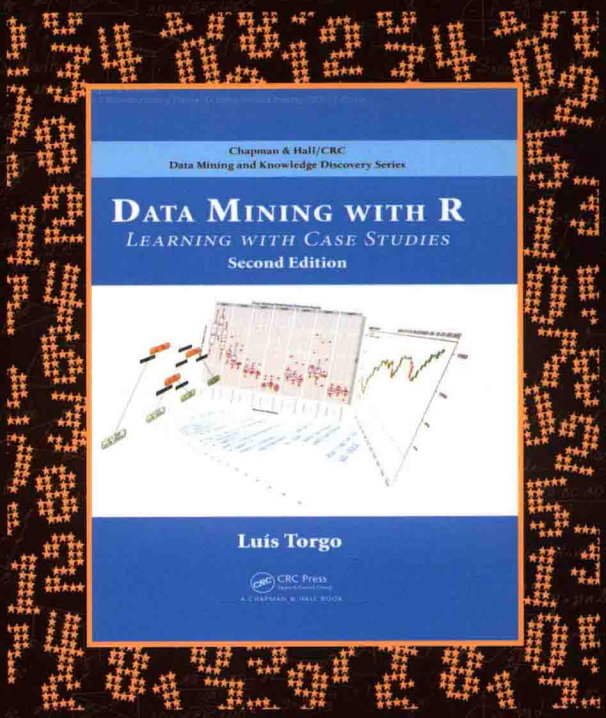
HZ BOOKS
华章IT
CRC Press
Taylor & Francis Group

数据科学与工程技术丛书

数据挖掘与R语言

（原书第2版）

[葡] 路易斯·托尔戈（Luís Torgo）著
李洪成 潘文捷 译



DATA MINING WITH R
LEARNING WITH CASE STUDIES
SECOND EDITION

机械工业出版社
China Machine Press

数据科学与工程技术丛书

DATA MINING WITH R
LEARNING WITH CASE STUDIES
SECOND EDITION

数据挖掘与R语言
（原书第2版）

[葡] 路易斯·托尔戈（Luís Torgo）著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘与R语言（原书第2版）/（葡）路易斯·托尔戈著；李洪成，潘文捷译．—北京：机械工业出版社，2018.3
（数据科学与工程技术丛书）
书名原文：Data Mining with R: Learning with Case Studies, Second Edition

ISBN 978-7-111-59666-0

I. 数… II. ①路… ②李… ③潘… III. ①数据采集 ②程序语言－程序设计 IV. ①TP274 ②TP312

中国版本图书馆CIP数据核字（2018）第065808号

本书版权登记号：图字 01-2017-7334

本书面向初学者，通过案例讲解如何用R进行数据挖掘。全书包括两部分，第一部分介绍R和数据挖掘的基础知识，第二部分为案例研究，通过预测海藻数量、预测股票市场收益、侦测欺诈交易以及微阵列样本分类四个案例培养读者构建解决方案的能力，掌握工具的使用技巧。

本书适合作为高校学生或业界新手了解R和数据挖掘的入门读本，其中的代码和数据均可免费下载。

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：张志铭　　责任校对：殷 虹
印　刷：北京文昌阁彩色印刷有限责任公司　　版　次：2018年5月第1版第1次印刷
开　本：185mm×260mm 1/16　　印　张：22.75（含0.25印张彩插）
书　号：ISBN 978-7-111-59666-0　　定　价：99.00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) 88378991 88361066　　投稿热线：(010) 88379604
购书热线：(010) 68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

