 19
- The Slope..... 20
- Solving a Simple Linear Regression 20
- More General Way of Solving a Simple Linear Regression 23
- Minimizing the Overall Sum of Squared Error 23
- Solving the Formula 24
- Working Details of Simple Linear Regression 25
- Complicating Simple Linear Regression a Little..... 26
- Arriving at Optimal Coefficient Values 29
- Introducing Root Mean Squared Error 29
- Running a Simple Linear Regression in R..... 30
- Residuals 31
- Coefficients..... 32
- SSE of Residuals (Residual Deviance) 34
- Null Deviance..... 34
- R Squared 34
- F-statistic 35
- Running a Simple Linear Regression in Python 36
- Common Pitfalls of Simple Linear Regression 37
- Multivariate Linear Regression 38
- Working details of Multivariate Linear Regression 40
- Multivariate Linear Regression in R..... 41
- Multivariate Linear Regression in Python..... 42

Issue of Having a Non-significant Variable in the Model	42
Issue of Multicollinearity	43
Mathematical Intuition of Multicollinearity	43
Further Points to Consider in Multivariate Linear Regression	44
Assumptions of Linear Regression	45
Summary.....	47
Chapter 3: Logistic Regression.....	49
Why Does Linear Regression Fail for Discrete Outcomes?	49
A More General Solution: Sigmoid Curve	51
Formalizing the Sigmoid Curve (Sigmoid Activation).....	52
From Sigmoid Curve to Logistic Regression.....	53
Interpreting the Logistic Regression	53
Working Details of Logistic Regression	54
Estimating Error.....	56
Least Squares Method and Assumption of Linearity	57
Running a Logistic Regression in R	59
Running a Logistic Regression in Python.....	61
Identifying the Measure of Interest.....	61
Common Pitfalls.....	68
Time Between Prediction and the Event Happening	69
Outliers in Independent variables.....	69
Summary.....	69
Chapter 4: Decision Tree.....	71
Components of a Decision Tree.....	73
Classification Decision Tree When There Are Multiple Discrete Independent Variables.....	74
Information Gain	75
Calculating Uncertainty: Entropy	75
Calculating Information Gain	76
Uncertainty in the Original Dataset.....	76
Measuring the Improvement in Uncertainty	77

TABLE OF CONTENTS

- Which Distinct Values Go to the Left and Right Nodes 79
- When Does the Splitting Process Stop? 84
- Classification Decision Tree for Continuous Independent Variables..... 85
- Classification Decision Tree When There Are Multiple Independent Variables 88
- Classification Decision Tree When There Are Continuous and Discrete Independent Variables..... 93
- What If the Response Variable Is Continuous? 94
 - Continuous Dependent Variable and Multiple Continuous Independent Variables 95
 - Continuous Dependent Variable and Discrete Independent Variable 97
 - Continuous Dependent Variable and Discrete, Continuous Independent Variables 98
- Implementing a Decision Tree in R..... 99
- Implementing a Decision Tree in Python 99
- Common Techniques in Tree Building 100
- Visualizing a Tree Build 101
- Impact of Outliers on Decision Trees..... 102
- Summary..... 103
- Chapter 5: Random Forest 105**
 - A Random Forest Scenario..... 105
 - Bagging 107
 - Working Details of a Random Forest 107
 - Implementing a Random Forest in R..... 108
 - Parameters to Tune in a Random Forest 112
 - Variation of AUC by Depth of Tree 114
 - Implementing a Random Forest in Python 116
 - Summary..... 116
- Chapter 6: Gradient Boosting Machine 117**
 - Gradient Boosting Machine 117
 - Working details of GBM..... 118
 - Shrinkage..... 123

AdaBoost.....	126
Theory of AdaBoost	126
Working Details of AdaBoost	127
Additional Functionality for GBM.....	132
Implementing GBM in Python.....	132
Implementing GBM in R	133
Summary.....	134
Chapter 7: Artificial Neural Network	135
Structure of a Neural Network	136
Working Details of Training a Neural Network	138
Forward Propagation	138
Applying the Activation Function	141
Back Propagation	146
Working Out Back Propagation.....	146
Stochastic Gradient Descent	148
Diving Deep into Gradient Descent.....	148
Why Have a Learning Rate?.....	152
Batch Training	152
The Concept of Softmax	153
Different Loss Optimization Functions	155
Scaling a Dataset.....	156
Implementing Neural Network in Python	157
Avoiding Over-fitting using Regularization.....	160
Assigning Weightage to Regularization term	162
Implementing Neural Network in R.....	163
Summary.....	165

