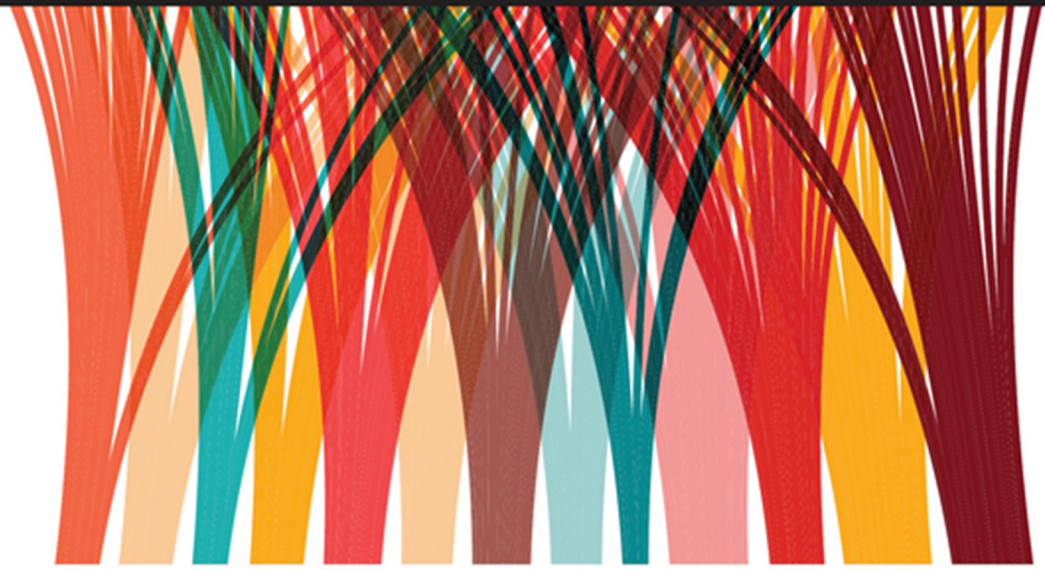
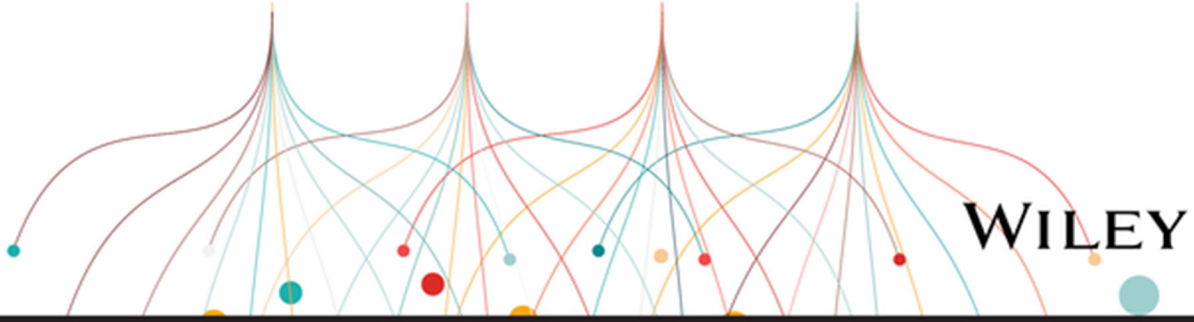


PRACTICAL MACHINE LEARNING IN R

FRED NWANGANGA • MIKE CHAPPLE



WILEY

Practical Machine Learning in R

Practical Machine Learning in R

FRED NWANGANGA
MIKE CHAPPLE

WILEY

Published simultaneously in Canada

ISBN: 978-1-119-59151-1
ISBN: 978-1-119-59153-5 (ebk)
ISBN: 978-1-119-59157-3 (ebk)

Manufactured in the United States of America

Library of Congress Control Number: 2020933607

To my parents, Grace and Friday. I would not be who I am without you. Thanks for always being there. I miss you.

Your loving son,
Chuka

To Ricky. I am so proud of the young man you've become.

Love,
Dad

About the Authors

Fred Nwanganga is an assistant teaching professor of business analytics at the University of Notre Dame's Mendoza College of Business, where he teaches both graduate and undergraduate courses in data management, machine learning, and unstructured data analytics. He has more than 15 years of technology leadership experience in both the private sector and higher education. Fred holds a PhD in computer science and engineering from the University of Notre Dame.

Mike Chapple is an associate teaching professor of information technology, analytics, and operations at the University of Notre Dame's Mendoza College of Business. Mike has more than 20 years of technology experience in the public and private sectors. He serves as academic director of the university's Master of Science in Business Analytics Program and is the author of more than 25 books. Mike earned his PhD in computer science from Notre Dame.

