HZ BOOKS 华章IT

[PACKT] PUBLISHING

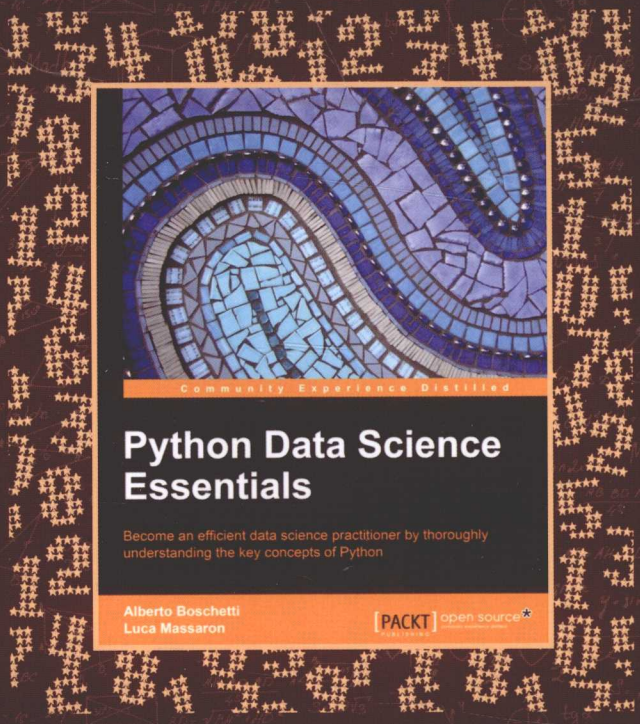
数据科学与工程技术丛书

数据科学导论

Python语言实现

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

于俊伟 靳小波 译



PYTHON DATA SCIENCE ESSENTIALS

机械工业出版社
China Machine Press

科学与工程技术丛书

PYTHON DATA SCIENCE ESSENTIALS

数据科学导论

Python语言实现

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

于俊伟 靳小波 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 语言实现 /（意）阿尔贝托·博斯凯蒂（Alberto Boschetti），（意）卢卡·马萨罗（Luca Massaron）著；于俊伟，靳小波译．—北京：机械工业出版社，2016.7
（数据科学与工程技术丛书）
书名原文：Python Data Science Essentials

ISBN 978-7-111-54434-0

I. 数… II. ①阿… ②卢… ③于… ④靳… III. 数据处理软件－程序设计 IV. TP274

中国版本图书馆 CIP 数据核字（2016）第 176067 号

本书版权登记号：图字：01-2015-7673

数据科学导论：Python 语言实现

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：余　洁　　责任校对：董纪丽
印　　刷：北京瑞德印刷有限公司　　版　　次：2016 年 8 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：12.25（含彩插 0.25 印张）
书　　号：ISBN 978-7-111-54434-0　　定　　价：49.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