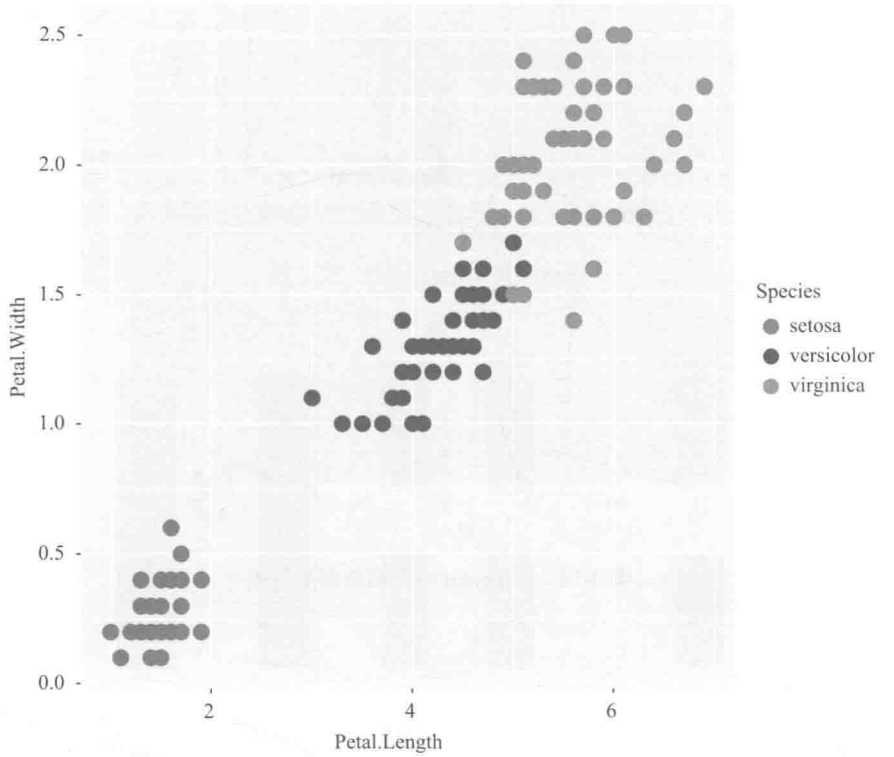


图 3-6　使用 ggplot 函数对 iris 数据集进行映射的示例

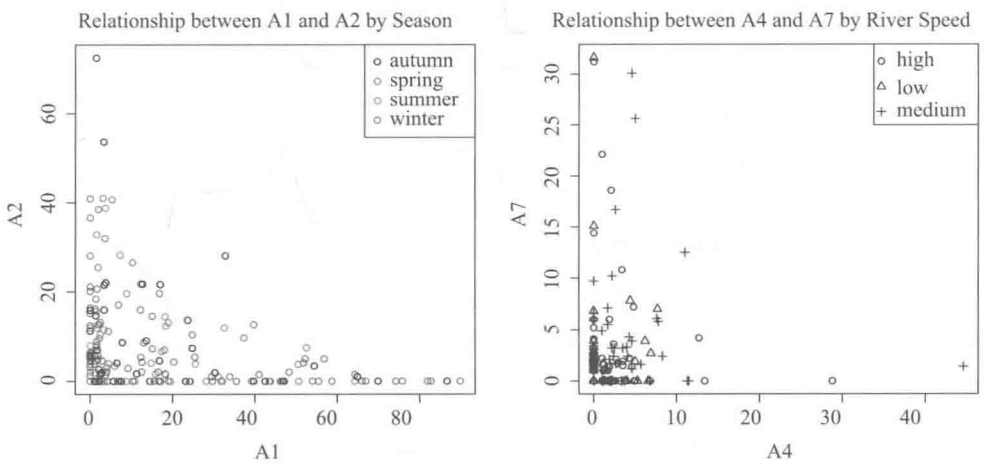


图 3-13　根据名义变量区分描点的两个散点图

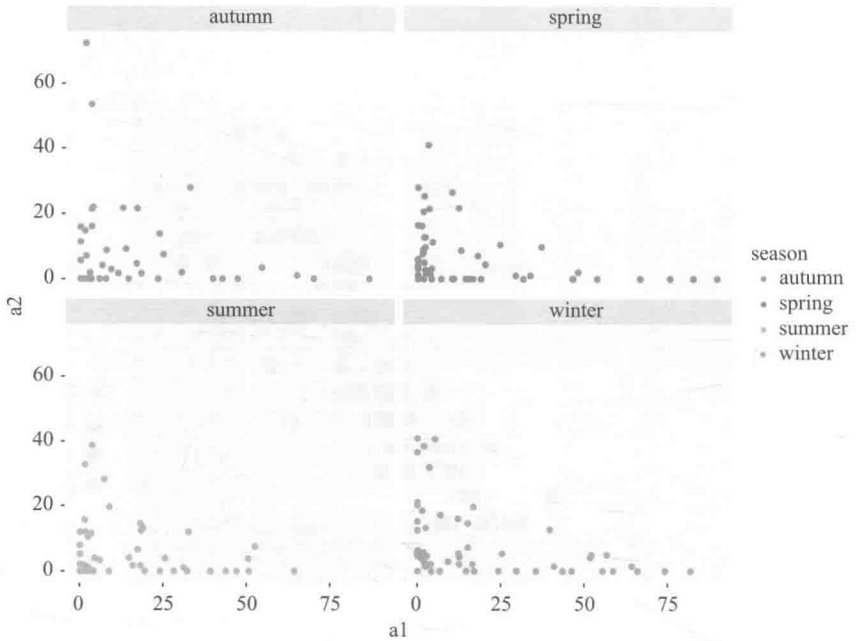


图 3-14　在 ggplot 中分面绘制散点图

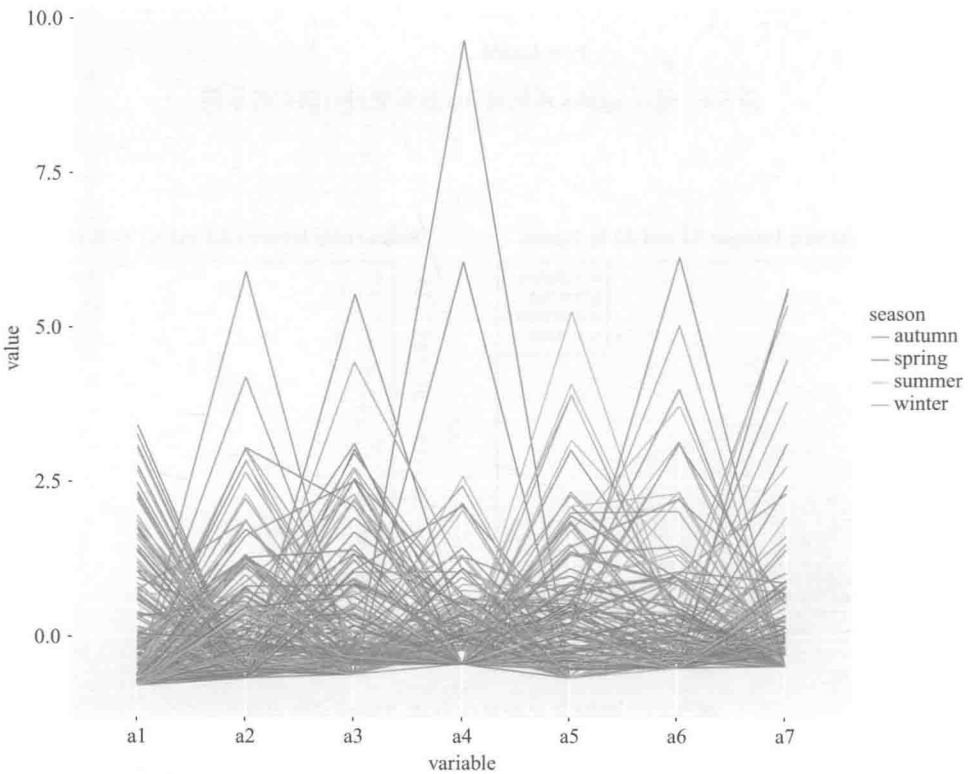


