Sinan Ozdemir 著

数据科学原理

（影印版）

Principles of Data Science

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Packt

数据科学原理(影印版)

Principles of Data Science

Sinan Ozdemir 著

南京　东南大学出版社

图书在版编目(CIP)数据

数据科学原理:英文/(美)思南·约茨德米尔(Sinan Ozdemir)著. —影印本. —南京:东南大学出版社,2017.10

书名原文:Principles of Data Science

ISBN 978-7-5641-7364-7

Ⅰ.①数… Ⅱ.①思… Ⅲ.①数据处理—英文 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2017)第 196705 号

图字:10-2017-119 号

数据科学原理(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:24.25
字　　数:475 千字
版　　次:2017 年 10 月第 1 版
印　　次:2017 年 10 月第 1 次印刷
书　　号:ISBN 978-7-5641-7364-7
定　　价:92.00 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
Sinan Ozdemir

Reviewers
Samir Madhavan
Oleg Okun

Acquisition Editor
Sonali Vernekar

Content Development Editor
Samantha Gonsalves

Technical Editor
Anushree Arun Tendulkar

Copy Editor
Shaila Kusanale

Project Coordinator
Devanshi Doshi

Proofreaders
Safis Editing

Indexer
Tejal Daruwale Soni

Graphics
Jason Monteiro

Production Coordinator
Melwyn Dsa

Cover Work
Melwyn Dsa

About the Author

Sinan Ozdemir is a data scientist, startup founder, and educator living in the San Francisco Bay Area with his dog, Charlie; cat, Euclid; and bearded dragon, Fiero. He spent his academic career studying pure mathematics at Johns Hopkins University before transitioning to education. He spent several years conducting lectures on data science at Johns Hopkins University and at the General Assembly before founding his own start-up, Legion Analytics, which uses artificial intelligence and data science to power enterprise sales teams.

After completing the Fellowship at the Y Combinator accelerator, Sinan has spent most of his days working on his fast-growing company, while creating educational material for data science.

I would like to thank my parents and my sister for supporting me through life, and also, my various mentors, including Dr. Pam Sheff of Johns Hopkins University and Nathan Neal, the chapter adviser of my collegiate leadership fraternity, Sigma Chi.

Thank you to Packt Publishing for giving me this opportunity to share the principles of data science and my excitement for how this field will impact all of our lives in the coming years.

About the Reviewers

Samir Madhavan has over six years of rich data science experience in the industry and has also written a book called *Mastering Python for Data Science*. He started his career with Mindtree, where he was a part of the fraud detection algorithm team for the UID (Unique Identification) project, called Aadhar, which is the equivalent of a Social Security number for India. After this, he joined Flutura Decision Sciences and Analytics as the first employee, where he was part of the core team that helped the organization scale to an over a hundred members. As a part of Flutura, he helped establish big data and machine learning practice within Flutura and also helped out in business development. At present, he is leading the analytics team for a Boston-based pharma tech company called Zapprx, and is helping the firm to create data-driven products that will be sold to its customers.

Oleg Okun is a machine learning expert and an author/editor of four books, numerous journal articles, and conference papers. His career spans more than a quarter of a century. He was employed in both academia and industry in his mother country, Belarus, and abroad (Finland, Sweden, and Germany). His work experience includes document image analysis, fingerprint biometrics, bioinformatics, online/offline marketing analytics, and credit-scoring analytics.

He is interested in all aspects of distributed machine learning and the Internet of Things. Oleg currently lives and works in Hamburg, Germany.

I would like to express my deepest gratitude to my parents for everything that they have done for me.

---

www.PacktPub.com

eBooks, discount offers, and more

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 `customercare@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/mapt`

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

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

Preface

The topic of this book is data science, which is a field of study and application that has been growing rapidly for the past several decades. As a growing field, it is gaining a lot of attention in both the media as well as in the job market. The United States recently appointed its first ever chief data scientist, DJ Patil. This move was modeled after tech companies who, honestly, only recently started hiring massive data teams. These skills are in high demand and their applications extend much further than today's job market.

This book will attempt to bridge the gap between math/programming/domain expertise. Most people today have expertise in at least one of these (maybe two), but proper data science requires a little bit of all three. We will dive into topics from all three areas and solve complex problems. We will clean, explore, and analyze data in order to derive scientific and accurate conclusions. Machine learning and deep learning techniques will be applied to solve complex data tasks.