华章IT

HZBOOKS | Information Technology

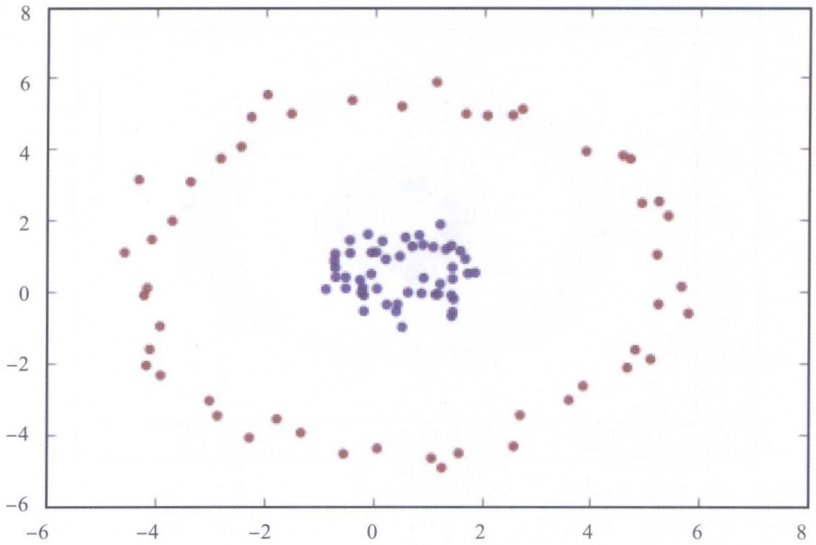


图 1

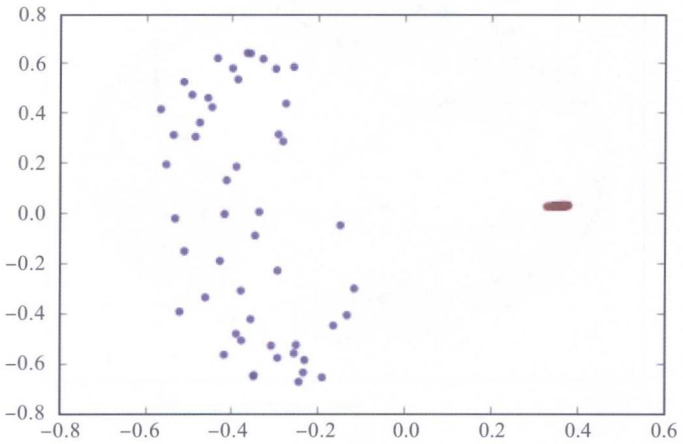


图 2

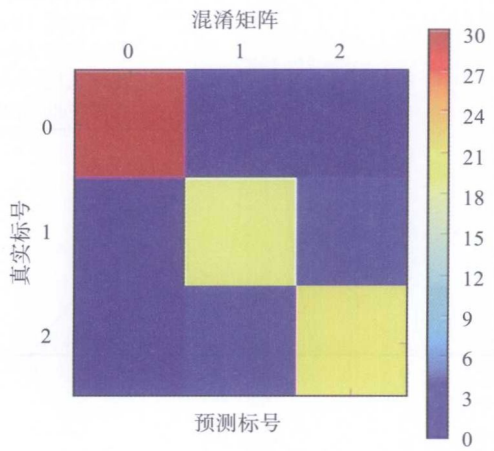


图 3

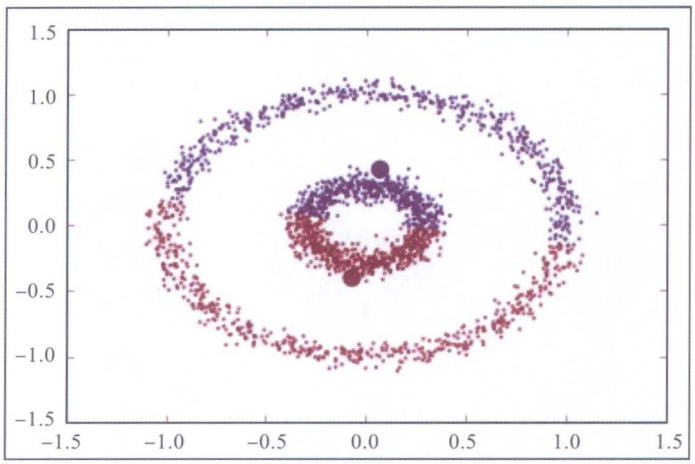

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

图 4

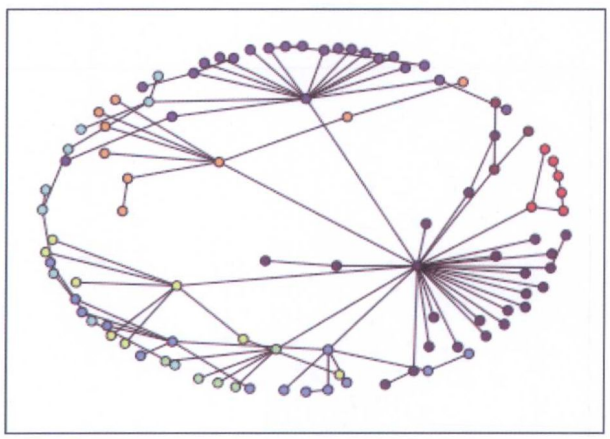


图 5

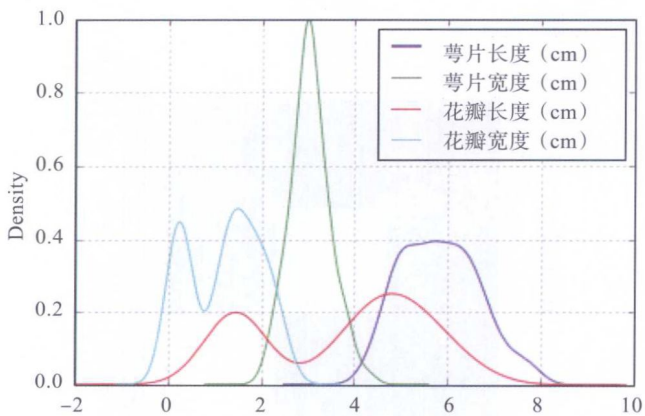

1.0
0.8
0.6
0.4
0.2
0.0
Density
−2
0
2
4
6
8
10
萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）

