Advanced Data Analytics Using Python

With Machine Learning, Deep Learning and NLP Examples

Sayan Mukhopadhyay

Apress®

Advanced Data Analytics Using Python

Sayan Mukhopadhyay
Kolkata, West Bengal, India

ISBN-13 (pbk): 978-1-4842-3449-5 ISBN-13 (electronic): 978-1-4842-3450-1
https://doi.org/10.1007/978-1-4842-3450-1

Library of Congress Control Number: 2018937906

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Celestin
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 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 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/978-1-4842-3449-5. For more detailed information, please visit www.apress.com/source-code.

Printed on acid-free paper

This is dedicated to all my math teachers, especially to Kalyan Chakraborty.

Table of Contents

About the Author xi

About the Technical Reviewer xiii

Acknowledgments xv

Chapter 1: Introduction 1

Why Python? 1

When to Avoid Using Python 2

OOP in Python 3

Calling Other Languages in Python 12

Exposing the Python Model as a Microservice 14

High-Performance API and Concurrent Programming 17

Chapter 2: ETL with Python (Structured Data) 23

MySQL 23

How to Install MySQLdb? 23

Database Connection 24

INSERT Operation 24

READ Operation 25

DELETE Operation 26

UPDATE Operation 27

COMMIT Operation 28

ROLL-BACK Operation 28

Elasticsearch....31
Connection Layer API....33
Neo4j Python Driver....34
neo4j-rest-client....35
In-Memory Database....35
MongoDB (Python Edition)....36
Import Data into the Collection....36
Create a Connection Using pymongo....37
Access Database Objects....37
Insert Data....38
Update Data....38
Remove Data....38
Pandas....38
ETL with Python (Unstructured Data)....40
E-mail Parsing....40
Topical Crawling....42

Chapter 3: Supervised Learning Using Python....49

Dimensionality Reduction with Python....49
Correlation Analysis....50
Principal Component Analysis....53
Mutual Information....56
Classifications with Python....57
Semisupervised Learning....58
Decision Tree....59
Which Attribute Comes First?....59
Random Forest Classifier....60

Naive Bayes Classifier....61
Support Vector Machine....62
Nearest Neighbor Classifier....64
Sentiment Analysis....65
Image Recognition....67
Regression with Python....67
Least Square Estimation....68
Logistic Regression....69
Classification and Regression....70
Intentionally Bias the Model to Over-Fit or Under-Fit....71
Dealing with Categorical Data....73

Chapter 4: Unsupervised Learning: Clustering....77

K-Means Clustering....78
Choosing K: The Elbow Method....82
Distance or Similarity Measure....82
Properties....82
General and Euclidean Distance....83
Squared Euclidean Distance....84
Distance Between String-Edit Distance....85
Similarity in the Context of Document....87
Types of Similarity....87
What Is Hierarchical Clustering?....88
Bottom-Up Approach....89
Distance Between Clusters....90
Top-Down Approach....92
Graph Theoretical Approach....97
How Do You Know If the Clustering Result Is Good?....97

Chapter 5: Deep Learning and Neural Networks.....99

Backpropagation.....100

Backpropagation Approach.....100

Generalized Delta Rule.....100

Update of Output Layer Weights.....101

Update of Hidden Layer Weights.....102

BPN Summary.....103

Backpropagation Algorithm.....104

Other Algorithms.....106

TensorFlow.....106

Recurrent Neural Network.....113

Chapter 6: Time Series.....121

Classification of Variation.....121

Analyzing a Series Containing a Trend.....121

Curve Fitting.....122

Removing Trends from a Time Series.....123

Analyzing a Series Containing Seasonality.....124

Removing Seasonality from a Time Series.....125

By Filtering.....125

By Differencing.....126

Transformation.....126

To Stabilize the Variance.....126

To Make the Seasonal Effect Additive.....127

To Make the Data Distribution Normal.....127

Stationary Time Series.....128

Stationary Process.....128

Autocorrelation and the Correlogram.....129

Estimating Autocovariance and Autocorrelation Functions.....129

Time-Series Analysis with Python....130
Useful Methods....131
Autoregressive Processes....133
Estimating Parameters of an AR Process....134
Mixed ARMA Models....137
Integrated ARMA Models....138
The Fourier Transform....140
An Exceptional Scenario....141
Missing Data....143

Chapter 7: Analytics at Scale....145

Hadoop....145
MapReduce Programming....145
Partitioning Function....146
Combiner Function....147
HDFS File System....159
MapReduce Design Pattern....159
Spark....166
Analytics in the Cloud....168
Internet of Things....179

Index....181

About the Author

Sayan Mukhopadhyay has more than 13 years of industry experience and has been associated with companies such as Credit Suisse, PayPal, CA Technologies, CSC, and Mphasis. He has a deep understanding of applications for data analysis in domains such as investment banking, online payments, online advertisement, IT infrastructure, and retail. His area of expertise is in applying high-performance computing in distributed and data-driven environments such as real-time analysis, high-frequency trading, and so on.

He earned his engineering degree in electronics and instrumentation from Jadavpur University and his master's degree in research in computational and data science from IISc in Bangalore.

About the Technical Reviewer

Sundar Rajan Raman has more than 14 years of full stack IT experience in machine learning, deep learning, and natural language processing. He has six years of big data development and architect experience, including working with Hadoop and its ecosystems as well as other NoSQL technologies such as MongoDB and Cassandra. In fact, he has been the technical reviewer of several books on these topics.

He is also interested in strategizing using Design Thinking principles and in coaching and mentoring people.

Acknowledgments

Thanks to Labonic Chakraborty (Ripa) and Kusumika Mukherjee.