HZ BOOKS 华章IT

[PACKT] PUBLISHING

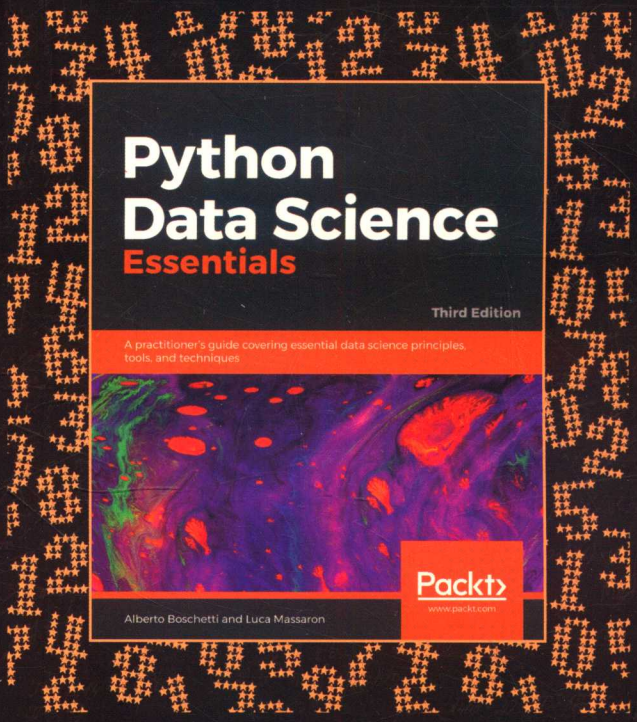
数据科学与工程技术丛书

数据科学导论

Python语言

（原书第3版）

[意] 阿尔贝托·博斯凯蒂（Alberto Boschetti）
卢卡·马萨罗（Luca Massaron） 著
于俊伟 译



PYTHON DATA SCIENCE ESSENTIALS

THIRD EDITION

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PYTHON DATA SCIENCE ESSENTIALS

THIRD EDITION

数据科学导论

Python语言

（原书第3版）

[意] 阿尔贝托·博斯凯蒂（Alberto Boschetti）
卢卡·马萨罗（Luca Massaron）
著

于俊伟 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据科学导论：Python 语言（原书第 3 版）/（意）阿尔贝托·博斯凯蒂，（意）卢卡·马萨罗著；于俊伟译．—北京：机械工业出版社，2020.2
（数据科学与工程技术丛书）
书名原文：Python Data Science Essentials, Third Edition