图 6

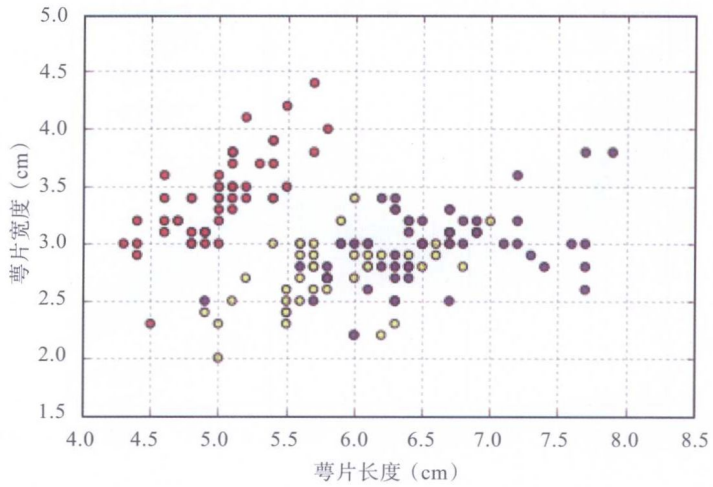


图　7

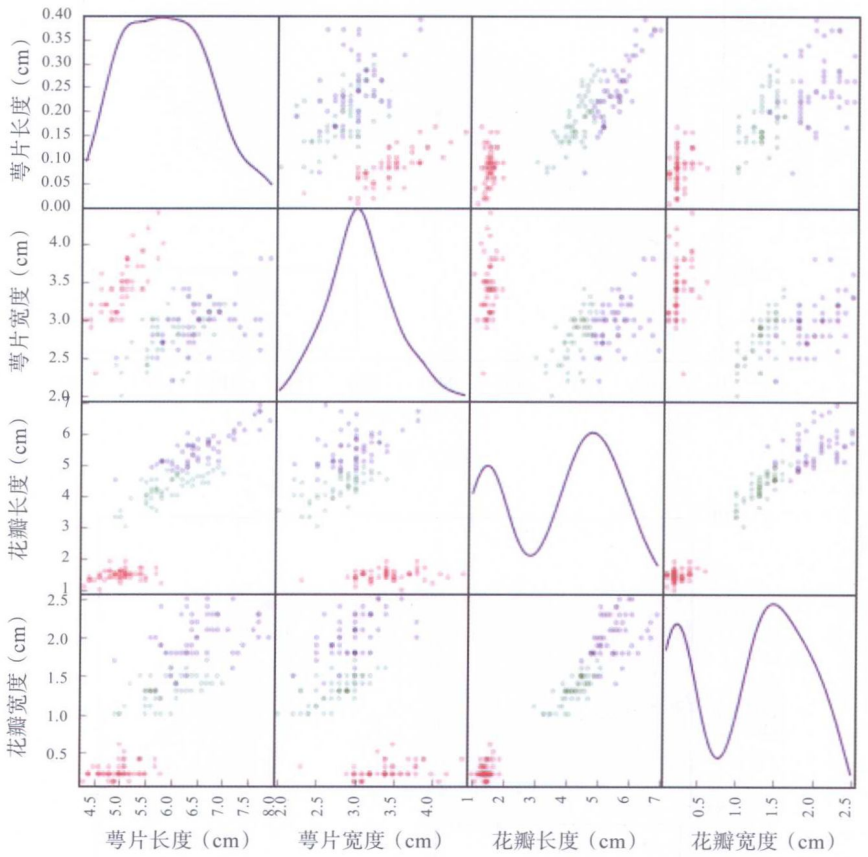


图　8

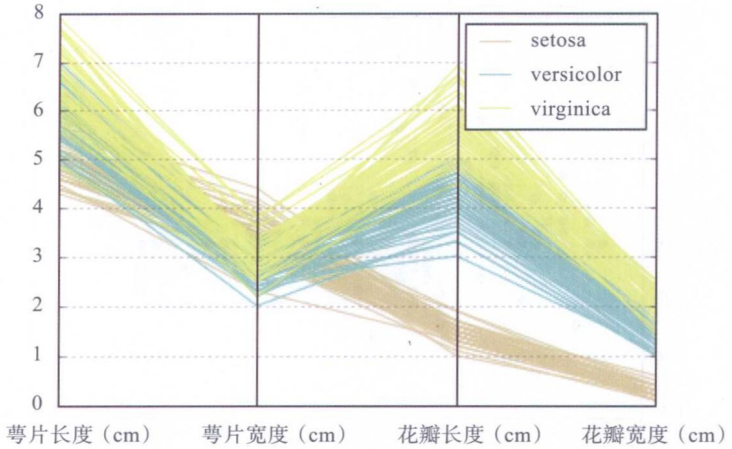

8
7
6
5
4
3
2
1
0
setosa
versicolor
virginica
萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）

图　9

