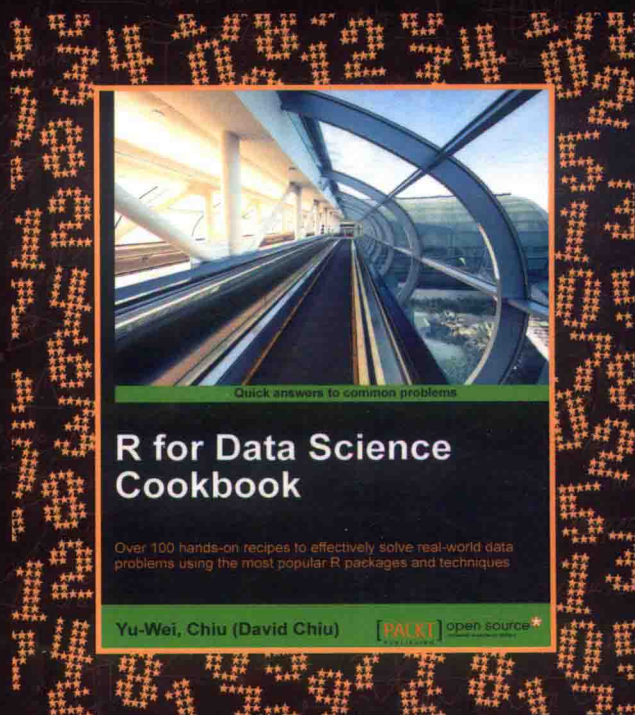


# 数据科学

## R语言实现

丘祐玮 (David Chiu) 著

魏博 译 刘冰 审校



# R FOR DATA SCIENCE COOKBOOK



机械工业出版社  
China Machine Press

数据科学与工程技术丛书

R FOR DATA SCIENCE COOKBOOK

# 数据科学

## R语言实现

丘祐玮 (David Chiu) 著

魏博 译 刘冰 审校



机械工业出版社  
China Machine Press

## 图书在版编目 (CIP) 数据

数据科学: R 语言实现 / 丘祐玮著; 魏博译. —北京: 机械工业出版社, 2017.5

(数据科学与工程技术丛书)

书名原文: R for Data Science Cookbook

ISBN 978-7-111-56831-5

I. 数… II. ①丘… ②魏… III. 程序语言-程序设计 IV. TP312

中国版本图书馆 CIP 数据核字 (2017) 第 092039 号

本书版权登记号: 图字: 01-2017-0155

Yu-Wei, Chiu(David Chiu): *R for Data Science Cookbook* (ISBN: 9781784390815).

Copyright © 2016 Packt Publishing. First published in the English language under the title “R for Data Science Cookbook”.

All rights reserved.

Chinese simplified language edition published by China Machine Press.

Copyright © 2017 by China Machine Press.

本书中文简体字版由 Packt Publishing 授权机械工业出版社独家出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

本书介绍了 R 语言在数据科学领域应用的方方面面, 包括数据处理、数据操作、数据可视化、概率模拟、序列预测、频繁项集挖掘、监督式算法和非监督式算法等。本书注重对 R 语言程序包的介绍和使用, 可操作性强, 并且提供了关于高级技术的扩展阅读资料, 供学有余力的读者钻研探讨。本书适合数据科学相关专业的本科生和研究生使用, 也可作为数据领域从业人员的参考书。

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: 和 静

责任校对: 殷 虹

印 刷: 北京诚信伟业印刷有限公司

版 次: 2017 年 6 月第 1 版第 1 次印刷

开 本: 185mm × 260mm 1/16

印 张: 20.25

书 号: ISBN 978-7-111-56831-5

定 价: 79.00 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

读者热线: (010) 68326294 88379649 68995259

读者信箱: hzjsj@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

## 推 荐 序

对于互联网来说，当今是一个数据为王的时代。特别是对于中国互联网从业者来说，这个时代来得很快、很迅猛。Web 2.0 的到来让用户数据呈现出爆炸式的增长态势。如今，在搜索引擎、推荐系统、广告投放、语音识别、图像识别、自然语言理解、对话系统、交通物流、互联网金融等领域都可以看到数据科学发挥功效的身影。而 AlphaGo 的胜利又掀起了人类对机器智能的大讨论。数据科学让人类在这个时代重新看到了 0 和 1 所蕴藏的巨大价值，找到了新的技术和战略高地。