ISBN 978-7-111-64669-3

I. 数… II. ① 阿… ② 卢… ③ 于… III. 软件工具－程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2020）第 023683 号

本书版权登记号：图字 01-2018-8349

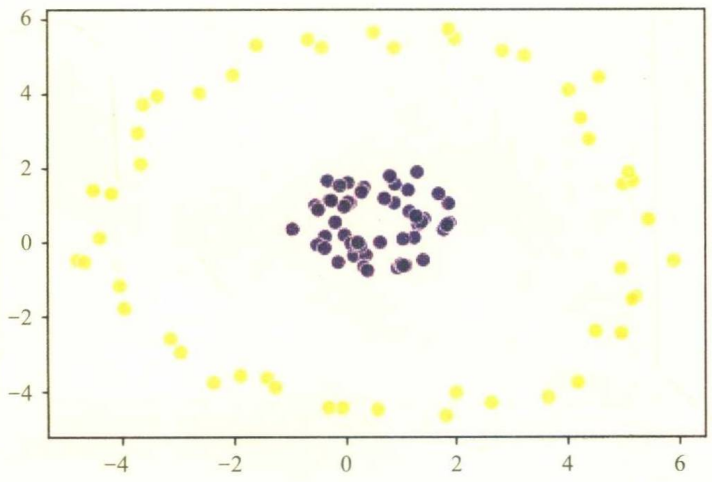
数据科学导论：Python 语言（原书第 3 版）

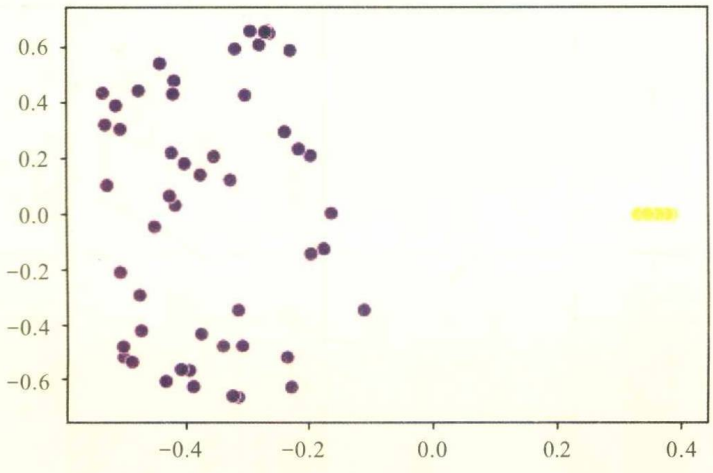
出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：罗丹琪　　责任校对：李秋荣
印　刷：大厂回族自治县益利印刷有限公司　　版　次：2020 年 3 月第 1 版第 1 次印刷
开　本：185mm×260mm 1/16　　印　张：18.75（含彩插 0.25 印张）
书　号：ISBN 978-7-111-64669-3　　定　价：79.00 元

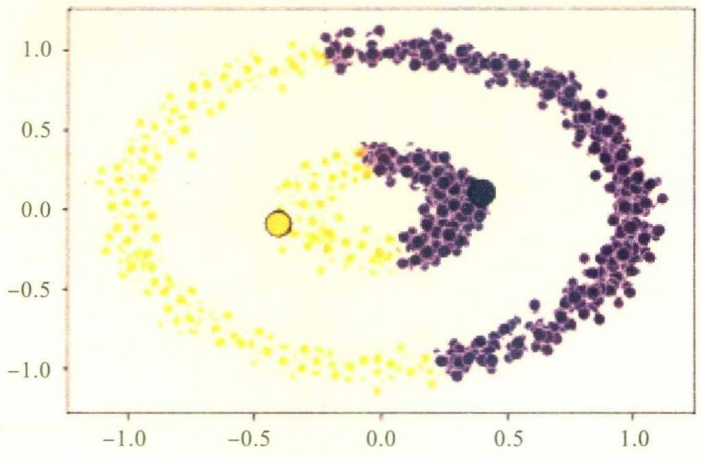
客服电话：（010）88361066 88379833 68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

