

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/342599493>

Practical Machine Learning for Data Analysis Using Python

Book · June 2020

CITATIONS

157

READS

3,648

1 author:



[Abdulhamit Subasi](#)

University of Turku

310 PUBLICATIONS 14,105 CITATIONS

SEE PROFILE

PRACTICAL
MACHINE
LEARNING FOR
DATA ANALYSIS
USING PYTHON

Page left intentionally blank

PRACTICAL MACHINE LEARNING FOR DATA ANALYSIS USING PYTHON

ABDULHAMIT SUBASI

*Professor of Information Systems at Effat University,
Jeddah, Saudi Arabia*



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2020 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-821379-7

For information on all Academic Press publications
visit our website at <https://www.elsevier.com/books-and-journals>

Publisher: Mara Conner
Editorial Project Manager: Rafael G. Trombaco
Production Project Manager: Paul Prasad Chandramohan
Designer: Christian Bilbow

Typeset by Thomson Digital



*A huge thank to my parents for always expecting me
to do my best, and telling me I could accomplish anything,
no matter what it was.*

To my wife, Rahime, for her patience and support.

*To my wonderful children,
Seyma Nur, Tuba Nur and Muhammed Enes.
You are always in my heart and the joys in my life.*

*To those who read this book,
and appreciate the work that goes into it, thank you.
If you have any feedback, please let me know.*

Abdulhamit Subasi

Page left intentionally blank

Contents

Preface	xi
Acknowledgments	xiii
1. Introduction	1
1.1 What is machine learning?	1
1.1.1 Why is machine learning needed?	2
1.1.2 Making data-driven decisions	3
1.1.3 Definitions and key terminology	4
1.1.4 Key tasks of machine learning	6
1.1.5 Machine learning techniques	7
1.2 Machine learning framework	7
1.2.1 Data collection	8
1.2.2 Data description	8
1.2.3 Exploratory data analysis	9
1.2.4 Data quality analysis	9
1.2.5 Data preparation	9
1.2.6 Data integration	10
1.2.7 Data wrangling	10
1.2.8 Feature scaling and feature extraction	10
1.2.9 Feature selection and dimension reduction	10
1.2.10 Modeling	11
1.2.11 Selecting modeling techniques	11
1.2.12 Model building	11
1.2.13 Model assessment and tuning	11
1.2.14 Implementation and examining the created model	12
1.2.15 Supervised machine learning framework	12
1.2.16 Unsupervised machine learning framework	13
1.3 Performance evaluation	14
1.3.1 Confusion matrix	16
1.3.2 F-measure analysis	18
1.3.3 ROC analysis	18
1.3.4 Kappa statistic	19
1.3.5 What is measured	20
1.3.6 How they are measured	20
1.3.7 How to interpret estimates	21
1.3.8 <i>k</i> -Fold cross-validation in scikit-learn	21
1.3.9 How to choose the right algorithm	22
1.4 The Python machine learning environment	22
1.4.1 Pitfalls	23
1.4.2 Drawbacks	24
1.4.3 The NumPy library	24
1.4.4 Pandas	24
1.5 Summary	25
References	26
2. Data preprocessing	27
2.1 Introduction	27
2.2 Feature extraction and transformation	28
2.2.1 Types of features	29
2.2.2 Statistical features	29
2.2.3 Structured features	32
2.2.4 Feature transformations	33
2.2.5 Thresholding and discretization	33
2.2.6 Data manipulation	33
2.2.7 Standardization	34
2.2.8 Normalization and calibration	39
2.2.9 Incomplete features	40
2.2.10 Feature extraction methods	42
2.2.11 Feature extraction using wavelet transform	44
2.3 Dimension reduction	53
2.3.1 Feature construction and selection	55
2.3.2 Univariate feature selection	56
2.3.3 Recursive feature elimination	60
2.3.4 Feature selection from a model	61
2.3.5 Principle component analysis (PCA)	62
2.3.6 Incremental PCA	67
2.3.7 Kernel principal component analysis	68
2.3.8 Neighborhood components analysis	70
2.3.9 Independent component analysis	72
2.3.10 Linear discriminant analysis (LDA)	76
2.3.11 Entropy	78
2.4 Clustering for feature extraction and dimension reduction	79
References	88