然而机遇总是与挑战共存。海量数据通常意味着人们要花费巨大的精力才能把数据转变为价值，无论是数据记录、数据处理、数据算法，还是数据可视化，每一个环节都会影响最后的产出，而每一个环节都不能轻而易举地完成。市面上关于数据科学的书籍林林总总，最多的要数 Python 派和 R 派。Python 派起源于工程人员，不管是理论还是实践，读者可选的种类还是很多的。而 R 语言是由统计学者开发的，其出身和基因也决定了它在工程应用上的暂时落后，相关书籍也较少。对于我国数据科学从业者来说，R 也是一门相对陌生的语言。听说魏博在翻译这本有关 R 语言开发应用的书籍，我非常欣慰。人们在 R 语言的实战书籍中又多了一种选择。同时，我也希望本书可以带动大家使用 R 进行实际开发的热情。学科和行业的发展需要百家争鸣。数据科学的利器越多，挑战和壁垒就会越容易攻克，人们获得数据价值的机会也就越大。

本书介绍了 R 语言在数据科学领域应用的方方面面，包括数据处理、数据操作、数据可视化、概率模拟、序列预测、频繁项集挖掘、监督式算法和非监督式算法等。本书注重 R 语言程序包的介绍和使用，这可以让读者省去很多无谓的精力，避免“重复造轮子”。特别是对于“唯快不破”的互联网行业来说，能够尽快使用成熟的开发包，是一件很幸福的事情。对于一名希望提升自己数据科学战斗力的读者来说，这本书也是不二选择。

魏博毕业于中国科学院数学所，读博期间就开始主攻用户需求建模和自动推理，后来负责阿里巴巴优酷大文娱视频搜索引擎的算法优化工作，目前在欧普拉软件新闻推荐部门负责用户画像算法。他在理解用户行为和用户数据挖掘方面有较深的造诣。同时他也是一名资深的 R 语言用户和 R 语言传播者。这本书的翻译也是对他自身能力的考验和体现。英文原版书有 400 多页，他能够在工作之余，牺牲节假日，4 个月来静下心来完成翻译实属不易。同时，R 语言在国内的发展还有很长的路要走，特别是一些面向开发的技术术语的

中文译法还没有达成广泛共识，在翻译过程中经常需要揣度和推敲。希望本书能够成为广大 R 语言数据科学爱好者共同阅读、共同推广的资源。

再次感谢魏博能够把机械工业出版社的优质外文资源翻译成书。也希望这本书能够成为大家手边时常翻阅的经典数据科学资料。

胡睿

前微软资深研发总监

现欧普拉软件技术（北京）有限公司副总裁

## 译者序

“数据科学”也许是近几年最热门的科技名词。从各种各样的互联网应用到 AlphaGo 的胜利，世界正在见证着数据科学带来的深刻影响。特别是我国在“互联网+”大潮推动下，海量数据每时每刻都在产生。无论是学术界，还是工业界，大量的科研人员都在这块数据金矿上挥洒青春和汗水。从数据中发现规律并应用到产品开发或者战略决策中，逐渐成为各行各业及各个企业公司的标准日常活动。同时，数据科学对于统计学、数据挖掘/机器学习算法，以及数据可视化等的需求也变得日渐迫切。一名掌握数据存储知识、数据科学算法、数据项目管理和数据可视化技术的人员，通常是多家用人单位的必争之才。