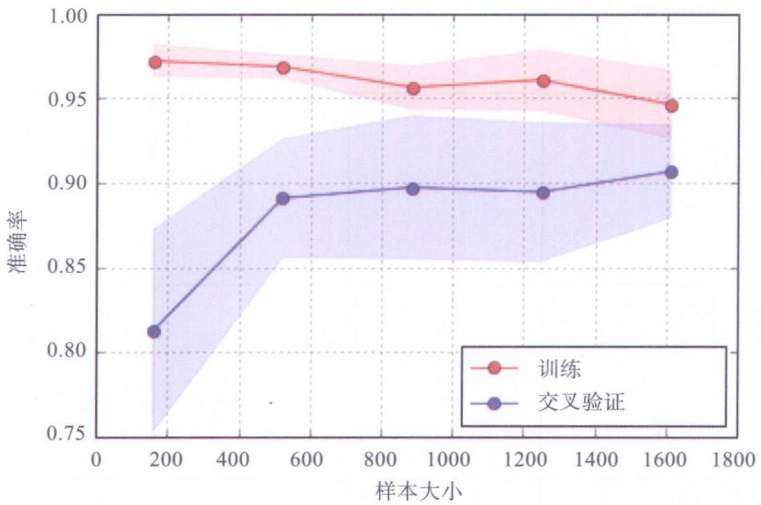

1.00
0.95
0.90
0.85
0.80
0.75
准确率
训练
交叉验证
0
200
400
600
800
1000
1200
1400
1600
1800
样本大小

图　10

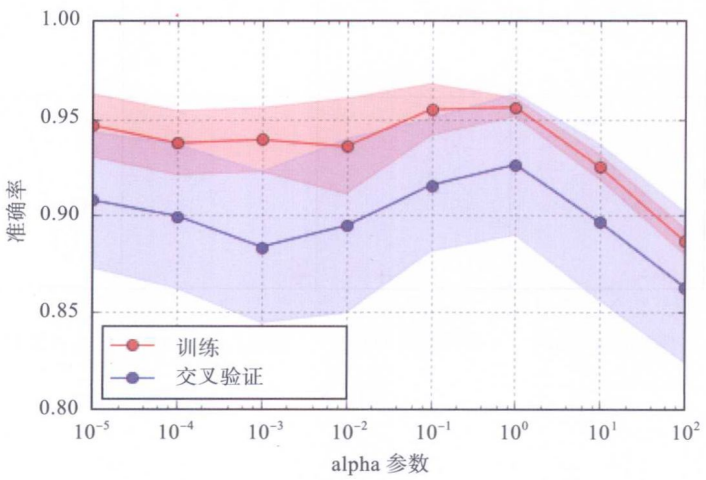

1.00
0.95
0.90
0.85
0.80
准确率
训练
交叉验证
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
alpha 参数