3. Machine learning techniques	91	3.8 Instance-based learning	200
3.1 Introduction	91	3.9 Summary	201
3.2 What is machine learning?	92	References	201
3.2.1 Understanding machine learning	92	4. Classification examples for healthcare	203
3.2.2 What makes machines learn?	92	4.1 Introduction	203
3.2.3 Machine learning is a multidisciplinary field	93	4.2 EEG signal analysis	203
3.2.4 Machine learning problem	94	4.2.1 Epileptic seizure prediction and detection	204
3.2.5 Goals of learning	95	4.2.2 Emotion recognition	225
3.2.6 Challenges in machine learning	95	4.2.3 Classification of focal and nonfocal epileptic EEG signals	233
3.3 Python libraries	96	4.2.4 Migraine detection	246
3.3.1 Scikit-learn	96	4.3 EMG signal analysis	252
3.3.2 TensorFlow	99	4.3.1 Diagnosis of neuromuscular disorders	253
3.3.3 Keras	99	4.3.2 EMG signals in prosthesis control	262
3.3.4 Building a model with Keras	100	4.3.3 EMG signals in rehabilitation robotics	271
3.3.5 The natural language tool kit	100	4.4 ECG signal analysis	278
3.4 Learning scenarios	103	4.4.1 Diagnosis of heart arrhythmia	279
3.5 Supervised learning algorithms	104	4.5 Human activity recognition	288
3.5.1 Classification	105	4.5.1 Sensor-based human activity recognition	289
3.5.2 Forecasting, prediction, and regression	106	4.5.2 Smartphone-based recognition of human activities	292
3.5.3 Linear models	107	4.6 Microarray gene expression data classification for cancer detection	298
3.5.4 The perceptron	116	4.7 Breast cancer detection	300
3.5.5 Logistic regression	118	4.8 Classification of the cardiocogram data for anticipation of fetal risks	303
3.5.6 Linear discriminant analysis	120	4.9 Diabetes detection	306
3.5.7 Artificial neural networks	124	4.10 Heart disease detection	311
3.5.8 k-Nearest neighbors	128	4.11 Diagnosis of chronic kidney disease (CKD)	314
3.5.9 Support vector machines	133	4.12 Summary	318
3.5.10 Decision tree classifiers	138	References	318
3.5.11 Naive Bayes	145	5. Other classification examples	323
3.5.12 Ensemble methods	148	5.1 Intrusion detection	323
3.5.13 Bagging	149	5.2 Phishing website detection	326
3.5.14 Random forest	154	5.3 Spam e-mail detection	330
3.5.15 Boosting	160	5.4 Credit scoring	334
3.5.16 Other ensemble methods	171	5.5 Credit card fraud detection	338
3.5.17 Deep learning	177	5.6 Handwritten digit recognition using CNN	346
3.5.18 Deep neural networks	179		
3.5.19 Recurrent neural networks	182		
3.5.20 Autoencoders	184		
3.5.21 Long short-term memory (LSTM) networks	184		
3.5.22 Convolutional neural networks	187		
3.6 Unsupervised learning	190		
3.6.1 K-means algorithm	191		
3.6.2 Silhouettes	193		
3.6.3 Anomaly detection	196		
3.6.4 Association rule-mining	199		
3.7 Reinforcement learning	199		

5.7 Fashion-MNIST image classification with CNN	355	7.2.2 Applications of cluster analysis	468
5.8 CIFAR image classification using CNN	364	7.2.3 Number of possible clustering	468
5.9 Text classification	372	7.2.4 Types of clustering algorithms	468
5.10 Summary	387	7.3 The k-means clustering algorithm	469
References	387	7.4 The k-medoids clustering algorithm	471
6. Regression examples	391	7.5 Hierarchical clustering	473
6.1 Introduction	391	7.5.1 Agglomerative clustering algorithm	473
6.2 Stock market price index return forecasting	392	7.5.2 Divisive clustering algorithm	476
6.3 Inflation forecasting	413	7.6 The fuzzy c-means clustering algorithm	481
6.4 Electrical load forecasting	415	7.7 Density-based clustering algorithms	483
6.5 Wind speed forecasting	424	7.7.1 The DBSCAN algorithm	484
6.6 Tourism demand forecasting	429	7.7.2 OPTICS clustering algorithms	486
6.7 House prices prediction	441	7.8 The expectation of maximization for Gaussian mixture model clustering	489
6.8 Bike usage prediction	457	7.9 Bayesian clustering	492
6.9 Summary	462	7.10 Silhouette analysis	494
References	462	7.11 Image segmentation with clustering	497
7. Clustering examples	465	7.12 Feature extraction with clustering	500
7.1 Introduction	465	7.13 Clustering for classification	507
7.2 Clustering	466	7.14 Summary	511
7.2.1 Evaluating the output of clustering methods	467	References	511
		Index	513

Page left intentionally blank

Preface

Rapid developments in machine learning solutions and adoption across various sectors of industry enable the learning of complex models of real-world problems from observed (training) data through systemic solutions in different fields. Significant time and effort are required to create effective machine learning models and achieve reliable outcomes. The main project concepts can be grasped by building robust data pipelines and analyzing and visualizing data using feature extraction, selection, and modeling. Therefore, the extensive need for a reliable machine learning solution involves a development framework that not only is suitable for immersive machine learning modeling but also succeeds in preprocessing, visualization, system integration, and robust support for runtime deployment and maintenance setting. Python is an innovative programming language with multipurpose features, simple implementation and integration, an active developer community, and an ever-increasing machine learning ecosystem, contributing to the expanding adoption of machine learning.