古语说：工欲善其事，必先利其器。处理海量、纷繁、动态的数据不是一件容易的事情。2013年，我在一篇科研论文上第一次结识了 R 语言。从那以后，我便开始了 R 语言道路上的各种探索。从最开始的编写算法脚本，到后来执著于 R markdown 的使用，再到后来开发 Shiny 小程序，以及在服务器上执行计算。对我来说，从来没有一种语言让我觉得在数据科学的道路上如此顺畅和友好。然而，遍寻各种书籍和网上教程，很少有一份可靠的实践开发资料可以让读者体会到实际项目的“火药味”。很多书籍都在着重介绍 R 语言的基础知识，对于实际开发和各种强大的程序包都鲜有系统的介绍。这与 R 语言在我国的普及和发展程度不无关系。我想是时候推出一本书，让大家见识 R 语言和程序包的超强能力了。

本书便是一本注重 R 语言程序包实际应用的书籍。本书结构严谨，阐述通俗易懂，内容涵盖了 R 语言的绝大多数应用场景。同时，对于众人皆知的基础知识，本书不再赘述；对于一些高级技术，本书还提供了一些扩展阅读的资料，供学有余力的读者钻研探讨。值得提出的是，本书的作者丘祐玮是中国台湾地区活跃的数据科学家。他不仅拥有自己的数据科学公司，同时还服务于各大数据科学论坛，与中国大陆的数据科学工作者也有合作，是一位理论与实践并重的数据科学家。相信每一位读者都可以通过这本书学习到他的宝贵经验。

2016年9月，我受机械工业出版社的委托，承担本书的翻译工作。整个翻译过程历时4个月。这个过程也是自我温习、自我提高的过程。特别是书中关于数据处理程序包的介绍，令我耳目一新。同时我的翻译工作也得到了阿里巴巴优酷大文娱同事和欧普拉同事

的鼓励和支持。感谢刘冰老师负责全书的审校。还特别要感谢我的爱人王丽萍几个月来对我的支持，希望这本书也是一份家庭礼物。

尽绵薄之力，终付梓成册。译者水平有限，疏漏在所难免，敬请读者批评指正。

魏博

欧普拉软件技术（北京）有限公司

# 前 言

大数据、物联网、人工智能已经变成近几年最热门的科技流行语。尽管大家用很多名词去定义这些技术，但是共通的思想是它们都是数据驱动的。人们并不满足于简单地拥有数据，因为发现其中的价值才是最本质的。因此数据科学家已经开始关注如何从原始数据中洞悉深层价值。

数据科学已经变成学术界和产业界最流行的话题。但是数据科学是一门非常宽泛的学科，学会掌握数据科学注定很有挑战性。初学者必须学习如何准备、处理、聚合和可视化数据。而更多高级技能包括机器学习，挖掘各种数据格式（文本、图像和视频），以及最重要的——使用数据产生商业价值。数据科学家的角色需要大量的努力，同时，一名成功的数据科学家也需要一个有力的工具来解决日常问题。

在这个领域中，数据科学家使用最广泛的工具是开源而且免费的 R 语言。作为一种机器语言，R 提供了许多数据处理函数、学习库和可视化函数，允许用户快速上手分析数据。R 可以帮助用户快速执行分析，并在不需要懂得复杂数学模型细节的前提下执行机器学习算法。

本书给出了实际方案，教你如何使用 R 语言将数据科学落地。全书共 12 章，每一章都分成几个简单的教程。通过每一个教程循序渐进的介绍，你可以使用 R 的程序包，掌握书中所教授的技术。

本书首先介绍如何创建 R 函数来避免不必要的代码重复。你会学到如何使用 R 程序包，在各种数据源上准备数据、处理数据和执行高级 ETL 操作。数据操作的一个例子是介绍如何使用 `dplyr` 和 `data.table` 程序包有效地处理大型数据结构。还有一章关注 `ggplot2`，介绍如何创建高级图形，进行数据展示。你也会学到如何使用 `ggvis` 程序包构建交互式报告。

本书也会介绍如何使用数据挖掘技术发现经常一起购买的产品。后面的章节还给出了财务数据的时间序列分析结果。还有一些章节会深入介绍机器学习技术，包括数据分类、回归、聚类和降维。我可以保证，本书会让你觉得，数据科学学习原来如此简单。