About the Technical Editors

Everaldo Aguiar received his PhD from the University of Notre Dame, where he was affiliated with the Interdisciplinary Center for Network Science and Applications. He is a former data science for social good fellow and now works as a principal data science manager at SAP Concur, where he leads a team of data scientists that develops, deploys, maintains, and evaluates machine learning solutions embedded into customer-facing products.

Seth Berry is an assistant teaching professor in the Information Technology, Analytics, and Operations Department at the University of Notre Dame. He is an avid R user (he is old enough to remember when using Tinn-R was a good idea) and enjoys just about any statistical programming task that comes his way. He is particularly interested in all forms of text analysis and how people's online behaviors can predict real-life decisions.

Acknowledgments

It takes a small army to put together a book, and we are grateful to the many people who collaborated with us on this one.

First and foremost, we thank our families, who once again put up with our nonsense as we were getting this book to press. We'd also like to thank our colleagues in the Information Technology, Analytics, and Operations Department at the University of Notre Dame's Mendoza College of Business. Much of the content in this book started as collegial hallway conversations, and we are thankful to have you in our lives.

Jim Minatel, our acquisitions editor at Wiley, was instrumental in getting this book underway. Mike has worked with Jim for many years and is thankful for his unwavering support. This is Fred's first collaboration with Wiley, and it truly has been a remarkable and rewarding experience.

Our agent, Carole Jelen of Waterside Productions, continues to be a valuable partner, helping us develop new opportunities, including this one.

Our technical editors, Seth Berry and Everaldo Aguiar, gave us invaluable feedback as we worked our way through this book. Thank you for your meaningful contributions to this work.

Our research assistants, Nicholas Schmit and Yun "Jessica" Yan, did an awesome job with literature review and putting together some of the supplemental material for the book.

We'd also like to thank the support crew at Wiley, particularly Kezia Endsley, our project editor, and Vasanth Koilraj, our production editor. You were the glue that kept this project on schedule.

—Fred and Mike

Contents at a Glance

Contents

Introduction

Machine learning is changing the world. Every organization, large and small, seeks to extract knowledge from the massive amounts of information that they store and process on a daily basis. The tantalizing desire to predict the future drives the work of business analysts and data scientists in fields ranging from marketing to healthcare. Our goal with this book is to make the tools of analytics approachable for a broad audience.

The R programming language is a purpose-specific language designed to facilitate statistical analysis and machine learning. We choose it for this book not only due to its strong popularity in the field but also because of its intuitive nature, particularly for individuals approaching it as their first programming language.

There are many books on the market that cover practical applications of machine learning, designed for businesspeople and onlookers. Likewise, there are many deeply technical resources that dive into the mathematics and computer science of machine learning. In this book, we strive to bridge these two worlds. We attempt to bring the reader an intuitive introduction to machine learning with an eye on the practical applications of machine learning in today's world. At the same time, we don't shy away from code. As we do in our undergraduate and graduate courses, we seek to make the R programming language accessible to everyone. Our hope is that you will read this book with your laptop open next to you, following along with our examples and trying your hand at the exercises.

Best of luck as you begin your machine learning adventure!

WHAT DOES THIS BOOK COVER?

This book provides an introduction to machine learning using the R programming language.

Chapter 1: What Is Machine Learning? This chapter introduces the world of machine learning and describes how machine learning allows the discovery of knowledge in data. In this chapter, we explain the differences between unsupervised learning, supervised learning, and reinforcement learning. We describe the differences between classification and regression problems and explain how to measure the effectiveness of machine learning algorithms.

Chapter 2: Introduction to R and RStudio In this chapter, we introduce the R programming language and the toolset that we will be using throughout the rest of the book. We approach R from the beginner's mind-set, explain the use of the RStudio integrated development environment, and walk readers through the creation and execution of their first R scripts. We also explain the use of packages to redistribute R code and the use of different data types in R.

Chapter 3: Managing Data This chapter introduces readers to the concepts of data management and the use of R to collect and manage data. We introduce the tidyverse, a collection of R packages designed to facilitate the analytics process, and we describe different approaches to describing and visualizing data in R. We also cover how to clean, transform, and reduce data to prepare it for machine learning.

Chapter 4: Linear Regression In this chapter, we dive into the world of supervised machine learning as we explore linear regression. We explain the underlying statistical principles behind regression and demonstrate how to fit simple and complex regression models in R. We also explain how to evaluate, interpret, and apply the results of regression models.

