O'REILLY®

TURING 图灵程序设计丛书

数据科学入门

Data Science from Scratch

[美] Joel Grus 著
高蓉 韩波 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

数据科学入门

Data Science from Scratch

First Principles with Python

[美] Joel Grus 著
高蓉 韩波 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

数据科学入门 /（美）格鲁斯（Grus, J.）著 ；高蓉，韩波译. -- 北京 ：人民邮电出版社，2016.3
（图灵程序设计丛书）
ISBN 978-7-115-41741-1

Ⅰ. ①数… Ⅱ. ①格… ②高… ③韩… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第025033号

内 容 提 要

本书基于易于理解且具有数据科学相关的丰富的库的Python语言环境，从零开始讲解数据科学工作。具体内容包括：Python速成，可视化数据，线性代数，统计，概率，假设与推断，梯度下降法，如何获取数据，k近邻法，朴素贝叶斯算法，等等。作者借助大量具体例子以及数据挖掘、统计学、机器学习等领域的重要概念，详细展示了什么是数据科学。

本书适合有志成为数据科学工作者以及想了解数据科学的读者阅读。

◆ 著 [美] Joel Grus
译 高 蓉 韩 波
责任编辑 朱 巍
执行编辑 张 曼
责任印制 杨林杰

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：19
字数：451千字 2016年3月第1版
印数：1-3 500册 2016年3月北京第1次印刷
著作权合同登记号 图字：01-2015-8112号

定价：69.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第 8052 号

版权声明

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版、在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——*Wired*

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

前言

数据科学

有人称数据科学家为“21 世纪头号性感职业”（https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/）。虽说如此称呼有些夸张，但这个名称对数据科学的推崇却一点也没错，这是一个蓬勃发展、前途无限的行业。很多分析师都预言，未来十年会需要比现在多得多的数据科学工作者。

那么，什么是数据科学？唯有正确理解数据科学，才能培养出数据科学家。根据广受业界赞誉的文氏图（http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram），数据科学是以下几个方面的交叉：

- 黑客技能
- 数学和统计学知识
- 专业技能

我原本很想写一本能涵盖以上三个方面的书，但很快意识到仅关于专业技能的撰写就会耗费上万页笔墨，于是及时放弃转而专注于前两个方面。我的目标有两个：一是帮助读者掌握从事数据科学工作所必需的黑客技能；二是帮助读者熟悉数学和统计学，这是数据科学的核心。

对一本书来说，这两个愿望有点大了。学习黑客技能的最好方法就是钻研技术。通过阅读本书，你可以理解我钻研技术的方式，但相同的方式对你未必最适合；你可以理解我使用的一些工具，但相同的工具对你来说未必最顺手；你可以理解我如何解决数据问题，但相同的方式对你来说未必最有效。举例的目的和希望是启发你以自己的方式和方法完成工作。本书涵盖的所有代码和数据都可以从 GitHub 上下载。

同样，学习数学的最好方式就是研习数学。当然本书并不是一部数学著作，我们在本书中大半也不会“研习数学”，我想强调的是数学知识对从事数据科学工作至关重要。不理解

概率、统计、线性代数，就无法真正开始数据科学工作。在需要的地方，书中会引入数学方程式、数学直觉、数学公理，以及借以阐释大数学思想的卡通漫画。有我在，别怕！

总之，数据科学相当有趣（尤其和税务筹划或者煤矿开采等其他工作相比）。

从零开始

很多很多的数据科学库、框架、模块、工具箱可以有效地实现数据科学大部分常见的（和不常见的）算法与技术。如果你是一位数据科学家，就会非常熟悉 NumPy、scikit-learn、pandas 以及其他库。这些库对数据科学工作至关重要。如果还没有真正理解数据科学，运用这些库也是开始数据科学工作的好方式。

在本书中，我们从零开始着手数据科学工作。这意味着为了获得更好的理解，我们需要自己亲手构建工具和实现算法。我花费了很多心思选择注释良好、简洁易读的实现范例。在大部分情形下，所建立的工具意义清晰但实用性有限，它们对规模较小的示例数据集运转良好，但对“网络级别”的数据集就束手无策了。

在全书中，我会向读者指出相应的库，用以将相应技术运用于大规模数据集，但本书中我们不会使用它们。

对学习数据科学，一直有这样一种积极的争论，即什么样的语言环境最好？许多人认为是统计语言 R。（我们说，他们错了。）还有一些人认为是 Java 或者 Scala。而我认为，Python 才是最佳选择！

对于学习和从事数据科学工作，Python 具有几大优势：

- 免费；
- 编程相对简单（尤其是也易于理解）；
- 具有很多数据科学相关的库。

我不敢说 Python 是我最爱的编程语言，因为的确存在其他一些更舒适、设计更棒、编程更有乐趣的语言。但是，每当着手一个新的数据科学项目时，我最终使用的是 Python；每当需要快速构建某个有效程序的原型时，我使用的是 Python；每当需要用简洁易懂的方式表达数据科学概念时，我使用的还是 Python。于是，本书也采用 Python。

但是，教授 Python 不是本书的目的（尽管通过学习本书你会学到一些 Python 知识）。本书会用一章快速介绍 Python 的重要特征，这些特征与本书目的紧密相关。倘若读者没有 Python 基础（或编程基础），那需要再补充阅读一些关于 Python 的入门指导。