## 主要内容

第 1 章介绍如何创建 R 函数。该章会介绍 R 函数的基本构成、环境和参数匹配。我



们还会介绍高级技术，例如闭包、函数式编程和如何处理错误。

第 2 章介绍如何使用 R 读取结构化和非结构化的数据。该章首先介绍从文本文件中读取数据。然后，介绍如何把 R 和数据库连接起来。最后，你会学到如何编写网络爬虫，爬取网页和社交网络上的非结构化数据。

第 3 章介绍分析前的数据准备工作。在该章中，我们会介绍数据预处理过程，使用基本的 R 函数，进行例如类型转换、添加、过滤、舍弃、重塑和缺失值估计。

第 4 章介绍如何使用高级程序包 `data.table` 和 `dplyr` 有效而且高效地操作数据。`data.table` 提供了快速加载和聚合大型数据的可能。`dplyr` 程序包提供了以类似 SQL 的语法操作数据的能力。

第 5 章介绍使用 `ggplot2` 可视化数据。首先介绍 `ggplot2` 的基本构成。然后，介绍高级技术，使用 `ggplot2` 函数创建复杂的图形。最后，介绍如何使用 `ggmap` 构建地图。

第 6 章展示如何使用 R 创建一份专业的报告。首先，讨论如何使用 R `markdown` 语法，嵌入 R 代码块。然后，介绍如何使用 `ggvis` 添加交互式图表。最后，介绍如何创建和发布 R Shiny 报告。

第 7 章关注如何从不同的概率分布上抽样数据。作为一个具体的例子，我们会介绍如何使用概率函数模拟随机交易过程。

第 8 章首先讨论点估计和置信区间。然后，介绍参数和非参数检验方法。最后，介绍如何使用 ANOVA 分析工程师的收入是否会随着头衔和地区的变化而不同。

第 9 章介绍用于发现交易数据中关联项和暗藏的频率模式的常用方法。在该章中，我们会使用一个实际例子，以便你可以学到如何在实际的数据集中执行规则和模式挖掘。

第 10 章首先介绍如何从财务数据集中创建和操作时间序列。然后介绍如何使用 `HoltWinters` 和 `ARIMA` 预测时间序列。该章会通过一个实际例子介绍如何使用 `ARIMA` 预测股票价格。

第 11 章介绍如何构建基于标注训练数据的预测模型。你会学到如何使用回归模型理解数值关系，并使用拟合模型进行连续值预测。对于分类任务，你会学到如何拟合数据，生成一个树形分类器。

第 12 章介绍未标注数据的隐含结构。首先，介绍如何使用聚类方法对位置临近的旅馆进行分组。然后，介绍如何使用 `PCA` 方法选取和抽取经济自由度数据集中的特征。

## 机器环境

要学习本书中的例子，你需要一台可以访问互联网的计算机，而且可以安装 R 环境。你可以通过 <http://www.cran.rproject.org/> 下载 R 安装文件。具体安装信息可以在第 1 章中找到。

