



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

Copyright © 2020 by John Wiley & Sons, Inc., Indianapolis, Indiana

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

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2020933607

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

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

About the Authors	vii	Chapter 6 <i>k</i> -Nearest Neighbors	223
About the Technical Editors	ix	Chapter 7 Naïve Bayes	251
Acknowledgments	xi	Chapter 8 Decision Trees	277
Introduction	xxi		
PART I: Getting Started	1	PART IV: Evaluating and Improving Performance	305
Chapter 1 What Is Machine Learning?	3	Chapter 9 Evaluating Performance	307
Chapter 2 Introduction to R and RStudio	25	Chapter 10 Improving Performance	341
Chapter 3 Managing Data	53	PART V: Unsupervised Learning	367
PART II: Regression	101	Chapter 11 Discovering Patterns with Association Rules	369
Chapter 4 Linear Regression	103	Chapter 12 Grouping Data with Clustering	395
Chapter 5 Logistic Regression	165		
PART III: Classification	221	Index	421

Contents

About the Authors	vii	Chapter 2	
About the Technical Editors	ix	Introduction to R and RStudio	25
Acknowledgments	xi	Welcome to R	26
Introduction	xxi	R and RStudio Components	27
		The R Language	27
		RStudio	28
		RStudio Desktop	28
		RStudio Server	29
		Exploring the RStudio Environment	29
		R Packages	38
		The CRAN Repository	38
		Installing Packages	38
		Loading Packages	39
		Package Documentation	40
		Writing and Running an R Script	41
		Data Types in R	44
		Vectors	45
		Testing Data Types	47
		Converting Data Types	50
		Missing Values	51
		Exercises	52
PART I:		Chapter 3	
Getting Started	1	Managing Data	53
Chapter 1		The Tidyverse	54
What Is Machine Learning?	3	Data Collection	55
Discovering Knowledge in Data	5	Key Considerations	55
Introducing Algorithms	5	Collecting Ground Truth Data	55
Artificial Intelligence, Machine Learning, and Deep Learning	6	Data Relevance	55
Machine Learning Techniques	7		
Supervised Learning	8		
Unsupervised Learning	12		
Model Selection	14		
Classification Techniques	14		
Regression Techniques	15		
Similarity Learning Techniques	16		
Model Evaluation	16		
Classification Errors	17		
Regression Errors	19		
Types of Error	20		
Partitioning Datasets	22		
Holdout Method	23		
Cross-Validation Methods	23		
Exercises	24		

Quantity of Data	56	Regression	114
Ethics	56	Simple Linear Regression	115
Importing the Data	56	Ordinary Least Squares Method	116
Reading Comma-Delimited Files	56	Simple Linear Regression Model	119
Reading Other Delimited Files	60	Evaluating the Model	120
Data Exploration	60	Residuals	121
Describing the Data	61	Coefficients	121
Instance	61	Diagnostics	122
Feature	61	Multiple Linear Regression	124
Dimensionality	62	The Multiple Linear	
Sparsity and Density	62	Regression Model	124
Resolution	62	Evaluating the Model	125
Descriptive Statistics	63	Residual Diagnostics	127
Visualizing the Data	69	Influential Point Analysis	130
Comparison	69	Multicollinearity	133
Relationship	70	Improving the Model	135
Distribution	72	Considering Nonlinear	
Composition	73	Relationships	135
Data Preparation	74	Considering Categorical	
Cleaning the Data	75	Variables	137
Missing Values	75	Considering Interactions	
Noise	79	Between Variables	139
Outliers	81	Selecting the Important	
Class Imbalance	82	Variables	141
Transforming the Data	84	Strengths and Weaknesses	146
Normalization	84	Case Study: Predicting Blood	
Discretization	89	Pressure	147
Dummy Coding	89	Importing the Data	148
Reducing the Data	92	Exploring the Data	149
Sampling	92	Fitting the Simple Linear	
Dimensionality Reduction	99	Regression Model	151
Exercises	100	Fitting the Multiple Linear	
		Regression Model	152
		Exercises	161
PART II:		Chapter 5	
Regression	101	Logistic Regression	165
Chapter 4		Prospecting for Potential Donors	166
Linear Regression	103	Classification	169
Bicycle Rentals and Regression	104	Logistic Regression	170
Relationships Between Variables	106	Odds Ratio	172
Correlation	106		