本书数据科学导论的其余部分采取了类似的书写方式，在必要或需要阐明时才深入细节，否则省略细节留给读者自己去挖掘（或者在维基百科上查阅）。

过去我曾培训过许多数据科学家。不是每个人都会努力变成改变世界的明星级数据忍者，但所有人都通过培训成为了更棒的数据科学家。我越来越相信，任何拥有一定数学基础和编程技术的人，只要再匹配一些基本材料就可以从事数据科学工作。必需品是好奇心、勤奋工作的态度，还有本书。没错，就是本书！

本书排版约定

本书使用了下列排版约定。

- 楷体
  表示新术语。
- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。
- 等宽粗体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。
- 等宽斜体（*`constant width italic`*）
  表示应该由用户输入的值或根据上下文确定的值替换的文本。

该图标表示提示或建议。

该图标表示一般注释。

该图标表示警告或警示。

示例代码的使用

本书的补充材料（示例代码、练习等）都可以从 GitHub 下载：https://github.com/joelgrus/

data-science-from-scratch。

本书提供代码的目的是帮你快速完成工作。一般情况下，你可以在你的程序或文档中使用本书中的代码，而不必取得我们的许可，除非你想复制书中很大一部分代码。例如，你在编写程序时，用到了本书中的几个代码片段，这不必取得我们的许可。但若将 O'Reilly 图书中的代码制成光盘并进行出售或传播，则需获得我们的许可。引用示例代码或书中内容来解答问题无需许可。将书中很大一部分的示例代码用于你个人的产品文档，这需要我们的许可。

如果你引用了本书的内容并标明版权归属声明，我们对此表示感谢，但这不是强制的。版权归属声明通常包括：标题、作者、出版社和 ISBN，例如："*Data Science from Scratch* by Joel Grus (O'Reilly). Copyright 2015 Joel Grus, 978-1-4919-0142-7"。

如果你认为你对示例代码的使用已经超出上述范围，或者你对是否需要获得示例代码的授权还不清楚，请随时联系我们：permissions@oreilly.com 。

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）是应运而生的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。技术专家、软件开发人员、Web 设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将 Safari Books Online 视作获取资料的首选渠道。

对于组织团体（https://www.safaribooksonline.com/enterprise/）、政府机构（https://www.safaribooksonline.com/government/）、教育机构（https://www.safaribooksonline.com/academic-public-library/）和个人，Safari Books Online 提供各种产品组合和灵活的定价策略（https://www.safaribooksonline.com/pricing/）。用户可通过一个功能完备的数据库检索系统访问 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 以及其他几十家出版社（https://www.safaribooksonline.com/our-library/）的上千种图书、培训视频和正式出版之前的书稿。要了解 Safari Books Online 的更多信息，我们网上见。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网站地址是：
http://shop.oreilly.com/product/0636920033400.do

对于本书的评论和技术性问题，请发送电子邮件到：bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：
http://www.oreilly.com

我们在 Facebook 的地址如下：http://facebook.com/oreilly

请关注我们的 Twitter 动态：http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：http://www.youtube.com/oreillymedia

致谢

首先，我深深地感谢 Mike Loukides 接受我写作本书的提议（并对本书的篇幅提出了合理的建议）。其实他可以选择更轻松的方式，例如可以说："那个总发样章给我的是什么人？我该怎样拒绝他？"但他没有，我感激不尽！同样，我很感谢我的编辑 Marie Beaugureau，她在整个出版过程中给予我指导，并最终使本书呈现出比我独立完成更棒的状态。

如果我从未学习数据科学，怎么能写出这样一本书？如果没有 Dave Hsu、Igor Tatarinov、John Rauser 和 Farecast 群组其他人的影响，我不大可能学习数据科学（当时甚至还没有数据科学这个名称）。还要感谢免费大型公开课程项目 Coursera 的优秀教师们。

同样，我很感谢本书的试读者和评论者。Jay Fundling 找出了许多错误，并指出很多含混的表达，因而本书得到了很大的改善，非常感谢。Debashis Ghosh 全面检查了本书的统计学部分。本书原本表达了肯定 Python 否定 R 的看法，Andrew Musselman 建议淡化这种立场，我后来体会到这是金玉良言。我也非常感谢 Trey Causey、Ryan Matthew Balfanz、Loris Mularoni、Núria Pujol、Rob Jefferson、Mary Pat Campbell、Zach Geary 和 Wendy Grus

给我的宝贵反馈。当然，本书其余的问题，责任在我。

我非常感谢 Twitter 上的数据科学社区，它让我接触到很多新奇的概念，让我认识了很多大牛。我深感自己的欠缺，需要写一本书来弥补。（再次）特别感谢 Trey Causey，他（并非刻意地）提醒我在书中加一章线性代数的内容。同样感谢 Sean J. Taylor，他（并非刻意地）指出了数据处理一章中的若干重大缺漏。

最后，向 Ganga 和 Madeline 致以我无限的感恩。比写一本书更痛苦的事，莫过于和写这本书的人朝夕相对。本书得以完成，全赖家人的支持与鼓励。

目录