本书中的例子是在 Microsoft Windows 和 R 3.2.4 的基础上编写和测试的。这些例子

也可以在 Mac OS X 或者类似于 UNIX 的操作系统下，使用最新的 R 版本编译通过。

## 读者人群

本书是面向已经熟悉 R 语言的基础操作，但是希望学习如何使用 R 程序包有效而且高效地分析现实世界数据问题的读者。

## 行文结构

在本书中，你会发现一些标题经常出现（如准备工作、实现步骤、运行原理、更多技能和扩展阅读）。

为了清楚地介绍如何完成每一个教程的学习，我们使用以下行文结构：

### 准备工作

这个部分会告诉你当前教程需要的东西，并介绍如何安装软件和本教程所需的基础环境。

### 实现步骤

这个部分包括本教程所需的步骤。

### 运行原理

这个部分通常包含对前一个部分的具体解释。

### 更多技能

这个部分包含本教程的额外信息，以便扩展读者关于当前教程的知识面。

### 扩展阅读

这个部分提供了指向其他有用信息的链接。

## 本书约定

在本书中，你会发现文本有多种风格，以提供不同的信息。这里给出一些风格的例子，并给出其中的含义。

程序包和函数名风格如下：“你可以安装加载程序包 `RCurl`”。

代码块设置如下：

```
> install.packages("RCurl")
```

```
> library(RCurl)
```

屏幕中的词语，例如菜单或者对话框，文本做如下展示：“在 R 中，缺失值使用 NA（不适用）标记，不可能的值用 NaN（不是一个值）标记”。

## 样例源码下载

你可以从 <http://www.packtpub.com> 通过个人账号下载你所购买书籍的样例源码。如果你是从其他途径购买的，可以访问 <http://www.packtpub.com/support>，完成账号注册，就可以直接通过邮件方式获得相关文件。

你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，下载本书的源代码。

## 下载书中彩图

我们还为读者准备了一个 PDF 文件，该文件包含了本书所有截图和样图，可以更好地帮助读者理解输出的变化。你可以从以下地址下载：[http://www.packtpub.com/sites/default/files/downloads/RforDataScienceCookbook\\_ColorImages.pdf](http://www.packtpub.com/sites/default/files/downloads/RforDataScienceCookbook_ColorImages.pdf)。

# 目 录

|   |  |
|---|--|
| 推荐序                                     |  |
| 译者序                                     |  |
| 前言                                      |  |
| <b>第 1 章 R 中的函数</b> ..... 1             |  |
| 1.1 引言..... 1                           |  |
| 1.2 创建 R 函数..... 2                      |  |
| 1.3 匹配参数..... 3                         |  |
| 1.4 理解环境..... 5                         |  |
| 1.5 使用词法域..... 8                        |  |
| 1.6 理解闭包..... 10                        |  |
| 1.7 执行延迟计算..... 12                      |  |
| 1.8 创建中缀操作符..... 13                     |  |
| 1.9 使用替代函数..... 15                      |  |
| 1.10 处理函数中的错误..... 17                   |  |
| 1.11 调试函数..... 21                       |  |
| <b>第 2 章 数据抽取、转换和加载</b> ..... 28        |  |
| 2.1 引言..... 28                          |  |
| 2.2 下载公开数据..... 28                      |  |
| 2.3 读取和写入 CSV 文件..... 31                |  |
| 2.4 扫描文本文件..... 32                      |  |
| 2.5 使用 Excel 文件..... 34                 |  |
| 2.6 从数据库中读取数据..... 36                   |  |
| 2.7 爬取网络数据..... 38                      |  |
| 2.8 获取 Facebook 数据..... 44              |  |
| 2.9 使用 twitterR..... 49                 |  |
| <b>第 3 章 数据预处理和准备</b> ..... 53          |  |
| 3.1 引言..... 53                          |  |
| 3.2 重命名数据变量..... 53                     |  |
| 3.3 转换数据类型..... 55                      |  |
| 3.4 使用日期格式..... 57                      |  |
| 3.5 添加新的记录..... 58                      |  |
| 3.6 过滤数据..... 60                        |  |
| 3.7 舍弃数据..... 63                        |  |
| 3.8 合并数据..... 64                        |  |
| 3.9 排列数据..... 65                        |  |
| 3.10 重塑数据..... 67                       |  |
| 3.11 检测缺失数据..... 69                     |  |
| 3.12 估计缺失数据..... 71                     |  |
| <b>第 4 章 数据操作</b> ..... 74              |  |
| 4.1 引言..... 74                          |  |
| 4.2 使用 data.table 加强 data.frame..... 74 |  |
| 4.3 使用 data.table 管理数据..... 77          |  |
| 4.4 使用 data.table 执行快速聚合..... 82        |  |
| 4.5 使用 data.table 合并大型数据集..... 85       |  |
| 4.6 使用 dplyr 进行数据抽取和切片..... 88          |  |
| 4.7 使用 dplyr 进行数据抽样..... 91             |  |
| 4.8 使用 dplyr 选取列..... 92                |  |
| 4.9 使用 dplyr 进行链式操作..... 94             |  |