图　11

译 者 序

我们正处于一个快速发展的信息化时代，人们每天都在生产着各种类型的数据，与此同时，数据也在极大地影响着我们的生活。于是，数据成为与能源同等重要的资源。掌握了数据获取、处理、建模、分析等过程的理论和方法，无疑就是掌握了打开这种新型资源的钥匙。

数据科学是融合多种学科的新的知识领域，一般要求学习者或从业者具备统计学等数学知识、计算机相关学科专业知识和特定业务领域的知识。目前，数据科学领域的研究和应用备受瞩目，吸引了众多研究者、实践者和从业者的参与，他们都在积极探索数据科学的基本理论、研究方法和技术应用。

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，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；不仅包含数据加载、运算和改写等基本数据准备过程，还有特征选择、维数约简等高级数据操作方法；建立了由训练、验证、测试等过程组成的数据科学流程，结合示例深入浅出地讲解了多种机器学习算法；介绍了基于图模型的社会网络创建、分析和处理方法；最后是数据分析结果的可视化及相关工具使用方法的介绍。

本书作者是两位意大利数据科学专家，他们长期从事与数据科学相关的教学和科研工作，在 Python 社区、社交网络上也很活跃，发表了多篇学术论文和著作，对数据科学相关人员影响很大。本书是作者多年实践经验的总结，具有以下特点：1）循序渐进，深入浅出，让初学者不畏惧，让从业者得要领。2）理论与实践相结合，几乎所有算法和理论都辅以简洁的实例和说明，通过简单的几行代码即可验证。3）深入理解数据科学概念，轻松进行理论扩展，快速建立自己的工程，使读者做到学以致用，促进多种形式的科学研究和应用开发。

无论是作为数据科学和机器学习理论研究者的参考书，还是作为使用 Python 进行数据

科学应用开发人员的工具书，抑或作为有志成为数据科学家的初学者的指导书，本书都能提供非常有价值的参考。本书还可以作为高等院校相关学科本科生或研究生的学习教材，特别适合从事数据科学、信息处理和机器学习等方向的研究生进行学习和参考。

本书第 4 章由河南工业大学信息科学与工程学院靳小波博士翻译，其余章节由河南工业大学信息科学与工程学院于俊伟博士翻译。由于译者水平有限，加之时间仓促，错误和疏漏在所难免，恳请读者批评指正。

本书的翻译工作受到国家自然科学基金项目（61300123）的资助。还要感谢机械工业出版社华章公司的编辑为本书出版付出的辛勤劳动。

最后，要特别感谢爱人刘楠及女儿 Cynthia 对我工作的理解和支持！

于俊伟

2016 年 3 月

前　言

“千里之行，始于足下。”

——老子（公元前 604—531 年[⊖]）

数据科学属于相对较新的知识领域，它需要成功融合线性代数、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，它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、机器学习和算法问题求解提供了快速、可靠、跨平台、成熟的开发环境。无论之前数据科学应用中阻止你掌握 Python 的原因是什么，这些都将通过我们简单的分步化解和示例导向的方法来解决，我们将帮助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。

借助你现有的 Python 语法和结构知识（不要担心，如果你需要获取更多的 Python 知识，我们有一些 Python 教程），本书将从介绍建立基本的数据科学工具箱开始。接着，它将引导你进入完整的数据改写和预处理阶段。我们还需要花一定量的时间来解释数据类型的转换、修复、探索和处理等核心活动。然后，我们将演示高级数据科学操作，建立变量和假设选择的实验流程，优化超参数，有效地使用交叉验证和测试。最后，我们将完成数据科学精要的概述，介绍主要的机器学习算法、图的分析技术和所有用于呈现结果的可视化方法。

在数据科学项目的具体演示过程中，永远都伴有清晰的代码和简化的例子，以帮助你理解项目背后的机制和实际数据集。本书也会给你一些经验提示，帮助你立即上手当前的项目。准备好了吗？相信你已经准备踏上这个漫长而又值得期待的旅程了。

