Community Experience Distilled

R Machine Learning Essentials

Gain quick access to the machine learning concepts and practical applications using the R development environment

Michele Usuelli

[PACKT] PUBLISHING open source* community experience distilled

R Machine Learning Essentials

Gain quick access to the machine learning concepts and practical applications using the R development environment

Michele Usuelli

[PACKT] PUBLISHING open source community experience distilled

BIRMINGHAM - MUMBAI

R Machine Learning Essentials

First published: November 2014

Production reference: 1211114

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78398-774-0

www.packtpub.com

Credits

Author
Michele Usuelli

Reviewers
Eric Hare
Jithin S L
Jia Liu
Samir Madhavan
Raghavendra Prasad Narayan
Owen S. Vallis

Acquisition Editor
Subho Gupta

Content Development Editor
Amey Varangaonkar

Technical Editor
Mrunmayee Patil

Copy Editors
Alfida Paiva
Rashmi Sawant
Laxmi Subramanian

Project Coordinator
Leena Purkait

Proofreaders
Simran Bhogal
Ameesha Green
Paul Hindle

Indexer
Monica Ajmera Mehta

Graphics
Abhinash Sahu

Production Coordinator
Alwin Roy

Cover Work
Alwin Roy

About the Author

Michele Usuelli is a data scientist living in London. He has a background of and is passionate about statistics and computer science, and as part of his work, he has explored different software and tools for data analysis and machine learning, focusing on R.

Always wanting to share what he learned from his projects, Michele has written some articles on R-bloggers. R connected to Hadoop and some applications of R tools are the topics covered here.

Michele is passionate about cutting-edge technologies and fast-paced growing environments. Since the very beginning, his work took place in start-up environments. He started his career in one of the most innovative big data start-ups in Milan and worked for a top publishing company in the pricing and analytics division. Currently, he works for a leading R-based company.

I wouldn't have been able to write this book without my personal and professional growth in the last few years, and so I would like to thank all the people I worked with, and of course, my family and friends. I have worked with great people and learned a lot from them.

About the Reviewers

Eric Hare is a graduate from the Department of Statistics at Iowa State University. He graduated from the University of Washington in 2012 with a Bachelor's degree in Statistics and in Computer Engineering. He does research in statistical graphics, statistical computing, and data manipulation. He is currently working on a web application to analyze the statistical properties of Peptide Libraries.

Jithin S L completed his B.Tech in Information Technology from Loyola Institute of Technology and Science. He started his career in the field of analytics and then moved to various verticals of big data technology. He has worked with reputed organizations such as Thomson Reuters, IBM Corporation, and Flytxt, under different roles. He has worked in the banking, energy, healthcare, and telecom domains, and has handled global projects on big data technology.

He has submitted many research papers on technology and business at national and international conferences.

In Albert Einstein's words, *learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning*.

I surrender myself to God Almighty who helped me throughout these days to review this book in an effective way.

I dedicate my work on this book to my dad, Mr. N. Subbian Asari (late), my lovable mom, Mrs. M. Lekshmi, and my sweet sister, Ms. S.L Jishma, for coordinating and encouraging me to produce this book.

Last but not least, I would like to thank all my friends.

Jia Liu obtained her PhD degree in Statistics from Iowa State University. Her research interests are in mixed-effects model, Bayesian method, Bootstrap method, reliability, design of experiments, machine learning, and data mining. She has 3 years of working experience in the pharmaceutical industry.

Samir Madhavan has extensive experience in big data and machine learning. He has worked for the ubiquitous Unique Identification Project, Aadhar, where he was part of the team that helped in developing its fraud module. He was also part of the initial team when Flutura Decision Sciences and Analytics started off. He has created various analytical products, which are being used by the e-commerce, retail, and M2M industries.

Raghavendra Prasad Narayan has completed his Bachelor of Engineering in Electronics and Communication from VTU, Belgaum, and completed his Master's degree majoring in Knowledge Engineering from the National University of Singapore. Since 2009, his area of work has been machine learning and natural language processing (NLP). He has worked on the different problems of NLP, and to solve these problems, he has used the ML algorithms extensively (such as classification, clustering algorithms, feature selection/reduction methods, and graphical models). Other than NLP problems, he has also worked on social network analysis, stock market forecasting, yield predictions, and market mix modeling problems.

Currently, he is working at Meltwater Group in the Data Enrichment team as an NLP engineer.