|  |     |                              |     |
|--|-----|------------------------------|-----|
| 4.10 使用 <code>dplyr</code> 整理行             | 95  | 7.3 理解均匀分布                   | 171 |
| 4.11 使用 <code>dplyr</code> 消除重复行           | 97  | 7.4 生成二项随机变量                 | 173 |
| 4.12 使用 <code>dplyr</code> 添加新列            | 98  | 7.5 生成泊松随机变量                 | 175 |
| 4.13 使用 <code>dplyr</code> 汇总数据            | 99  | 7.6 从正态分布中抽样                 | 177 |
| 4.14 使用 <code>dplyr</code> 合并数据            | 102 | 7.7 从卡方分布中抽样                 | 183 |
| <b>第 5 章 使用 <code>ggplot2</code> 可视化数据</b> | 105 | 7.8 理解学生 $t$ -分布             | 185 |
| 5.1 引言                                     | 105 | 7.9 从数据集中抽样                  | 187 |
| 5.2 使用 <code>ggplot2</code> 创建基础图形         | 106 | 7.10 模拟随机过程                  | 188 |
| 5.3 改变美学映射                                 | 109 | <b>第 8 章 R 中的统计推断</b>        | 191 |
| 5.4 引入几何对象                                 | 112 | 8.1 引言                       | 191 |
| 5.5 执行变换                                   | 116 | 8.2 获取置信区间                   | 191 |
| 5.6 调整图形尺度                                 | 118 | 8.3 执行 $Z$ -检验               | 196 |
| 5.7 分面                                     | 120 | 8.4 执行学生 $T$ -检验             | 199 |
| 5.8 调整主题                                   | 122 | 8.5 执行精确二项检验                 | 202 |
| 5.9 组合图形                                   | 124 | 8.6 执行 Kolmogorov-Smirnov 检验 | 203 |
| 5.10 创建地图                                  | 126 | 8.7 使用 Pearson 卡方检验          | 205 |
| <b>第 6 章 制作交互式报告</b>                       | 131 | 8.8 理解 Wilcoxon 秩和检验         | 207 |
| 6.1 引言                                     | 131 | 8.9 执行单因素方差分析                | 209 |
| 6.2 创建 R Markdown 报告                       | 131 | 8.10 执行双因素方差分析               | 212 |
| 6.3 学习 markdown 语法                         | 135 | <b>第 9 章 R 语言规则和模式挖掘</b>     | 216 |
| 6.4 嵌入 R 代码块                               | 137 | 9.1 引言                       | 216 |
| 6.5 使用 <code>ggvis</code> 创建交互式图形          | 140 | 9.2 把数据转换为事务                 | 216 |
| 6.6 理解基础语法                                 | 143 | 9.3 展示事务和关联                  | 218 |
| 6.7 控制坐标轴和图例                               | 148 | 9.4 使用 Apriori 规则挖掘关联关系      | 220 |
| 6.8 使用尺度                                   | 153 | 9.5 对冗余规则剪枝                  | 223 |
| 6.9 给 <code>ggvis</code> 图形添加交互            | 154 | 9.6 可视化关联规则                  | 224 |
| 6.10 创建 R Shiny 文档                         | 159 | 9.7 使用 Eclat 挖掘频繁项集          | 226 |
| 6.11 发布 R Shiny 报告                         | 164 | 9.8 使用时序信息创建事务               | 228 |
| <b>第 7 章 概率分布模拟</b>                        | 169 | 9.9 使用 cSPADE 挖掘频繁序列模式       | 231 |
| 7.1 引言                                     | 169 | <b>第 10 章 R 语言时间序列挖掘</b>     | 235 |
| 7.2 生成随机样本                                 | 169 | 10.1 引言                      | 235 |