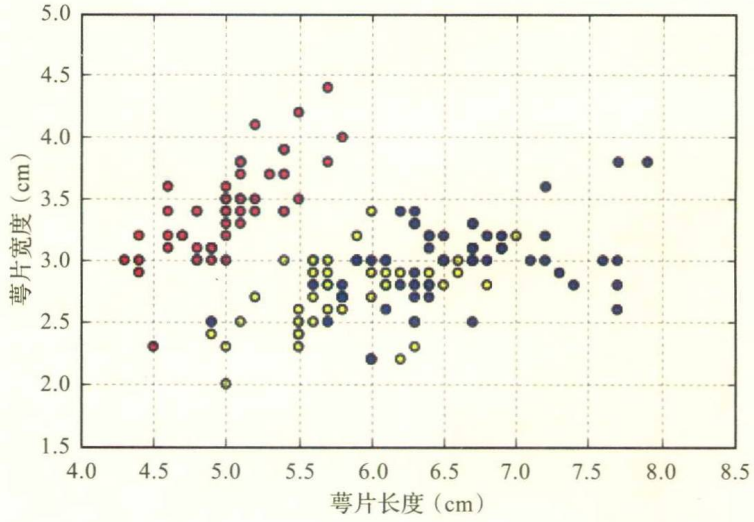
华章图书

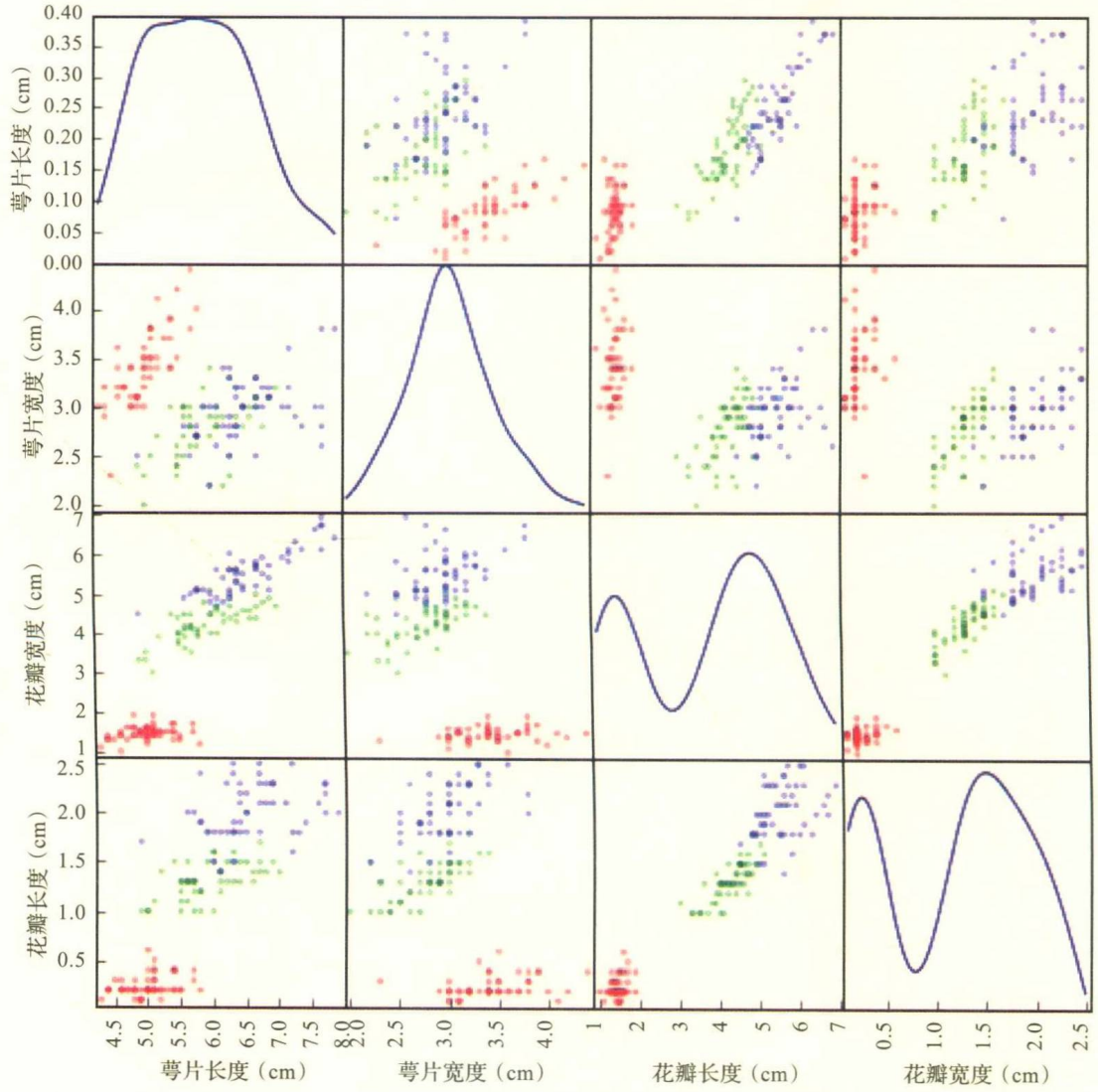
一本打开的书，
一扇开启的门，
通向科学殿堂的阶梯，
托起一流人才的基石。

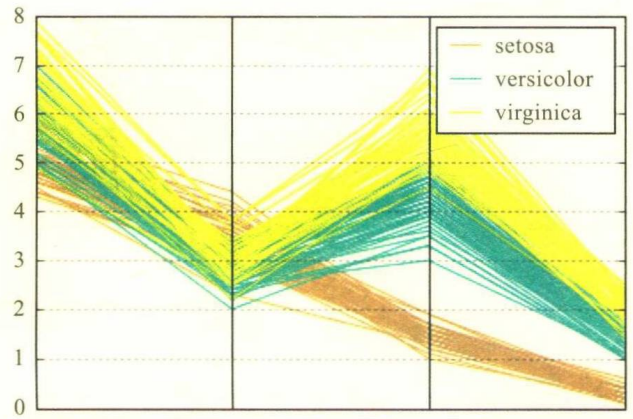

6
4
2
0
−2
−4
−4
−2
0
2
4
6

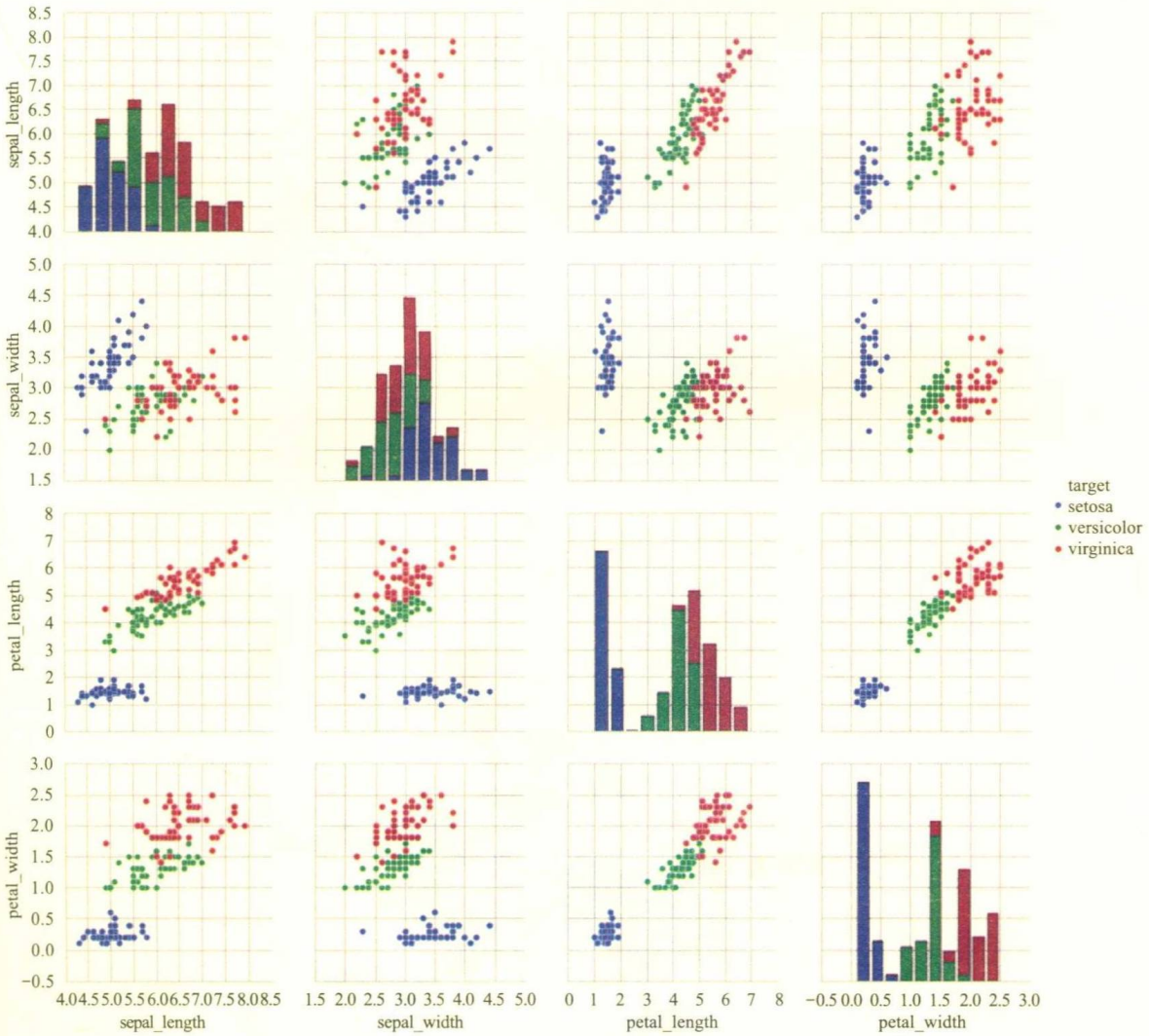

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4

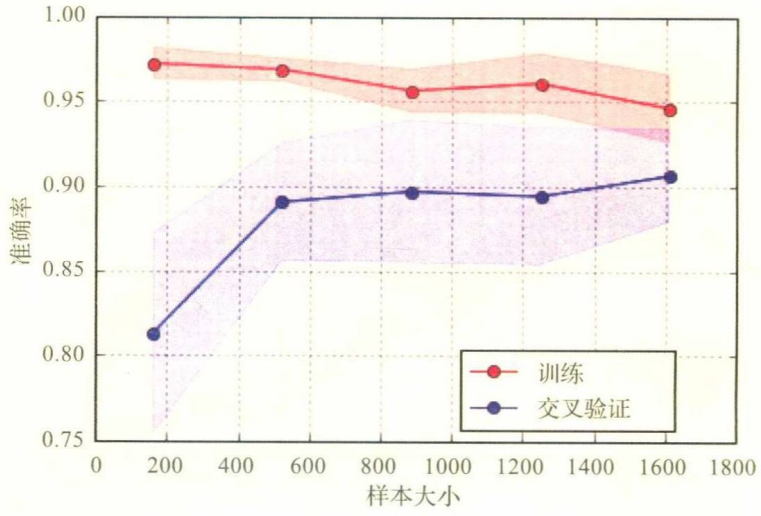

1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