Binomial Logistic Regression Model	176	Improving the Model	239
Dealing with Missing Data	178	Strengths and Weaknesses	241
Dealing with Outliers	182	Case Study: Revisiting the Donor Dataset	241
Splitting the Data	187	Importing the Data	241
Dealing with Class Imbalance	188	Exploring and Preparing the Data	242
Training a Model	190	Dealing with Missing Data	243
Evaluating the Model	190	Normalizing the Data	245
Coefficients	193	Splitting and Balancing the Data	246
Diagnostics	195	Building the Model	248
Predictive Accuracy	195	Evaluating the Model	248
Improving the Model	198	Exercises	249
Dealing with Multicollinearity	198		
Choosing a Cutoff Value	205	Chapter 7	
Strengths and Weaknesses	206	Naïve Bayes	251
Case Study: Income Prediction	207	Classifying Spam Email	252
Importing the Data	208	Naïve Bayes	253
Exploring and Preparing the Data	208	Probability	254
Training the Model	212	Joint Probability	255
Evaluating the Model	215	Conditional Probability	256
Exercises	216	Classification with Naïve Bayes	257
PART III:		Additive Smoothing	261
Classification	221	Naïve Bayes Model	263
Chapter 6		Splitting the Data	266
<i>k</i>-Nearest Neighbors	223	Training a Model	267
Detecting Heart Disease	224	Evaluating the Model	267
<i>k</i> -Nearest Neighbors	226	Strengths and Weaknesses of the Naïve Bayes Classifier	269
Finding the Nearest Neighbors	228	Case Study: Revisiting the Heart Disease Detection Problem	269
Labeling Unlabeled Data	230	Importing the Data	270
Choosing an Appropriate <i>k</i>	231	Exploring and Preparing the Data	270
<i>k</i> -Nearest Neighbors Model	232	Building the Model	272
Dealing with Missing Data	234	Evaluating the Model	273
Normalizing the Data	234	Exercises	274
Dealing with Categorical Features	235		
Splitting the Data	237	Chapter 8	
Classifying Unlabeled Data	237	Decision Trees	277
Evaluating the Model	238	Predicting Build Permit Decisions	278

Decision Trees	279	Receiver Operating	
Recursive Partitioning	281	Characteristic Curve	333
Entropy	285	Area Under the Curve	336
Information Gain	286	Exercises	339
Gini Impurity	290	Chapter 10	
Pruning	290	Improving Performance	341
Building a Classification Tree		Parameter Tuning	342
Model	291	Automated Parameter Tuning	342
Splitting the Data	294	Customized Parameter Tuning	348
Training a Model	295	Ensemble Methods	354
Evaluating the Model	295	Bagging	355
Strengths and Weaknesses of		Boosting	358
the Decision Tree Model	298	Stacking	361
Case Study: Revisiting the Income		Exercises	366
Prediction Problem	299	PART V:	
Importing the Data	300	Unsupervised	
Exploring and Preparing the		Learning	367
Data	300	Chapter 11	
Building the Model	302	Discovering Patterns	
Evaluating the Model	302	with Association Rules	369
Exercises	304	Market Basket Analysis	370
PART IV:		Association Rules	371
Evaluating and		Identifying Strong Rules	373
Improving Performance	305	Support	373
Chapter 9		Confidence	373
Evaluating Performance	307	Lift	374
Estimating Future Performance	308	The Apriori Algorithm	374
Cross-Validation	311	Discovering Association Rules	376
<i>k</i> -Fold Cross-Validation	311	Generating the Rules	377
Leave-One-Out Cross-		Evaluating the Rules	382
Validation	315	Strengths and Weaknesses	386
Random Cross-Validation	316	Case Study: Identifying Grocery	
Bootstrap Sampling	318	Purchase Patterns	386
Beyond Predictive Accuracy	321	Importing the Data	387
Kappa	323	Exploring and Preparing the Data	387
Precision and Recall	326	Generating the Rules	389
Sensitivity and Specificity	328	Evaluating the Rules	389
Visualizing Model Performance	332		

Exercises	392	The Average Silhouette	
Notes	393	Method	411
Chapter 12		The Gap Statistic	412
Grouping Data with		Strengths and Weaknesses of	
Clustering	395	<i>k</i> -Means Clustering	414
Clustering	396	Case Study: Segmenting	
<i>k</i> -Means Clustering	399	Shopping Mall Customers	415
Segmenting Colleges with <i>k</i> -Means		Exploring and Preparing the Data	415
Clustering	403	Clustering the Data	416
Creating the Clusters	404	Evaluating the Clusters	418
Analyzing the Clusters	407	Exercises	420
Choosing the Right Number of		Notes	420
Clusters	409	Index	421
The Elbow Method	409		

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?

Chapter 2: Introduction to R and RStudio

Chapter 3: Managing Data

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