图 3-18　一个平行坐标图

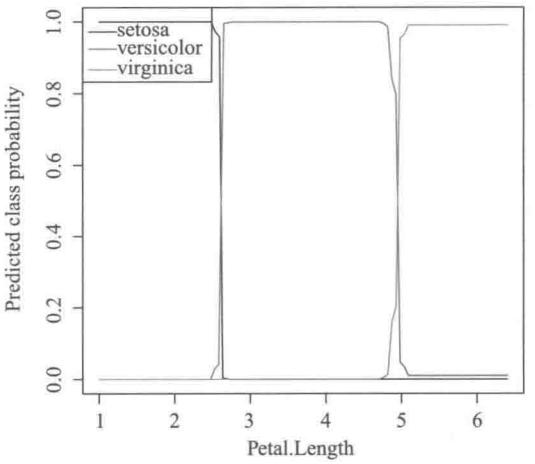


图 3-37　Petal.Length 的边际图

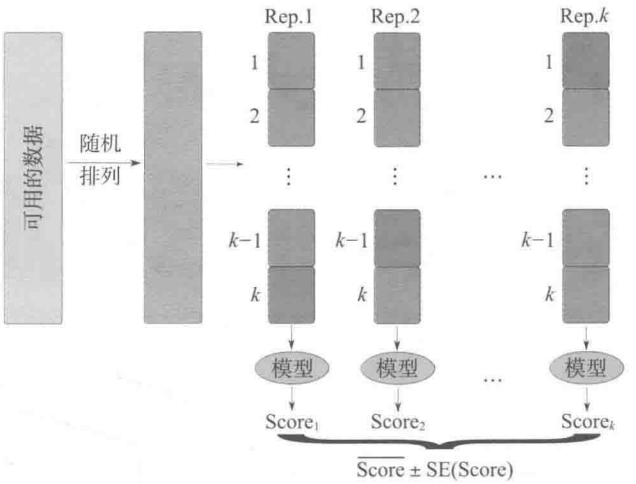


图 3-38　k 折交叉验证

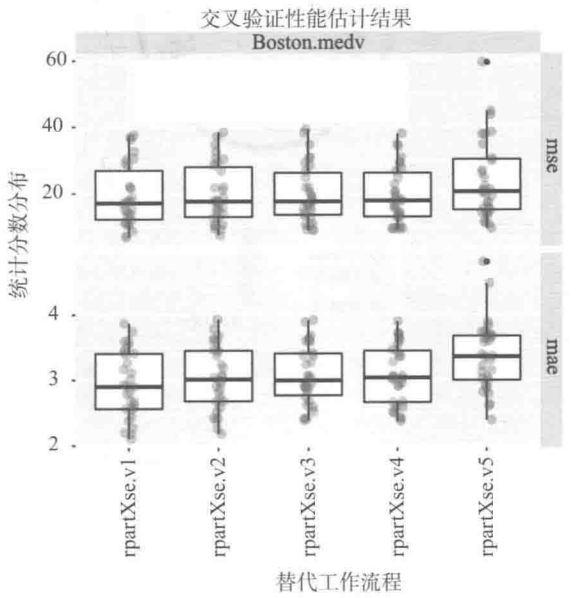


图 3-39　一个 10 折 CV 估计实验的结果

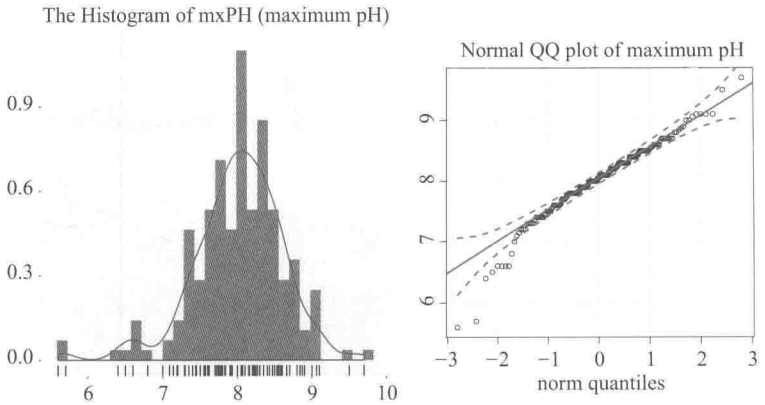