|                       |                      |     |                        |                             |     |
|-----------------------|----------------------|-----|------------------------|-----------------------------|-----|
| 10.2                  | 创建时间序列数据             | 235 | 11.11                  | 可视化递归分割树                    | 282 |
| 10.3                  | 绘制时间序列对象             | 238 | 11.12                  | 使用混淆矩阵度量模型性能                | 283 |
| 10.4                  | 分解时间序列               | 241 | 11.13                  | 使用 ROC 度量预测性能               | 285 |
| 10.5                  | 平滑时间序列               | 243 |                        |                             |     |
| 10.6                  | 预测时间序列               | 247 | <b>第 12 章 非监督式机器学习</b> |                             | 288 |
| 10.7                  | 选取 ARIMA 模型          | 251 | 12.1                   | 引言                          | 288 |
| 10.8                  | 创建 ARIMA 模型          | 255 | 12.2                   | 使用层次聚类法对数据聚类                | 288 |
| 10.9                  | 使用 ARIMA 模型预测        | 257 | 12.3                   | 切割树成聚类                      | 291 |
| 10.10                 | 使用 ARIMA 模型预测股票价格    | 260 | 12.4                   | 使用 k-means 方法对数据聚类          | 293 |
|                       |                      |     | 12.5                   | 使用基于密度的方法对数据聚类              | 294 |
| <b>第 11 章 监督式机器学习</b> |                      | 264 | 12.6                   | 从聚类中抽取轮廓信息                  | 296 |
| 11.1                  | 引言                   | 264 | 12.7                   | 比较多种聚类方法                    | 298 |
| 11.2                  | 使用 $l_m$ 拟合线性回归模型    | 264 | 12.8                   | 使用基于密度的聚类识别数字               | 299 |
| 11.3                  | 汇总线性模型拟合             | 266 | 12.9                   | 使用 k-means 聚类方法分组相似<br>文本文档 | 301 |
| 11.4                  | 使用线性回归来预测未知值         | 268 | 12.10                  | 使用主成分分析法进行数据<br>降维          | 303 |
| 11.5                  | 度量回归模型的性能            | 270 | 12.11                  | 使用陡坡图确定主成分数量                | 305 |
| 11.6                  | 执行多元回归分析             | 272 | 12.12                  | 使用 Kaiser 方法确定主成分<br>数量     | 306 |
| 11.7                  | 使用逐步回归选取最优拟合回归<br>模型 | 274 | 12.13                  | 使用双标图可视化多变元数据               | 308 |
| 11.8                  | 应用高斯模型泛化线性回归         | 276 |                        |                             |     |
| 11.9                  | 执行逻辑斯谛回归分析           | 277 |                        |                             |     |
| 11.10                 | 使用递归分割树构建分类模型        | 280 |                        |                             |     |

# 第 1 章

## R 中的函数

### 1.1 引言

R 语言是数据科学家的主流编程语言。基于著名的数据分析网站 KDnuggets 的民意测验，3 项（2012 年到 2014 年）的调查显示，R 语言在数据分析、数据挖掘和数据科学领域中最受欢迎的语言。对许多数据科学家来说，R 语言不仅是一门编程语言，而且相关软件还提供了交互式的开发环境，支持运行各种数据分析任务。

R 语言在数据操作和分析方面有许多优势。下面是 3 个最显著的优势。

- 开源并且免费：用户使用 SAS 或者 SPSS 需要购买使用许可。而用户使用 R 语言是免费的，并且可以方便地学习如何实现每一个函数源代码中的统计算法。
- 强大的数据分析函数：R 语言在数据科学领域很出名。许多生物学家、统计学家和编程人员在使用 CRAN（Comprehensive R Archive Network）全球发布之前，就把他们的模型封装在 R 程序包里了。这种机制允许任何用户通过 CRAN 包下载和安装，开展分析项目。
- 易于使用：由于 R 语言是一种自解释的高级语言，使用 R 语言编程非常简单。R 语言用户只需要知道如何使用 R 函数，并借助强大的文档即可知道每一个变量如何工作，而不需要了解其背后复杂的数学知识，就可以轻松地执行高级数据分析任务。