Owen S. Vallis is currently a professor of Music Technology for the *Music Technology: Interaction, Intelligence, and Design* program at the California Institute of Arts. Owen is a musician, artist, and scientist interested in performance, sound, and technology. As a cofounder of Flipmu and The Noise Index, he explores a diverse range of projects including big data research and sound art installations. He produces, composes, and designs audio processors, and creates new hardware interfaces for musical performance.

Owen received his PhD in 2013 from the New Zealand School of Music, Victoria University, Wellington, and explored contemporary approaches to live computer music. During his graduate research, Owen focused on developing new musical interfaces, interactive musical agents, and large networked music ensembles. Owen graduated as a Bachelor of Arts in Music Technology from the California Institute of the Arts in 2008.

Having lived in Toronto, Canada; Wellington, New Zealand; Tokyo, Japan; San Francisco; Nashville; and Los Angeles; Owen has been able to develop a broad and interesting cross section of musical ideologies and aesthetics. Over the past 10 years, he has worked as a research scientist for Twitter; developed multitouch interfaces for Nokia research labs; worked for the leading ribbon microphone manufacturer Royer Labs; has had musical production featured in major motion films; built a recording facility; and produced, engineered, and mixed records in Tokyo, Nashville, and Los Angeles. Owen's work has been featured in Wired, Future Music, Pitchfork, XLR8R, Processing.org, and various computer arts magazines, and is shown at events such as NASA's Yuri's Night, Google I/O, and the New York Cutlog art festival.

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://www.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.

Table of Contents

Preface

When facing a business problem, machine learning allows you to develop powerful and effective data-driven solutions. The recent explosion of data volume and sources increased the effectiveness of solutions based on data, so this field is becoming more and more valuable. Developing a machine learning solution has specific requirements, and there are some software and tools that support it. A very good option is to use R, which is an open source programming language for statistics supported by a wide international community. The R structure is projected for statistical analysis, and the international community develops the most cutting-edge solutions. For these reasons, R allows you to develop powerful machine learning solutions using just a few lines of code.

There are machine learning tutorials, and they usually require some knowledge of the basics of statistics and computer science. This book is not just a tutorial. It doesn't even require a strong background in statistics or computer science. The target is not to provide you with a complete overview of all the techniques or to teach you how to build sophisticated solutions. This book is a path full of hands-on examples that provide you with the expertise to build a solution to a new problem. The aim is to show the most important concepts behind the approach in such a way that you have a deep understanding of machine learning and are able to identify and use the new algorithms.

What this book covers

Chapter 1, *Transforming Data into Actions*, shows you how new technologies allow you to solve business problems with a data-driven approach.

Chapter 2, *R – A Powerful Tool for Developing Machine Learning Algorithms*, explains why R is a great option for machine learning, and covers the basics of the software.

Chapter 3, *A Simple Machine Learning Analysis*, shows you a simple example of machine learning solutions.

Chapter 4, *Step 1 – Data Exploration and Feature Engineering*, shows you how to clean and transform the data before using machine learning algorithms.

Chapter 5, *Step 2 – Applying Machine Learning Techniques*, shows you how to apply machine learning algorithms to solve the problem.

Chapter 6, *Step 3 – Validating the Results*, shows you how to measure an algorithm's accuracy in order to tune its parameters.

Chapter 7, *Overview of Machine Learning Techniques*, presents the main branches of machine learning algorithms.

Chapter 8, *Machine Learning Examples Applicable to Businesses*, shows you how to solve a business problem using machine learning.

What you need for this book

The only software that you need to run the code is R, preferably 3.0.0+. It is highly recommended, although not necessary, that you install the RStudio Desktop IDE.

Who this book is for

This book is intended for those who want to learn how to perform some machine learning using R, in order to gain insight from their data and to find the solution to some real-life problems. Perhaps you already know a bit about machine learning but have never used R, or perhaps you know a little R but are new to machine learning. In either case, this book will get you up and running quickly. It would be helpful to have a bit of familiarity with basic programming concepts, but no prior experience is required.

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: "Load the `randomForest` package containing the `random forest` algorithm."

A block of code is set as follows:

```
[default]
arrayFeatures <- names(dtBank)
arrayFeatures <- arrayFeatures[arrayFeatures != 'output']
formulaAll <- paste('output', '~')
formulaAll <- paste(formulaAll, arrayFeatures[1])
for(nameFeature in arrayFeatures[-1]){
  formulaAll <- paste(formulaAll, '+', nameFeature)
}
formulaAll <- formula(formulaAll)
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
n1 + n2
[1] 5
n1 * n2
[1] 6
```

New terms and **important words** are shown in bold.

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.