Intelligent structures and data-driven enterprises are becoming a reality, and the developments in techniques and technologies are enabling this to happen. With data being of utmost importance, the market for machine learning and data science practitioners has never been larger than it is now. In fact, the world is facing a shortage of data scientists and machine learning experts. Arguably the most demanding job in the 21st century involves developing some significant expertise in this domain.

Machine learning techniques are computing algorithms, including artificial neural networks, k-nearest neighbor, support vector machines, decision tree algorithms, and deep learning. Machine learning applications are currently of great interest in economics, security, healthcare, biomedicine, and biomedical engineering. This book describes how to use machine learning techniques to analyze the data in these fields.

The author of this book has a great deal of practical experience in the implementation of real-world problems utilizing Python and its machine learning ecosystem. *Practical Machine Learning for Data Analysis Using Python* aims to improve the skill levels of readers and qualify them to create practical machine learning solutions. Moreover, this book is a problem solver's guide for building intelligent real-world systems. It offers a systematic framework that includes principles, procedures, practical examples, and code. The book also contributes to the critical skills needed by its readers to understand and solve various machine learning problems.

This book is an excellent reference for readers developing machine learning techniques by using real-world case studies in the Python machine learning environment. It focuses on building a foundation of machine learning knowledge to solve different case studies from different fields in the real world, including biomedical signal analysis, healthcare, security, economy, and finance. In addition, it focuses on a broad variety of models for machine learning, including regression, classification, clustering, and forecasting.

This book consists of seven chapters. Chapter 1 gives an introduction to data analysis using machine learning techniques. Chapter 2 provides an overview of data pre-processing such as feature extraction, transformation, feature selection, and dimension reduction. Chapter 3 offers an overview of machine learning techniques such as naïve Bayes, k-nearest neighbor, artificial neural networks, support vector machines, decision tree, random forest, bagging, boosting, stacking, voting, deep neural network, recurrent neural network, and convolutional neural networks, for forecasting, prediction, and classification. Chapter 4 presents classification examples for healthcare. It includes electrocardiogram (ECG), electroencephalogram (EEG), and electromyogram (EMG) signal-processing techniques commonly used in the analysis and recognition of biomedical signals. In addition, it presents several medical data classifications, such as human activity recognition, microarray gene expression data classification for cancer detection, breast cancer detection, diabetes detection, and heart disease detection. Chapter 5 considers several applications, including intrusion detection, phishing website detection, spam e-mail detection, credit scoring, credit card fraud detection, handwritten digit recognition, image classification, and text classification. Chapter 6 provides regression examples, such as stock market

analysis, economic variable forecasting, electrical load forecasting, wind speed forecasting, tourism demand forecasting, and house prices prediction. Chapter 7 includes several examples related to unsupervised learning (clustering).

The main intent of this book is to help a wide range of readers to solve their own real-world problems, including IT professionals, analysts, developers, data scientists, and engineers. Furthermore, this book is intended to be a useful textbook for postgraduate and research students working in the areas of data science and machine learning. It also formulates a basis for researchers who are interested in applying machine learning methods to data analysis. In addition, this book will help a broad readership, including researchers, professionals, academics, and graduate students from a wide range of disciplines, who are beginning to look for applications in biomedical signal analysis, healthcare data analysis, financial and economic data forecasting, computer security, and more.

Executing the code examples provided in this book requires Python 3.x or higher versions to be installed on macOS, Linux, or Microsoft Windows. The examples throughout the book frequently utilize the essential libraries of Python, such as SciPy, NumPy, Scikit-learn, matplotlib, pandas, OpenCV, TensorFlow, and Keras, for scientific computing.

Acknowledgments

First of all, I would like to thank my publisher Elsevier and its team of dedicated professionals who have made this book-writing journey very simple and effortless, as well as all those who have worked in the background to make this book a success.

I would like to thank Sara Pianavilla and Rafael Trombaco for their great support.

Also, I would like to thank Paul Prasad Chandramohan for being patient in getting everything necessary to complete this book.

Abdulhamit Subasi