R语言机器学习

第2版（影印版）

Machine Learning with R

Second Edition

Brett Lantz 著

[PACKT] PUBLISHING

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

R 语言机器学习

第 2 版

(影印版)

Brett Lantz 著

南京　东南大学出版社

图书在版编目(CIP)数据

R语言机器学习:第2版:英文/(美)布雷特·兰茨(Brett Lantz)著. —影印本. —南京:东南大学出版社,2017.4

书名原文:Machine Learning with R, Second Edition

ISBN 978-7-5641-7071-4

Ⅰ.①R… Ⅱ.①布… Ⅲ.①程序语言—程序设计—英文 Ⅳ.①TP312

中国版本图书馆CIP数据核字(2017)第051938号

R语言机器学习 第2版(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼2号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787毫米×980毫米　16开本
印　　张:28.25
字　　数:553千字
版　　次:2017年4月第1版
印　　次:2017年4月第1次印刷
书　　号:ISBN 978-7-5641-7071-4
定　　价:87.00元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Credits

Author
Brett Lantz

Reviewers
Vijayakumar Nattamai Jawaharlal
Kent S. Johnson
Mzabalazo Z. Ngwenya
Anuj Saxena

Commissioning Editor
Ashwin Nair

Acquisition Editor
James Jones

Content Development Editor
Natasha D'Souza

Technical Editor
Rahul C. Shah

Copy Editors
Akshata Lobo
Swati Priya

Project Coordinator
Vijay Kushlani

Proofreader
Safis Editing

Indexer
Monica Ajmera Mehta

Production Coordinator
Arvindkumar Gupta

Cover Work
Arvindkumar Gupta

About the Author

Brett Lantz has spent more than 10 years using innovative data methods to understand human behavior. A trained sociologist, he was first enchanted by machine learning while studying a large database of teenagers' social networking website profiles. Since then, Brett has worked on interdisciplinary studies of cellular telephone calls, medical billing data, and philanthropic activity, among others. When not spending time with family, following college sports, or being entertained by his dachshunds, he maintains `http://dataspelunking.com/`, a website dedicated to sharing knowledge about the search for insight in data.

This book could not have been written without the support of my friends and family. In particular, my wife, Jessica, deserves many thanks for her endless patience and encouragement. My son, Will, who was born in the midst of the first edition and supplied much-needed diversions while writing this edition, will be a big brother shortly after this book is published. In spite of cautionary tales about correlation and causation, it seems that every time I expand my written library, my family likewise expands! I dedicate this book to my children in the hope that one day they will be inspired to tackle big challenges and follow their curiosity wherever it may lead.

I am also indebted to many others who supported this book indirectly. My interactions with educators, peers, and collaborators at the University of Michigan, the University of Notre Dame, and the University of Central Florida seeded many of the ideas I attempted to express in the text; any lack of clarity in their expression is purely mine. Additionally, without the work of the broader community of researchers who shared their expertise in publications, lectures, and source code, this book might not have existed at all. Finally, I appreciate the efforts of the R team and all those who have contributed to R packages, whose work has helped bring machine learning to the masses. I sincerely hope that my work is likewise a valuable piece in this mosaic.

About the Reviewers

Vijayakumar Nattamai Jawaharlal is a software engineer with an experience of 2 decades in the IT industry. His background lies in machine learning, big data technologies, business intelligence, and data warehouse.

He develops scalable solutions for many distributed platforms, and is very passionate about scalable distributed machine learning.

Kent S. Johnson is a software developer who loves data analysis, statistics, and machine learning. He currently develops software to analyze tissue samples related to cancer research. According to him, a day spent with R and ggplot2 is a good day. For more information about him, visit `http://kentsjohnson.com`.

I'd like to thank, Gile, for always loving me.

Mzabalazo Z. Ngwenya holds a postgraduate degree in mathematical statistics from the University of Cape Town. He has worked extensively in the field of statistical consulting, and currently works as a biometrician at a research and development entity in South Africa. His areas of interest are primarily centered around statistical computing, and he has over 10 years of experience with R for data analysis and statistical research. Previously, he was involved in reviewing *Learning RStudio for R Statistical Computing, R Statistical Application Development by Example Beginner's Guide, R Graph Essentials, R Object-oriented Programming, Mastering Scientific Computing with R,* and *Machine Learning with R,* all by Packt Publishing.

Anuj Saxena is a data scientist at IGATE Corporation. He has an MS in analytics from the University of San Francisco and an MSc in Statistics from the NMIMS University in India. He is passionate about data science and likes using open source languages such as R and Python as primary tools for data science projects. In his spare time, he participates in predictive analytics competitions on kaggle.com. For more information about him, visit `http://www.anuj-saxena.com`.

I'd like to thank my father, Dr. Sharad Kumar, who inspired me at an early age to learn math and statistics and my mother, Mrs. Ranjana Saxena, who has been a backbone throughout my educational life.