这些优势使得复杂的数据分析变得更加简单易行。对此，R 语言用户都深信不疑。而且，R 语言尤其适合基础用户或者开发人员。对于一名 R 语言用户，我们只需知道函数如何工作，而不需要知道函数实现的具体知识。类似于 SPSS，我们可以通过 R 语言的交互式 shell，运行各种类型的数据分析任务。另外，作为一名 R 语言开发人员，我们可以编写函数来创建新的模型，甚至可以把实现的函数封装在包中。

本书并不会讲解如何从零开始编写 R 程序。相反，本书的目标是要讨论如何成为一名 R 语言开发人员。本章的主要目的是向用户展示如何定义函数，从而加速分析过程。我们首先介绍如何创建函数，然后介绍 R 环境，接着讲解如何创建匹配参数。本章的内容还会涵盖如何执行 R 语言函数式编程，如何创建高级函数，例如中缀操作符和替代，以及如何处理错误和调试函数。

## 1.2 创建 R 函数

R 语言是函数的集合；用户可以在项目中使用各个程序包中的内置函数，或者为专门的目的定义新的函数。在本教程中，我们会展示如何创建一个 R 函数。

### 准备工作

如果你是 R 语言新手，你可以在 R 语言的官方网站（<http://www.r-project.org/>）上查阅具体的介绍、历史和功能。开始下载安装 R 语言的时候，建议使用 R 语言综合文档网络（<http://cran.r-project.org/>）。

### 实现步骤

执行下列步骤来创建你的第一个 R 函数。

1. 在 R 控制台中键入下列代码，创建第一个函数：

```
>addnum<- function(x, y){
+ s <- x+y
+ return(s)
+ }
```

2. 使用下列命令，执行用户定义的函数 addnum：

```
>addnum (3,7)
[1] 10
```

或者，你也可以不使用 return 语句来定义函数：

```
>addnum2<- function(x, y){
+ x+y
+ }
```

3. 使用下列命令，执行用户定义的函数 addnum2：

```
>addnum2(3,7)
[1] 10
```

4. 你可以输入函数名来查看函数的定义：

```
>addnum2
function(x, y){
  x+y
}
```

5. 你可以使用 body 和 formals 来检查函数体和形参：

```
>body(addnum2)
{
  x + y
}
>formals(addnum2)
```



```
$x
$y
>args (addnum2)
function (x, y)
NULL
```

## 运行原理

R 函数是组织良好且可重用的代码块，通过重用代码，可以减少编程工作的重复性劳动。而且，通过对函数中语句的模块化，你的 R 代码会变得更具有可读性和可维护性。

借助这些步骤，你现在可以创建 `addnum` 和 `addnum2` R 函数，并且使用任何一个函数对两个输入参数实现加和运算。在 R 语言中，函数通常具有以下形式：

```
FunctionName<- function (arg1, arg2) {
body
return(expression)
}
```

`FunctionName` 是函数名，`arg1` 和 `arg2` 是参数。大括号内部是函数体，它是合法语句、表达式或者指认的集合。在函数底部，我们可以看到 `return` 语句，它把表达式结果返回给调用者并跳出函数。

`addnum` 函数拥有标准的语法，包含 `body` 和 `return` 语句。然而，你不一定非要在函数结尾的地方加入 `return` 语句。如 `addnum2` 函数，它本身就会把最后的表达式结果返回给调用者。

如果你想查看函数的具体细节，只需在交互式 shell 中输入函数名。可以进一步使用 `body` 和 `formals` 函数来查看函数体和形参。也可以使用 `args` 函数来获取函数的参数列表。

## 更多技能

如果你想查阅 R 函数的文档，可以使用 `help` 函数，或者只需在函数名前输入 `?`。例如，如果你想查阅 `sum` 函数的文档，你需要使用以下步骤：

```
>help (sum)
> ?sum
```

## 1.3 匹配参数

在 R 函数中，参数是激活函数的输入变量。我们可以给函数传递一般参数、命名参数、带有默认变量的参数，或者不确定数量的参数。在本教程中，我们会展示如何给定义好的函数传递各种各样的参数。

### 准备工作

确保你已经在操作系统中安装了 R 语言，完成了之前的步骤。