Downloading the color images of this book

We also provide you with a PDF file that has color images of the screenshots used in this book. The color images will help you better understand the changes in the output. You can download this file from `https://www.packtpub.com/sites/default/files/downloads/7740OS_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.

Citations and references

- *Chapter 4*, *Step 1 – Data Exploration and Feature Engineering*, *Chapter 5*, *Step 2 – Applying Machine Learning Techniques*, *Chapter 6*, *Step 3 – Validating the Results*, and flag dataset:

  Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

- *Chapter 8*, *Machine Learning Examples Applicable to Businesses*, and bank dataset:

  [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

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.

1
Transforming Data into Actions

To face a business problem, we need the knowledge and expertise to find its solution. In addition, we also require related data that will help in identifying its solution. This chapter shows how new technologies allow us to build powerful machines that learn from data to give support to business decisions.

The topics that will be covered in this chapter are as follows:

- A general idea for approaching business problems
- The new challenges relating to digital technologies
- How the new tools help in using information
- How the tools identify the information that is not evident
- How the tools can estimate the outcome of future events
- Why R?

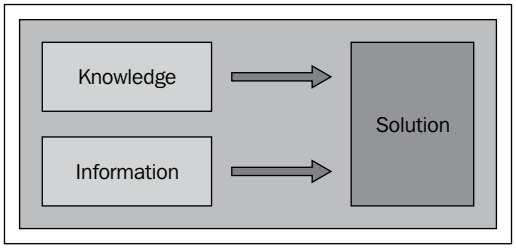
A data-driven approach in business decisions

Expertise and information play important roles in business decisions. This section shows how data-driven technologies changed the approach of facing challenges and improved their solutions.

Business decisions come from knowledge and expertise

The general idea for approaching business problems hasn't changed over the years, and it combines knowledge and information. Before using digital technologies, knowledge came from expertise provided by previous experiences and by other people. With regards to information, it was about analyzing the current situation and comparing it with past events.

A simple example is that of a fruit monger who wants to set the prices of their goods. The price of a product should maximize the profit, which depends on the sales volume and on the price itself. The dealer started their job working with their father who provided them with all their knowledge. Therefore, they already know the price of the different fruits. In addition, at the end of each day, they can observe the amount of each fruit that has been sold. Based on that, they can raise the price of fruits that sold very well and decrease the price of fruits that they didn't sell. This simple example shows how the fruit monger combines domain knowledge and information to solve their problem, as described in the following figure:



This simple example shows how a simple challenge requires a combination of knowledge and data.

The digital era provides more data and expertise

Although the general idea for approaching business problems hasn't changed, digital technologies are providing us with new powerful tools.

The Internet allows people to connect with each other and share their expertise in such a way that everyone has access to a huge set of information. Before the Internet, knowledge came from trusted people and books. Now, the spreading of information has allowed finding books and articles written by different people from every part of the world. In addition, websites and forums allow their users to connect with each other in order to share expertise and find quick answers.

Digital technologies keep track of different activities and produce a lot of related data. We talk about data referring to sets of information—quantitative or qualitative—which is processable by machines. Therefore, when facing a business problem, we can use lots of data from different sources. Some information might not be very relevant, but even after removing it, we often have a huge amount of data. Therefore, we have a lot of improvement potential for the results.

The changes derived from digital technologies involve the process of acquiring expertise and the nature of data. Therefore, the approach to problem solving presents new challenges.

A simple example of a company that faces a business problem is a car dealer who sells different used cars and wants to set the most relevant prices. The car dealer should determine the prices based on the car model, age, and other features. This example is meant to illustrate a possible situation and is not necessarily related to a real problem.

The car dealer needs to identify the best price for each car in order to maximize the revenue. Similar to the fruit monger, if the price of a car is too high, the car dealer won't sell it in a short time, so there will be an extra storage cost and the car will lose value. This leads to an extra cost and a decrease in the profit, thereby damaging the business. On the other hand, if the price is too low, the company will sell the car immediately. Although the storage cost is lower, the company hasn't made the best profit. In order to sell cars and maximize profit, the company wants to figure out the optimal prices.

Let's take a look at the expertise and information that help in finding the solution. The company can use:

- The knowledge of agents who have already sold different cars
- Information from the Internet
- The data about previous sales

The agents can use their past experience, so their knowledge helps in identifying the best prices. However, it's not enough to set the prices when the market changes quickly.

The Internet gives us a lot of information since there are many online shopping websites displaying the prices of used cars. Online shopping is different from the physical market, but an expert agent can take a look at the websites and compare the prices. In this way, the agent can combine their expertise with the online information and identify the right prices in a good way.