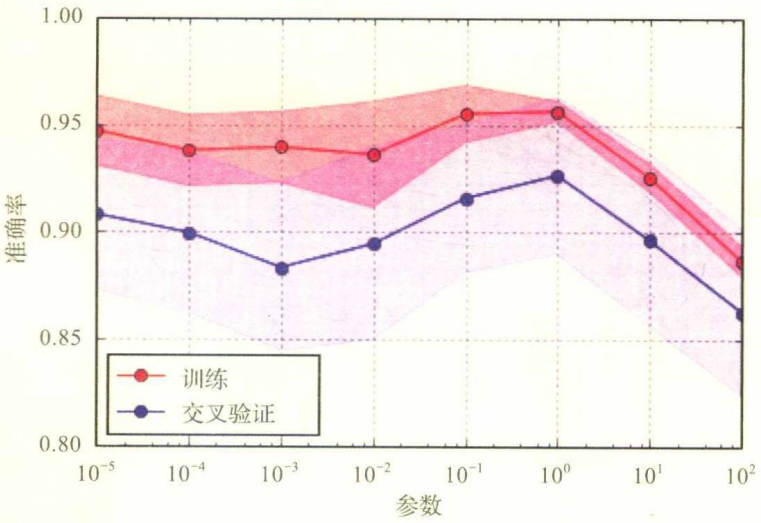

萼片宽度（cm）
萼片长度（cm）

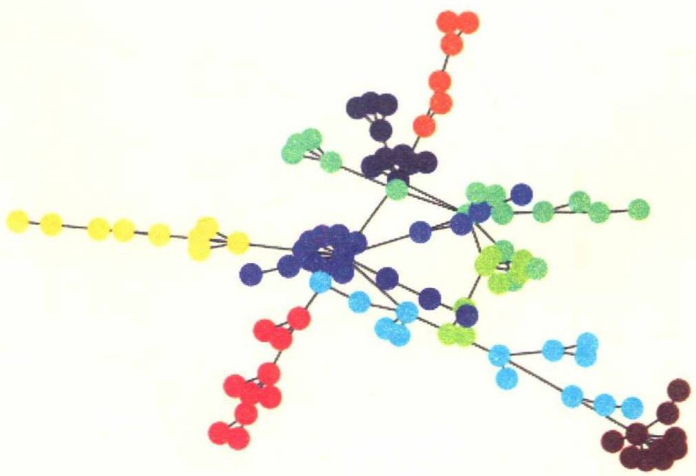

萼片长度（cm）
萼片宽度（cm）
花瓣宽度（cm）
花瓣长度（cm）
萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）


setosa
versicolor
virginica
萼片长度（cm）萼片宽度（cm）花瓣长度（cm）花瓣宽度（cm）


sepal_length
sepal_width
petal_length
petal_width
target
setosa
versicolor
virginica


1.00
0.95
0.90
0.85
0.80
0.75
准确率
0
200
400
600
800
1000
1200
1400
1600
1800
样本大小
训练
交叉验证


1.00
0.95
0.90
0.85
0.80
准确率
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
参数
训练
交叉验证



译 者 序

我们正处于一个快速发展的信息化时代，人们每天都会和各种各样的数据打交道，与此同时，数据也在极大地影响着我们的生活。数据科学就是这样一种涉及数据获取、处理、建模、分析和应用的新兴学科，吸引了众多科技公司、从业人员的关注和研究。

工欲善其事，必先利其器。那么，什么才是数据科学家最值得信赖的专业工具呢？Python 无疑是众多数据分析语言中最适合的一个。Python 是一种通用的、解释性和面向对象的语言，具有强大的数据分析和机器学习软件包，为解决各种数据科学问题提供了快速、可靠、成熟的开发环境。它易学易用，便于快速开发，有很好的交互式体验，并且已经征服了科学界，堪称解决数据科学问题的神器。