What this book covers

Chapter 1, How to Sound Like a Data Scientist, gives an introduction to the basic terminology used by data scientists and a look at the types of problem we will be solving throughout this book.

Chapter 2, Types of Data, looks at the different levels and types of data out there and how to manipulate each type. This chapter will begin to deal with the mathematics needed for data science.

Chapter 3, The Five Steps of Data Science, uncovers the five basic steps of performing data science, including data manipulation and cleaning, and sees examples of each step in detail.

Chapter 4, Basic Mathematics, helps us discover the basic mathematical principles that guide the actions of data scientists by seeing and solving examples in calculus, linear algebra, and more.

Chapter 5, Impossible or Improbable – a Gentle Introduction to Probability, is a beginner's look into probability theory and how it is used to gain an understanding of our random universe.

Chapter 6, Advanced Probability, uses principles from the previous chapter and introduces and applies theorems, such as the Bayes Theorem, in the hope of uncovering the hidden meaning in our world.

Chapter 7, Basic Statistics, deals with the types of problem that statistical inference attempts to explain, using the basics of experimentation, normalization, and random sampling.

Chapter 8, Advanced Statistics, uses hypothesis testing and confidence interval in order to gain insight from our experiments. Being able to pick which test is appropriate and how to interpret p-values and other results is very important as well.

Chapter 9, Communicating Data, explains how correlation and causation affect our interpretation of data. We will also be using visualizations in order to share our results with the world.

Chapter 10, How to Tell If Your Toaster Is Learning – Machine Learning Essentials, focuses on the definition of machine learning and looks at real-life examples of how and when machine learning is applied. A basic understanding of the relevance of model evaluation is introduced.

Chapter 11, Predictions Don't Grow on Trees, or Do They?, looks at more complicated machine learning models, such as decision trees and Bayesian-based predictions, in order to solve more complex data-related tasks.

Chapter 12, Beyond the Essentials, introduces some of the mysterious forces guiding data sciences, including bias and variance. Neural networks are introduced as a modern deep learning technique.

Chapter 13, Case Studies, uses an array of case studies in order to solidify the ideas of data science. We will be following the entire data science workflow from start to finish multiple times for different examples, including stock price prediction and handwriting detection.

What you need for this book

This book uses Python to complete all of its code examples. A machine (Linux/Mac/Windows OK) with access to a Unix-style Terminal with Python 2.7 installed is required. Installation of the Anaconda distribution is also recommended as it comes with most of the packages used in the examples.

Who this book is for

This book is for people who are looking to understand and utilize the basic practices of data science for any domain.

The reader should be fairly well acquainted with basic mathematics (algebra, perhaps probabilities) and should feel comfortable reading snippets of R/Python as well as pseudocode. The reader is not expected to have worked in a data field; however, they should have the urge to learn and apply the techniques put forth in this book to either their own datasets or those provided to them.

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: "For these operators, keep the `boolean` datatype in mind."

A block of code is set as follows:

```
tweet = "RT @j_o_n_dnger: $TWTR now top holding for
          Andor, unseating $AAPL"

words_in_tweet = first_tweet.split(' ') # list of words in tweet
```

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

```
for word in words_in_tweet:                # for each word in list
  if "$" in word:                          # if word has a "cashtag"
  print "THIS TWEET IS ABOUT", word  # alert the user
```

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 for this book from your account at `http://www.packtpub.com`. 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.

You can download the code files by following these steps:

1. Log in or register to our website using your e-mail address and password.
2. Hover the mouse pointer on the **SUPPORT** tab at the top.
3. Click on **Code Downloads & Errata**.

4. Enter the name of the book in the **Search** box.
5. Select the book for which you're looking to download the code files.
6. Choose from the drop-down menu where you purchased this book from.
7. Click on **Code Download**.

You can also download the code files by clicking on the **Code Files** button on the book's webpage at the Packt Publishing website. This page can be accessed by entering the book's name in the **Search** box. Please note that you need to be logged in to your Packt account.

Once the file is downloaded, please make sure that you unzip or extract the folder using the latest version of:

- WinRAR / 7-Zip for Windows
- Zipeg / iZip / UnRarX for Mac
- 7-Zip / PeaZip for Linux

The code bundle for the book is also hosted on GitHub at `https://github.com/PacktPublishing/Principles-of-Data-Science`. We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

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 `https://www.packtpub.com/sites/default/files/downloads/PrinciplesofDataScience_ColorImages.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 the 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