图 4-2　变量 mxPH 的直方图的“丰富”版本（左图）以及 Q-Q 图（右图）

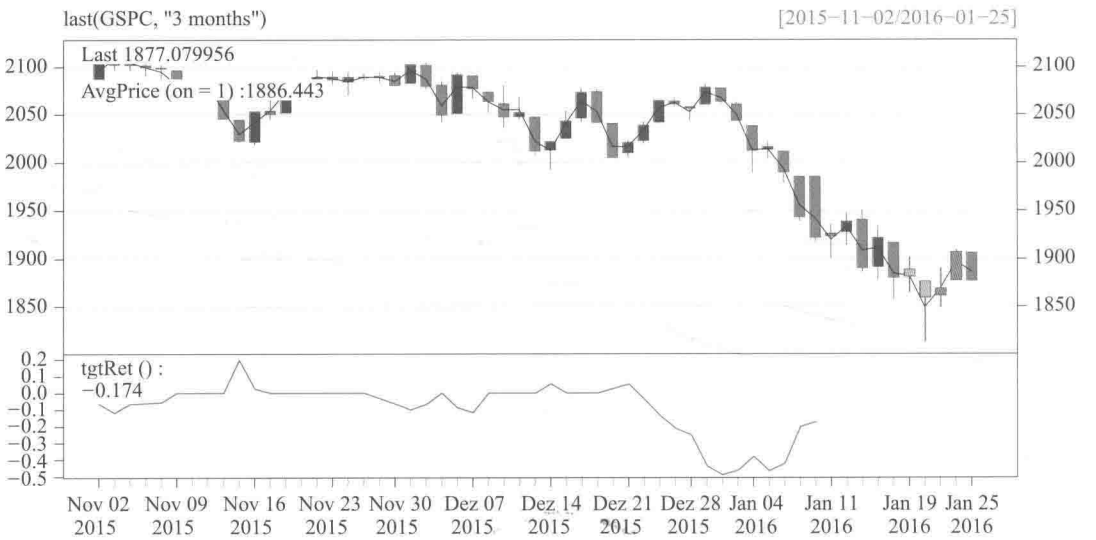


图 5-1　最后三个月的标准普尔 500 指数和我们的指标线图

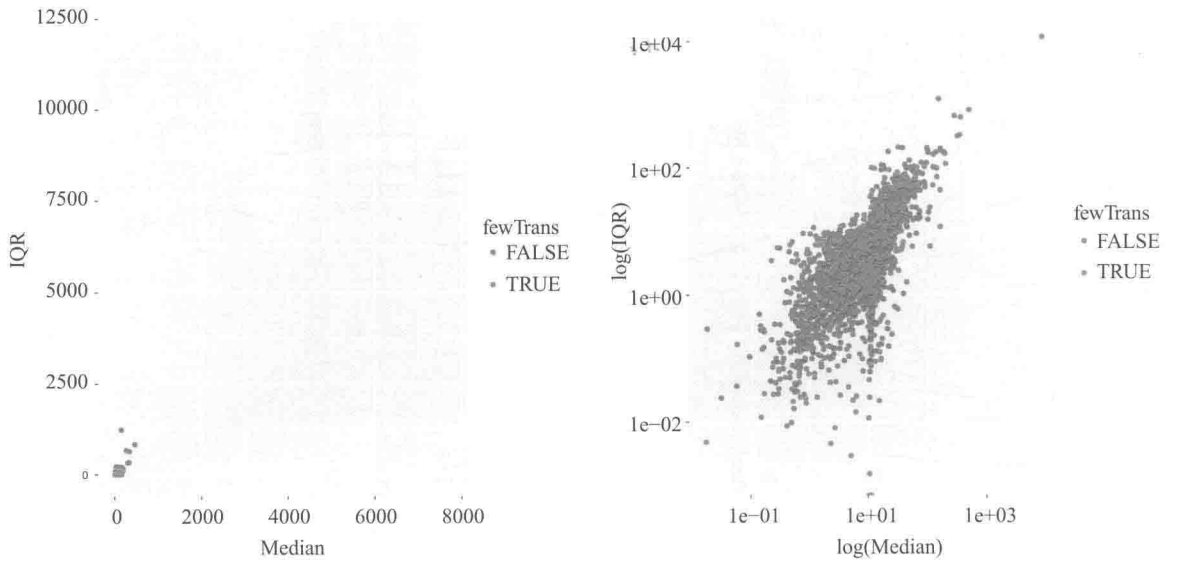


图 6-4　单位价格分布的某些特征

推 荐 序

Data mining has transformed the way that companies and other large organizations interact with their customers and manage complex processes. A profusion of data is now being put to good use to predict behavior and outcomes. On the software side, R has transformed the analytical landscape with its power and attractive pricing (free).

