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

ELSEVIER

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

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

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

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 preprocessing 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