本书介绍了进行数据科学分析和开发的所有关键要点，包括：Python 软件及相关工具包的安装和使用；数据加载、运算和改写等基本数据准备过程，以及特征选择、维数约简等高级数据操作方法；由训练、验证、测试等过程组成的数据科学流程，并结合示例深入浅出地讲解多种机器学习算法；数据可视化工具及数据学习表示方法；基于图模型的社交网络创建、分析和处理方法；当前最热的深度学习和大数据处理技术。

考虑到深度学习和大数据处理技术在图像识别、自然语言处理、物联网等领域的应用，本书第 3 版增加了两章单独对它们进行详细介绍。新增加的第 7 章深入介绍了深度学习原理，结合 Keras 库便捷地搭建各种深度学习网络，通过实例演示了深度学习在交通标志分类和电影评论情感预测中的应用。新增加的第 8 章介绍了基于 Spark 的大数据获取和处理方法，涵盖了大数据处理的分布式框架、数据格式、变量共享等内容，并在 KDD99 数据集上构建有效的机器学习算法。

本书提供了简洁的示例代码和多种实用数据集，可帮助你快速掌握数据科学和机器学习相关过程和算法原理。翻译过程中我们对第 2 版中的部分表述和代码格式进行了改进，对书中的代码和链接进行了测试和验证。本书可以作为高等院校数据科学、信息处理、机器学习和人工智能等学科的学习教材，也可以作为 Python 数据科学开发和应用人员的参考书。

本书由河南工业大学信息科学与工程学院于俊伟博士翻译。同时，感谢靳小波博士在前两版中文版中对机器学习部分做出的贡献；感谢薛程和赵胜等同学对新增加两章内容的整理和讨论；感谢刘琼、宋雪菲两位同学认真阅读了本书初稿，耐心细致地找出了

本书相对于第2版的改进，她们的工作很大程度上使本书避免了漏译和错译。本书翻译工作得到了河南省研究生教育教学改革研究与实践项目（2017SJGLX046Y）和河南工业大学第二批（2015年）青年骨干教师培育计划项目的资助。感谢机械工业出版社华章公司对本书出版的高度重视，特别感谢刘锋编辑对本书翻译和出版提供的帮助。

最后，还要感谢家人的爱和包容，谢谢你们的陪伴和支持！

于俊伟

2019年8月

前　言

“千里之行，始于足下。”

——老子（公元前 604 年—公元前 531 年）

数据科学属于一门相对较新的知识领域，它成功融合了线性代数、统计建模、可视化、计算语言学、图形分析、机器学习、商业智能、数据存储和检索等众多学科。

Python 编程语言在过去十年已经征服了科学界，现在是数据科学实践者不可或缺的工具，也是每一个有抱负的数据科学家的必备工具。Python 为数据分析、机器学习和算法求解提供了快速、可靠、跨平台、成熟的开发环境。无论之前在数据科学应用中阻止你掌握 Python 的原因是什么，我们将通过简单的分步化解和示例导向的方法帮你解决，以便你在演示数据集和实际数据集上使用最直接有效的 Python 工具。

作为第 3 版，本书对第 2 版的内容进行了更新和扩展。以最新的 Jupyter Notebook 和 JupyterLab 界面（结合可互换的内核，一个真正的多语言数据科学系统）为基础，本书包含了 Numpy、pandas 和 Scikit-learn 等库的所有主要更新。此外，本书还提供了不少新内容，包括新的梯度提升机器（GBM）算法（XGBoost、LightGBM 和 CatBoost）、深度学习（通过提供基于 Tensorflow 的 Keras 解决方案）、漂亮的数据可视化（主要使用 seaborn）和 Web 部署（使用 bottle）等。

本书首先介绍了如何在最新版 Python 3.6 中安装基本的数据科学工具箱，本书采用单源方法（这意味着本书代码也可以在 Python 2.7 上重用）。接着，将引导你进入完整的数据改写和预处理阶段，主要阐述用于数据分析、探索或处理的数据加载、变换、修复等关键数据科学活动。最后，本书将完成数据科学精要的概述，介绍主要的机器学习算法、图分析技术、所有可视化工具和部署工具，其中可视化工具更易于向数据科学专家或商业用户展示数据处理结果。