The goal of this book is to introduce you quickly to these two worlds. This introduction is done via practical case studies so you can place the learning in a realistic context without getting lost in a detailed discussion of statistical theory or the conceptual foundations of computer science. The free tools MySQL (for database manipulation) and R (for analysis) are used throughout. So this is very much a hands-on learning experience. You will gain the most if you install these tools and use them to work through the case studies in the book in detail.

The second edition retains the case-study approach of the first edition, but adds to it a 175-page survey of data mining methods and commentary on related tools in R. Part II then reinforces these concepts and tools in detail progressively throughout the case studies.

The author, Luís Torgo, has crafted this introduction based on a wealth of experience at the University of Porto, as well as teaching data mining courses in person and online.

数据挖掘改变了众多公司和大型机构与客户沟通的方式，同时也改变了他们管理复杂过程的方式。现在大量的数据被很好地用于预测行为模式和未知结果。从软件方面而言，R 以其强大的功能和诱人的价格（免费）改变了分析领域的蓝图。

本书的目的就是把读者快速引入这两个领域。结合书中的具体案例，读者可以在实际背景中进行学习，这样就不会在繁杂的统计理论或者计算机科学基础概念中不知所终。本书全部应用免费的工具——R（用于分析）和 MySQL（用于数据库操作），提供了丰富的动手学习的体验。若能安装这些工具并应用它们来详细分析书中的案例，你一定会收获满满。

第 2 版在保留第 1 版中案例学习方式的基础上，增加了 175 页对数据挖掘方法的回顾以及对相关 R 工具的讨论。第二部分通过案例学习逐步在细节上加深了对第一部分的概念和工

具的介绍。

本书作者 Luís Torgo 在波尔图大学从事教学工作多年，同时也经常在线下和线上教授数据挖掘课程，这些经验都融于第 2 版中，是一部精雕细琢的佳作。

Peter Bruce

美国统计教育学院

Statistics.com 在线课程网站总裁；《 Data Mining for Business Analytics : Concepts, Techniques and Applications in R》（ Wiley 2017）一书的作者（与 Shmueli、Yahav、Patel 和 Lichtendahl 合著）

中文版序

It is a big honor for me that the second edition of my book on Data Mining with R is also translated into the Chinese language, as it was the case of the first edition. The goal of this book is to bring R to the largest possible audience, trying to highlight the great capabilities of this environment for analyzing data. In this context, having the book translated to the most spoken language in the world is excellent. Hopefully, this new edition is at least as successful as the first.

My main goals in writing this new edition were: (i) upgrade the book with the large number of new packages and trends that have been appearing in this " fast paced world of R"; (ii) and address some of the questions raised by the readers. The first goal involved rewriting most of the code for solving the case studies, incorporating more recent packages that have become the standard in R (e.g. the excellent dplyr and ggplot2 packages). The second goal involved restructuring the book in two parts: (i) a first part containing an introduction to the two main topics of the book - R and data mining; and (ii) a second part with the case studies. The main novelty in this restructuring is the new chapter containing an introduction to the main data mining topics that appears in the first part. The goal of this chapter is to provide the reader with a broad, bird-eye view of the main concepts in the field of data mining and how the concepts can be tackled in R. I sincerely hope that this new structure of the book provides a better way of learning how to use R for data mining.

In summary, I hope this new edition of the book in the Chinese language becomes a very useful book for different types of readers from practitioners to students trying to delve into the world of data mining through this outstanding tool that is R.