TABLE OF CONTENTS

- Chapter 8: Word2vec 167**
 - Hand-Building a Word Vector 168
 - Methods of Building a Word Vector 173
 - Issues to Watch For in a Word2vec Model 174
 - Frequent Words 174
 - Negative Sampling 175
 - Implementing Word2vec in Python 175
 - Summary..... 178

- Chapter 9: Convolutional Neural Network 179**
 - The Problem with Traditional NN..... 180
 - Scenario 1 183
 - Scenario 2 184
 - Scenario 3 185
 - Scenario 4 186
 - Understanding the Convolutional in CNN 187
 - From Convolution to Activation..... 189
 - From Convolution Activation to Pooling..... 189
 - How Do Convolution and Pooling Help?..... 190
 - Creating CNNs with Code..... 190
 - Working Details of CNN..... 194
 - Deep Diving into Convolutions/Kernels..... 203
 - From Convolution and Pooling to Flattening: Fully Connected Layer 205
 - From One Fully Connected Layer to Another 206
 - From Fully Connected Layer to Output Layer 206
 - Connecting the Dots: Feed Forward Network 206
 - Other Details of CNN 207
 - Backward Propagation in CNN 209
 - Putting It All Together..... 210

Data Augmentation	212
Implementing CNN in R.....	214
Summary.....	215
Chapter 10: Recurrent Neural Network	217
Understanding the Architecture	218
Interpreting an RNN	219
Working Details of RNN.....	220
Time Step 1	224
Time Step 2	224
Time Step 3	225
Implementing RNN: SimpleRNN	227
Compiling a Model.....	228
Verifying the Output of RNN.....	230
Implementing RNN: Text Generation	234
Embedding Layer in RNN	238
Issues with Traditional RNN	243
The Problem of Vanishing Gradient	244
The Problem of Exploding Gradients	245
LSTM	245
Implementing Basic LSTM in keras.....	247
Implementing LSTM for Sentiment Classification.....	255
Implementing RNN in R.....	256
Summary.....	257
Chapter 11: Clustering.....	259
Intuition of clustering.....	259
Building Store Clusters for Performance Comparison.....	260
Ideal Clustering.....	261
Striking a Balance Between No Clustering and Too Much Clustering: K-means Clustering.....	262

TABLE OF CONTENTS

- The Process of Clustering 264
- Working Details of K-means Clustering Algorithm 268
 - Applying the K-means Algorithm on a Dataset 269
 - Properties of the K-means Clustering Algorithm 271
- Implementing K-means Clustering in R 274
- Implementing K-means Clustering in Python 275
- Significance of the Major Metrics 276
- Identifying the Optimal K 276
- Top-Down Vs. Bottom-Up Clustering 278
 - Hierarchical Clustering 278
 - Major Drawback of Hierarchical Clustering 280
- Industry Use-Case of K-means Clustering 280
- Summary 281
- Chapter 12: Principal Component Analysis 283**
 - Intuition of PCA 283
 - Working Details of PCA 286
 - Scaling Data in PCA 291
 - Extending PCA to Multiple Variables 291
 - Implementing PCA in R 294
 - Implementing PCA in Python 295
 - Applying PCA to MNIST 296
 - Summary 297
- Chapter 13: Recommender Systems 299**
 - Understanding k-nearest Neighbors 300
 - Working Details of User-Based Collaborative Filtering 302
 - Euclidian Distance 303
 - Cosine Similarity 306
 - Issues with UBCF 311
 - Item-Based Collaborative Filtering 312
 - Implementing Collaborative Filtering in R 313