I'd also like to thank my wonderful professors at the University of San Francisco and the NMIMS University who triggered my interest in this field and taught me the power of data and how it can be used to tell a wonderful story.

---

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.PacktPub.com`.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Preface

Machine learning, at its core, is concerned with the algorithms that transform information into actionable intelligence. This fact makes machine learning well-suited to the present-day era of big data. Without machine learning, it would be nearly impossible to keep up with the massive stream of information.

Given the growing prominence of R—a cross-platform, zero-cost statistical programming environment—there has never been a better time to start using machine learning. R offers a powerful but easy-to-learn set of tools that can assist you with finding data insights.

By combining hands-on case studies with the essential theory that you need to understand how things work under the hood, this book provides all the knowledge that you will need to start applying machine learning to your own projects.

What this book covers

Chapter 1, Introducing Machine Learning, presents the terminology and concepts that define and distinguish machine learners, as well as a method for matching a learning task with the appropriate algorithm.

Chapter 2, Managing and Understanding Data, provides an opportunity to get your hands dirty working with data in R. Essential data structures and procedures used for loading, exploring, and understanding data are discussed.

Chapter 3, Lazy Learning – Classification Using Nearest Neighbors, teaches you how to understand and apply a simple yet powerful machine learning algorithm to your first real-world task—identifying malignant samples of cancer.

Chapter 4, Probabilistic Learning – Classification Using Naive Bayes, reveals the essential concepts of probability that are used in the cutting-edge spam filtering systems. You'll learn the basics of text mining in the process of building your own spam filter.

Chapter 5, Divide and Conquer – Classification Using Decision Trees and Rules, explores a couple of learning algorithms whose predictions are not only accurate, but also easily explained. We'll apply these methods to tasks where transparency is important.

Chapter 6, Forecasting Numeric Data – Regression Methods, introduces machine learning algorithms used for making numeric predictions. As these techniques are heavily embedded in the field of statistics, you will also learn the essential metrics needed to make sense of numeric relationships.

Chapter 7, Black Box Methods – Neural Networks and Support Vector Machines, covers two complex but powerful machine learning algorithms. Though the math may appear intimidating, we will work through examples that illustrate their inner workings in simple terms.

Chapter 8, Finding Patterns – Market Basket Analysis Using Association Rules, exposes the algorithm used in the recommendation systems employed by many retailers. If you've ever wondered how retailers seem to know your purchasing habits better than you know yourself, this chapter will reveal their secrets.

Chapter 9, Finding Groups of Data – Clustering with k-means, is devoted to a procedure that locates clusters of related items. We'll utilize this algorithm to identify profiles within an online community.

Chapter 10, Evaluating Model Performance, provides information on measuring the success of a machine learning project and obtaining a reliable estimate of the learner's performance on future data.

Chapter 11, Improving Model Performance, reveals the methods employed by the teams at the top of machine learning competition leaderboards. If you have a competitive streak, or simply want to get the most out of your data, you'll need to add these techniques to your repertoire.

Chapter 12, Specialized Machine Learning Topics, explores the frontiers of machine learning. From working with big data to making R work faster, the topics covered will help you push the boundaries of what is possible with R.

What you need for this book

The examples in this book were written for and tested with R version 3.2.0 on Microsoft Windows and Mac OS X, though they are likely to work with any recent version of R.

Who this book is for

This book is intended for anybody hoping to use data for action. Perhaps you already know a bit about machine learning, but have never used R; or perhaps you know a little about R, but are new to machine learning. In any case, this book will get you up and running quickly. It would be helpful to have a bit of familiarity with basic math and programming concepts, but no prior experience is required. All you need is curiosity.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "The most direct way to install a package is via the `install.packages()` function."

A block of code is set as follows:

```
subject_name,temperature,flu_status,gender,blood_type
John Doe,        98.1,       FALSE,     MALE,       O
Jane Doe,        98.6,       FALSE,     FEMALE,     AB
Steve Graves,    101.4,      TRUE,      MALE,       A
```

Any command-line input or output is written as follows:

```
> summary(wbcd_z$area_mean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4530 -0.6666 -0.2949  0.0000  0.3632  5.2460
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "The **Task Views** link on the left side of the CRAN page provides a curated list of packages."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at `http://www.packtpub.com` for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

New to the second edition of this book, the example code is also available via GitHub at `https://github.com/dataspelunking/MLwR/`. Check here for the most up-to-date R code, as well as issue tracking and a public wiki. Please join the community!

Downloading the color images of this book

We also provide you with a PDF file that has color images of the screenshots/diagrams used in this book. The color images will help you better understand the changes in the output. You can download this file from `http://www.packtpub.com/sites/default/files/downloads/Machine_Learning_With_R_Second_Edition_ColoredImages.pdf`.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at `questions@packtpub.com`, and we will do our best to address the problem.

Table of Contents