和第 1 版一样，这本书的第 2 版再次被译成中文，这是我莫大的荣幸。本书中文版将把 R 呈现给更大的潜在读者群，更好地展现 R 语言环境下强大的数据分析功能。就此而言，把本书翻译为世界上使用人口最多的语言是一件很棒的事情。希望本书的新版本也能像第 1 版那样成功。

我写作新版本的主要目的如下：根据快节奏 R 语言世界中的大量新添加包和发展趋势来更新第 1 版；解决读者提出的一些问题。第一个目的需要应用最近出现并成为 R 标准的一些添加包，例如出色的 dplyr 和 ggplot2 添加包，重新编写案例研究中的大部分代码。第二个目的则需要重新规划本书的结构，现在分为两个部分：第一部分包含对书中两个主要主题——R 语言和数据挖掘的介绍；第二部分是案例研究。重新划分部分后，主要的新内容是在第一部分加入了介绍数据挖掘主题的一章。这一章的主要目的是为读者提供数据挖掘领域的主要概念和应用 R 来解决这些概念的宏观视角。我衷心希望本书的新结构能给读者提供学习用 R 进行数据挖掘的更好方式。

总之，不管是实际工作人员还是学生，所有希望通过 R 这个杰出工具研究数据挖掘的读者都可以阅读本书。我希望本书第 2 版能够成为对不同类型读者都十分有用的书籍。

Luís Torgo

波尔图，2018 年 1 月 12 日

译 者 序

随着大数据的概念变得越来越流行，对数据的探索、分析和预测成为大数据领域的基本技能之一。作为探索和分析数据的基本理论和工具，数据挖掘是近几年热门的技术之一。R作为功能强大并且免费的数据分析工具，在数据分析和挖掘领域获得了越来越多用户的青睐。本书介绍了R语言以及数据挖掘的基本知识，并应用R来进行实际数据案例的分析和挖掘，从数据中获取可以付诸行动的决策。

和第1版比较，本书增加了全新的一章对数据挖掘的基本知识进行介绍。全书分为两个部分，第一部分介绍R语言的基本知识以及数据挖掘的主要理论与方法，第二部分是案例研究。第二部分的每一章都详细介绍了一个案例，包括预测海藻数量、预测股票市场收益、侦测欺诈交易和微阵列样本分类。书中应用各种模型进行分析和挖掘，并对各个数据挖掘模型的性能进行分析和比较。

R是一款十分优秀的数据分析和挖掘软件，有大量的添加包（Package），现已成为主流的数据分析和挖掘软件之一。本书以实际的案例为主线，应用R语言进行系统的分析和预测，由浅入深，脉络清晰。读者不需要具有R语言和数据挖掘的预备知识就可以阅读本书。不管是R初学者还是熟练的R用户，都能从书中找到对自己有用的内容。本书案例分析所应用的方法和技能都是可以应用到实际数据挖掘实践中的，数据分析从业者将会发现本书是进行数据挖掘工作的有益参考。

我们有幸受机械工业出版社委托将此书译成中文，希望中文版的出版能够给国内读者学习R与机器学习带来方便。

本书第1版连续多年在亚马逊等网站的同类作品中成为畅销的书籍之一，希望第2版也能够和第1版一样受到读者的欢迎。

由于时间和水平所限，译文中难免会有不当之处，希望同行和读者多加指正。

李洪成

前　言

本书的主要目的是向读者介绍如何用 R 进行数据挖掘。R 是一种可以自由下载[⊖]的语言，它提供统计计算和绘图环境，这些功能和大量的添加包使其成为一款优秀的软件，取代了很多昂贵的数据挖掘工具。

本书的目的不是介绍数据挖掘的各个方面。许多已有的书籍已经覆盖了数据挖掘领域，而本书是用几个案例来向读者介绍 R 的数据挖掘能力。显然，这几个案例不能代表我们在现实世界中碰到的所有数据挖掘问题。同时，我们给出的解决方案也不是最完整的方案。本书通过这些实际案例向读者介绍如何用 R 进行数据挖掘，因此案例分析目的是展示用 R 进行信息提取的例子，而不是提供数据挖掘案例的完整分析报告。它们可以作为任何数据挖掘项目的可能思路，或者作为开发数据挖掘项目解决方案的基础。尽管如此，我们尽力尝试覆盖多方面的问题，以展示由数据大小、数据类型、分析目标和分析工具所带来的不同挑战。然而，这里的实践方式也是有代价的。实际上，作为具体案例研究的一种形式，为了让读者在自己的计算机上执行我们所描述的步骤，我们也做了某些妥协。也就是说，我们不能处理太大的问题，这些问题要求的计算机资源不是每个人都具备的。尽管这样，我们认为本书涵盖的问题也不算小，并且我们还对由不同数据类型和维度带来的问题给出了解决方案。