Implementing Collaborative Filtering in Python	314
Working Details of Matrix Factorization	315
Implementing Matrix Factorization in Python	321
Implementing Matrix Factorization in R.....	324
Summary.....	325
Chapter 14: Implementing Algorithms in the Cloud	327
Google Cloud Platform	327
Microsoft Azure Cloud Platform	331
Amazon Web Services.....	333
Transferring Files to the Cloud Instance	340
Running Instance Jupyter Notebooks from Your Local Machine.....	342
Installing R on the Instance.....	343
Summary.....	344
Appendix: Basics of Excel, R, and Python.....	345
Basics of Excel	345
Basics of R.....	347
Downloading R	348
Installing and Configuring RStudio	348
Getting Started with RStudio	349
Basics of Python	356
Downloading and installing Python	356
Basic operations in Python	358
Numpy	360
Number generation using Numpy.....	361
Slicing and indexing	362
Pandas.....	363
Indexing and slicing using Pandas	363
Summarizing data	364
Index.....	365

About the Author



V Kishore Ayyadevara is passionate about all things data. He has been working at the intersection of technology, data, and machine learning to identify, communicate, and solve business problems for more than a decade.

He's worked for American Express in risk management, in Amazon's supply chain analytics teams, and is currently leading data product development for a startup. In this role, he is responsible for implementing a variety of analytical solutions and building strong data science teams. He received his MBA from IIM Calcutta.

Kishore is an active learner, and his interests include identifying business problems that can be solved using data, simplifying the complexity within data science, and in transferring techniques across domains to achieve quantifiable business results.

He can be reached at www.linkedin.com/in/kishore-ayyadevara/

About the Technical Reviewer



Manohar Swamynathan is a data science practitioner and an avid programmer, with more than 13 years of experience in various data science-related areas, including data warehousing, business intelligence (BI), analytical tool development, ad-hoc analysis, predictive modeling, data science product development, consulting, formulating strategy, and executing analytics programs. He's made a career covering the lifecycle of data across different domains, including the US mortgage banking, retail/e-commerce, insurance, and industrial IoT. He has a bachelor's degree with a specialization in physics, mathematics, and computers, and a master's degree in project management. He currently lives in Bengaluru, the Silicon Valley of India.

He is the author of the book *Mastering Machine Learning with Python in Six Steps* (Apress, 2017). You can learn more about his various other activities on his website: www.mswamynathan.com.

Acknowledgments

I am grateful to my wife, Sindhura, for her love and constant support and for being a source of inspiration all through.

Sincere thanks to the Apress team, Celestin, Divya, and Matt, for their support and belief in me. Special thanks to Manohar for his review and helpful feedback. This book would not have been in this shape, without the great support from Arockia Rajan and Corbin Collins.

Thanks to Santanu Pattanayak and Antonio Gulli, who reviewed a few chapters, and also a few individuals in my organization who helped me considerably in proofreading and initial reviews: Praveen Balireddy, Arunjith, Navatha Komatireddy, Aravind Atreya, and Anugna Reddy.

Introduction

Machine learning techniques are being adopted for a variety of applications. With an increase in the adoption of machine learning techniques, it is very important for the developers of machine learning applications to understand *what* the underlying algorithms are learning, and more importantly, to understand *how* the various algorithms are learning the patterns from raw data so that they can be leveraged even more effectively.

This book is intended for data scientists and analysts who are interested in looking under the hood of various machine learning algorithms. This book will give you the confidence and skills when developing the major machine learning models and when evaluating a model that is presented to you.

True to the spirit of understanding what the machine learning algorithms are learning and how they are learning them, we first build the algorithms in Excel so that we can peek inside the black box of how the algorithms are working. In this way, the reader learns how the various levers in an algorithm impact the final result.

Once we've seen how the algorithms work, we implement them in both Python and R. However, this is *not* a book on Python or R, and I expect the reader to have some familiarity with programming. That said, the basics of Excel, Python, and R are explained in the appendix.

Chapter 1 introduces the basic terminology of data science and discusses the typical workflow of a data science project.

Chapters 2–10 cover some of the major supervised machine learning and deep learning algorithms used in industry.

Chapters 11 and 12 discuss the major unsupervised learning algorithms.

In Chapter 13, we implement the various techniques used in recommender systems to predict the likelihood of a user liking an item.

Finally, Chapter 14 looks at using the three major cloud service providers: Google Cloud Platform, Microsoft Azure, and Amazon Web Services.

All the datasets used in the book and the code snippets are available on GitHub at <https://github.com/kishore-ayyadevara/Pro-Machine-Learning>.