本书内容

第 1 章介绍所有必需的基础工具（用于交互计算的 shell 命令、库和数据集），使用 Python 可以立即开始数据科学分析。

第 2 章阐明如何加载要处理的数据，当数据太大计算机不能处理时要采用替代技术。本

⊖ 目前国内比较认可的老子生卒年分别是公元前 571 年和公元前 471 年。译者有幸生于老子故里，对老子的传说和史料有所了解，但众多考证都只能给出一个大概的年限。这里译者对作者严谨的引述表示敬意，或许以后利用数据科学技术能从众多史料中挖掘出更确切的老子生平。——译者注

章介绍了所有主要的数据操作和转换技术。

第 3 章提供了高级数据探索和操作技术，使用复杂的数据操作进行特征创建和精简、数据异常检测、验证技术应用等。

第 4 章带你学习 Scikit-learn 库中最重要的学习算法，演示了实际应用以及为了获得每种机器学习技术的最佳结果，指出了应该重点检查的关键数值和要调试的参数。

第 5 章详细介绍了一些实用又有效的数据处理技术，用于处理表示社会实体之间的关系或相互作用的数据。

第 6 章利用图形化表示完善数据科学概述。如果你想形象地表示复杂的数据结构、机器学习过程和结果，这些可视化技术是不可或缺的。

阅读准备

本书提到的 Python 及其他数据科学工具，从 IPython 到 Scikit-learn 都能在网上免费下载。要运行本书附带的源代码，需要一台带有 Windows、Linux 或 Mac OS 操作系统的计算机。本书将分步介绍 Python 解释器以及运行示例所需要的其他工具和数据的安装过程。

读者对象

本书基于你已经具备的一些核心技能，能使你变成高效的数据科学从业者。因此，我们假定你具有编程和统计学方面的基础知识。

本书提供的示例代码不需要你精通 Python 语言，但是假设你至少了解一些基础知识，如 Python 脚本编写、列表和字典数据结构、类对象的工作原理等。在阅读本书之前，花几个小时学习一下第 1 章推荐的网络课程，就可以快速获得这些知识，当然也可以学习其他相关教程。

本书并不需要高级数据科学的概念，我们提供的信息足够帮助你理解本书示例用到的核心概念。

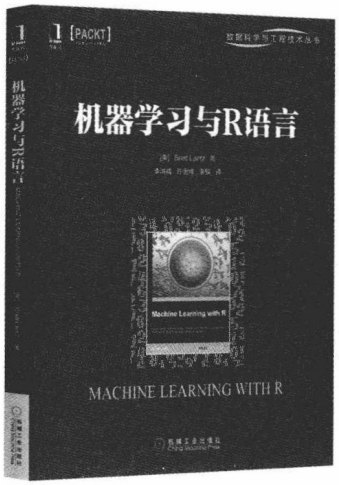
总的来说，本书适合以下人员：

- 有较少的 Python 编程经验和数据分析知识，但还没有数据科学算法等专业知识，有志于成为数据科学家的新手。
- 能熟练运用 R 和 Matlab 等工具进行统计建模、愿意利用 Python 进行数据科学处理的数据分析师。
- 有意学习数据操作和机器学习、不断拓展知识面的开发者和程序员。

代码下载

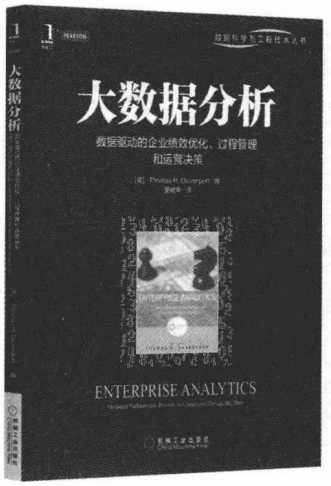
读者可登录华章网站 http://www.hzbook.com 下载本书示例代码。