第 2 版大幅修改了案例研究的 R 代码，使其与 R 中出现的最新添加包同步更新。此外，我们决定将本书分为两部分：第一部分为材料介绍；第二部分为案例研究。第一部分用一个全新的章节来介绍数据挖掘，以补充已有的对 R 的介绍。这个想法是为读者提供数据挖掘领域的一种鸟瞰图，更深入地描述这个研究领域的主题。这些信息补充了案例分析中给出的简单描述。此外，它允许读者更好地将数据挖掘任务及方法论的更大图景与案例研究的解决方案区分开来。最后，如果需要更多关于案例研究中使用方法的细节，我们希望这个新章节可以作为读者的参考。

本书并不要求读者具有 R 的先验知识，没有学过 R 和数据挖掘的读者也可以学习书中的案例。书中的各个案例相互独立，读者可以从书中任何一个案例开始。当然，在第一个简单案例中，给出了一些基本的 R 知识，这意味着，如果你没有学过 R，至少应该从第一个案

⊖ 下载网址：http://www.R-project.org。

例开始学习。而且，第 1 章给出了 R 的简介，它可以帮助你理解后面的章节。我们没有假设你熟悉数据挖掘和统计技术，在每个案例中必要的地方，都对不同的数据挖掘技术进行了介绍。不过，第一部分的新章节介绍了数据挖掘，包括我们在案例研究中应用的方法以及数据挖掘中常用的其他方法的进一步信息。另外，在某些节的末尾，我们提供了“进一步阅读”资料，如果需要，可以参考它们。总之，本书的读者应该是数据分析工具的用户，而不是研究人员或者开发人员。同时，我们希望后者将阅读本书作为进入 R 和数据挖掘世界的一种方式，从而发现本书的用途。

本书配有一个免费的 R 代码集，可以从本书网站下载[⊖]。其中含有案例研究中的所有代码，这可以帮助你进行实践学习。我们强烈建议读者在阅读本书时安装 R 并试验书中的代码。而且，我们创建了一个名为 DMwR2 的 R 添加包，它包含本书用到的多个函数和以 R 格式保存的案例数据集。建议你按照本书的指示安装并加载该添加包（第 1 章给出了细节）。

⊖ 下载网址：http://ltorgo.github.io/DMwR2。

致　谢

首先要感谢我的家人，没有他们的帮助和支持，我是无法完成本书的。他们的支持、关怀和爱给我足够的安慰，使我可以克服在写作本书过程中遇到的困难。同样，也要感谢我的朋友，他们总是在我需要安慰的时候和我一起畅饮、交流，带给我轻松愉悦的写作心情。谢谢我的家人和朋友！谢谢你们！现在，我希望有更多的时间陪在你们身边。

我也要感谢我的所有同事和 LIAAD/INESC Tec LA 实验室对我的支持。同时，也要感谢波尔图大学对我的研究的支持，感谢科学院计算机科学系的同事为我提供的愉快的工作环境。写作本书的部分资助来自于葡萄牙自然科学基金（资助号：SFRH/BSAB/113896/2015）。

最后，感谢所有针对反馈意见改进第 1 版以及校对当前版本草稿的学生和同事们。特别要感谢在波尔图大学科学院攻读计算机科学硕士学位的数据挖掘专业的学生们，以及在纽约大学斯特恩商学院攻读商业分析科学硕士学位的“数据挖掘与 R 语言”课程的学生们——他们对我的教学材料的参与和反馈在本书的新版本中有很好的体现。

Luís Torgo

葡萄牙，波尔图

目　录