Chapter 5: Logistic Regression While linear regression is suitable for problems that require the prediction of numeric values, it is not well-suited to categorical predictions. In this chapter, we describe logistic regression, a categorical prediction technique. We discuss the use of generalized linear models and describe how to build logistic regression models in R. We also explain how to evaluate, interpret, and improve upon the results of a logistic regression model.

Chapter 6: *k*-Nearest Neighbors The *k*-nearest neighbors technique allows us to predict the classification of a data point based on the classifications of other, similar data points. In this chapter, we describe how the *k*-NN process works and demonstrate how to build a *k*-NN model in R. We also show how to apply that model, making predictions about the classifications of new data points.

Chapter 7: Naïve Bayes The naïve Bayes approach to classification uses a table of probabilities to predict the likelihood that an instance belongs to a particular class. In this chapter, we discuss the concepts of joint and conditional probability and describe how the Bayes classification approach functions. We demonstrate building a naïve Bayes classifier in R and use it to make predictions about previously unseen data.

Chapter 8: Decision Trees Decision trees are a popular modeling technique because they produce intuitive results. In this chapter, we describe the creation and interpretation of decision tree models. We also explain the process of growing a tree in R and using pruning to increase the generalizability of that model.

Chapter 9: Evaluating Performance No modeling technique is perfect. Each has its own strengths and weaknesses and brings different predictive power to different types of problems. In this chapter, we discuss the process of evaluating model performance. We introduce resampling techniques and explain how they can be used to estimate the future performance of a model. We also demonstrate how to visualize and evaluate model performance in R.

Chapter 10: Improving Performance Once we have tools to evaluate the performance of a model, we can then apply them to help improve model performance. In this chapter, we look at techniques for tuning machine learning models. We also demonstrate how we can enhance our predictive power by simultaneously harnessing the predictive capability of multiple models.

Chapter 11: Discovering Patterns with Association Rules Association rules help us discover patterns that exist within a dataset. In this chapter, we introduce the association rules approach and demonstrate how to generate association rules from a dataset in R. We also explain ways to evaluate and quantify the strength of association rules.

Chapter 12: Grouping Data with Clustering Clustering is an unsupervised learning technique that groups items based on their similarity to each other. In this chapter, we explain the way that the *k*-means clustering algorithm segments data and demonstrate the use of *k*-means clustering in R.

READER SUPPORT FOR THIS BOOK

In order to make the most of this book, we encourage you to make use of the student and instructor materials made available on the companion site. We also encourage you to provide us with meaningful feedback on ways in which we could improve the book.

Companion Download Files

As you work through the examples in this book, you may choose either to type in all the code manually or to use the source code files that accompany the book. If you choose to follow along with the examples, you will also want to use the same datasets we use throughout the book. All the source code and datasets used in this book are available for download from `www.wiley.com/go/pmlr`.

How to Contact the Publisher

If you believe you've found a mistake in this book, please bring it to our attention. At John Wiley & Sons, we understand how important it is to provide our customers with accurate content, but even with our best efforts an error may occur.

To submit your possible errata, please email it to our customer service team at `wileysupport@wiley.com` with the subject line "Possible Book Errata Submission."

PART I

Getting Started

Chapter 1

What Is Machine Learning?

Welcome to the world of *machine learning*! You're about to embark upon an exciting adventure discovering how data scientists use algorithms to uncover knowledge hidden within the troves of data that businesses, organizations, and individuals generate every day.

If you're like us, you often find yourself in situations where you are facing a mountain of data that you're certain contains important insights, but you just don't know how to extract that needle of knowledge from the proverbial haystack. That's where machine learning can help. This book is dedicated to providing you with the knowledge and skills you need to harness the power of machine learning algorithms. You'll learn about the different types of problems that are well-suited for machine learning solutions and

the different categories of machine learning techniques that are most appropriate for tackling different types of problems.

Most importantly, we're going to approach this complex, technical field with a practical mind-set. In this book, our purpose is not to dwell on the intricate mathematical details of these algorithms. Instead, we'll focus on how you can put those algorithms to work for you immediately. We'll also introduce you to the R programming language, which we believe is particularly well-suited to approaching machine learning problems from a practical standpoint. But don't worry about programming or R for now. We'll get to that in Chapter 2. For now, let's dive in and get a better understanding of how machine learning works.

By the end of this chapter, you will have learned the following:

- How machine learning allows the discovery of knowledge in data
- How unsupervised learning, supervised learning, and reinforcement learning techniques differ from each other
- How classification and regression problems differ from each other
- How to measure the effectiveness of machine learning algorithms
- How cross-validation improves the accuracy of machine learning models