推荐阅读



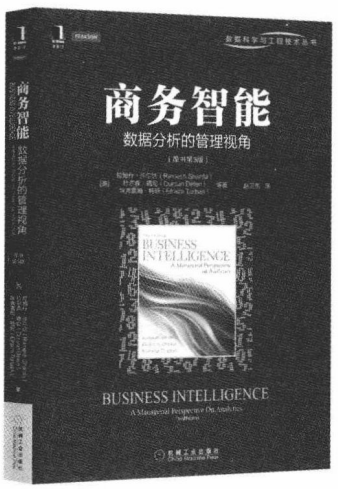
机器学习与R语言

作者：Brett Lantz ISBN：978-7-111-49157-6 定价：69.00元



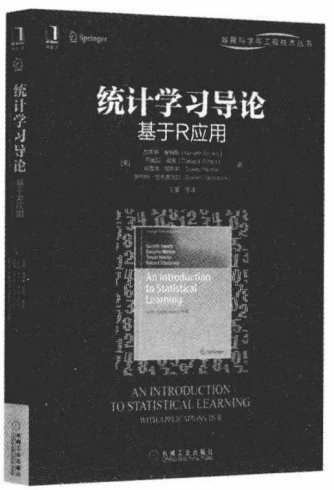
大数据分析：数据驱动的企业绩效优化、过程管理和运营决策

作者：Thomas H. Davenport ISBN：978-7-111-49184-2 定价：59.00元



商务智能：数据分析的管理视角（原书第3版）

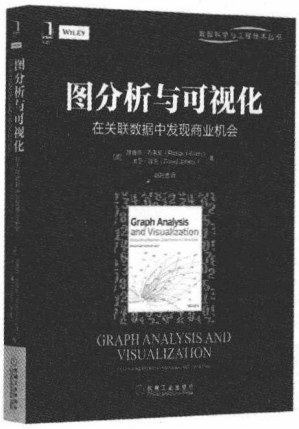
作者：Ramesh Sharda 等 ISBN：978-7-111-49439-3 定价：69.00元

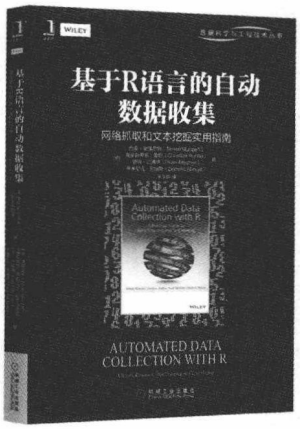


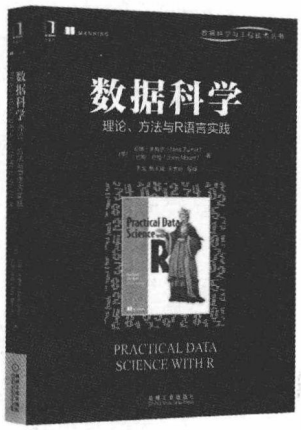
统计学习导论——基于R应用

作者：Gareth James 等 ISBN：978-7-111-49771-4 定价：79.00元

推荐阅读







图分析与可视化：在关联数据中发现商业机会

作者：理查德·布莱斯 ISBN：978-7-111-52692-6 定价：119.00元

本书将图与网络理论从实验室带到真实的世界中，深入探讨如何应用图和网络分析技术发现新业务和商业机会，并介绍了各种实用的方法和工具。作者Richard Brath和David Jonker运用高级专业知识，从真正的分析人员视角出发，通过体育、金融、营销、安全和社交媒体等领域的引人入胜的真实案例，全面讲解创建强大的可视化的过程。

基于R语言的自动数据收集：网络抓取和文本挖掘实用指南

作者：西蒙·蒙策尔特 等 ISBN：978-7-111-52750-3 定价：99.00元

本书由资深社会科学家撰写，从社会科学研究角度系统且深入阐释利用R语言进行自动化数据抓取和分析的工具、方法、原则和最佳实践。作者深入剖析自动化数据抓取和分析各个层面的问题，从网络和数据技术到网络抓取和文本挖掘的实用工具箱，重点阐释利用R语言进行自动化数据抓取和分析，能为社会科学研究者与开发人员设计、开发、维护和优化自动化数据抓取和分析提供有效指导。

数据科学：理论、方法与R语言实践

作者：尼娜·朱梅尔 等 ISBN：978-7-111-52926-2 定价：69.00元

本书讨论如何应用R程序设计语言和有用的统计技术处理日常的业务情况，并通过市场营销、商务智能和决策支持领域的示例，阐述了如何设计实验（比如A/B检验）、如何建立预测模型以及如何向不同层次的受众展示结果。

目　录