读者对象

如果你有志于成为数据科学家，并拥有一些数据分析和 Python 方面的基础知识，本书将助你在数据科学领域快速入门。对于有 R 语言或 MATLAB/GNU Octave 编程经验的数据分析人员，本书也可以作为一个全面的参考书，提高他们在数据操作和机器学习方面的技能。

本书内容

第 1 章介绍 Jupyter Notebook，演示怎样使用教程中的数据。

第 2 章介绍所有关键的数据操作和转换技术，重点介绍数据改写的最佳实践。

第 3 章讨论所有可能改进数据科学项目结果的操作，使读者能够进行高级数据操作。

第 4 章介绍 Scikit-learn 库中最重要的机器学习算法。向读者展示具体的实践应用，以及为了使每种学习技术获得最佳效果，应重点检查什么和调整哪些参数。

第 5 章为你提供基础和中高级图形表示技术，这对复杂数据结构和机器学习所得结果的表示和视觉理解是必不可少的。

第 6 章向读者提供处理社交关系和交互数据实用而有效的技术。

第 7 章演示了如何从零开始构建卷积神经网络，介绍了行业内增强深度学习模型的所有工具，解释了迁移学习的工作原理，以及如何使用递归神经网络进行文本分类和时间序列预测。

第 8 章介绍了一种处理数据的新方法——水平缩放大数据。这意味着要运行安装了 Hadoop 和 Spark 框架的机器集群。

附录包括一些 Python 示例和说明，重点介绍 Python 语言的主要特点，这些都是从事数据科学工作必须了解的。

阅读准备

为了充分利用本书，读者需要具备以下知识：

- 熟悉基本的 Python 语法和数据结构（比如，列表和字典）。
- 具有数据分析基础知识，特别是关于描述性统计方面的知识。

通过阅读本书，也可以帮你培养这两方面的技能，尽管本书没有过多地介绍其中的细节，而只提供了一些最基本的技术，数据科学家为了成功运行其项目必须了解这些技术。

你还需要做好以下准备：

- 一台装有 Windows、MacOS 或 Linux 操作系统的计算机，至少 8GB 的内存（如果你的计算机内存只有 4 GB，还是可以运行本书大部分示例的）。
- 如果加速第 7 章中的计算，最好安装一个 GPU。
- 安装 Python 3.6，推荐通过 Anaconda 来安装（https://www.anaconda.com/download/）。

下载示例代码及彩色图像

本书的示例代码及彩图，可以从 http://www.packtpub.com 通过个人账号下载，也可以访问华章图书官网 http://www.hzbook.com，通过注册并登录个人账号下载。

作者简介

阿尔贝托·博斯凯蒂（Alberto Boschetti）数据科学家、信号处理和统计学方面的专家。他是通信工程专业博士，现在在伦敦居住和工作。他主要从事自然语言处理、行为分析、机器学习和分布式处理等方面的挑战性工作。他对工作充满激情，经常参加学术聚会、研讨会及其他学术活动，紧跟数据科学技术发展的前沿。

我要感谢我的家人、朋友和同事！同时，也非常感谢开源社区！

卢卡·马萨罗（Luca Massaron）数据科学家、市场营销研究主导者，是多元统计分析、机器学习和客户洞察方面的专家。有十年以上解决实际问题的经验，使用推理、统计、数据挖掘和算法为利益相关者创造了巨大的价值。他是意大利网络受众分析的先锋，并在Kaggler上获得排名前十的佳绩，随后一直热心参与各种与数据及数据分析相关的活动，积极给新手和专业人员讲解数据驱动知识发现的潜力。他崇尚大道至简，坚信理解数据科学的精要能给你带来巨大收获。

致Yukiko和Amelia，谢谢你们的爱和包容。

“前路无止境，星云作伴长，双脚虽远行，终归还家乡。”

审阅者简介

Pietro Marinelli 一直致力于人工智能、文本分析和其他数据科学技术的研究，曾为多个行业的企业工作，在数据产品设计方面拥有 10 多年的经验。他提出了多种算法，从预测建模到高级仿真算法，以支持不同跨国公司高级管理者的业务决策。多年来，他一直名列 Kaggle 世界顶尖数据科学家之列，在意大利数据科学家中排名第三。

Matteo Malosetti 是一名数学工程师，在保险行业担任数据科学家。他热衷于自然语言处理应用和贝叶斯统计方面的研究工作。

目　录