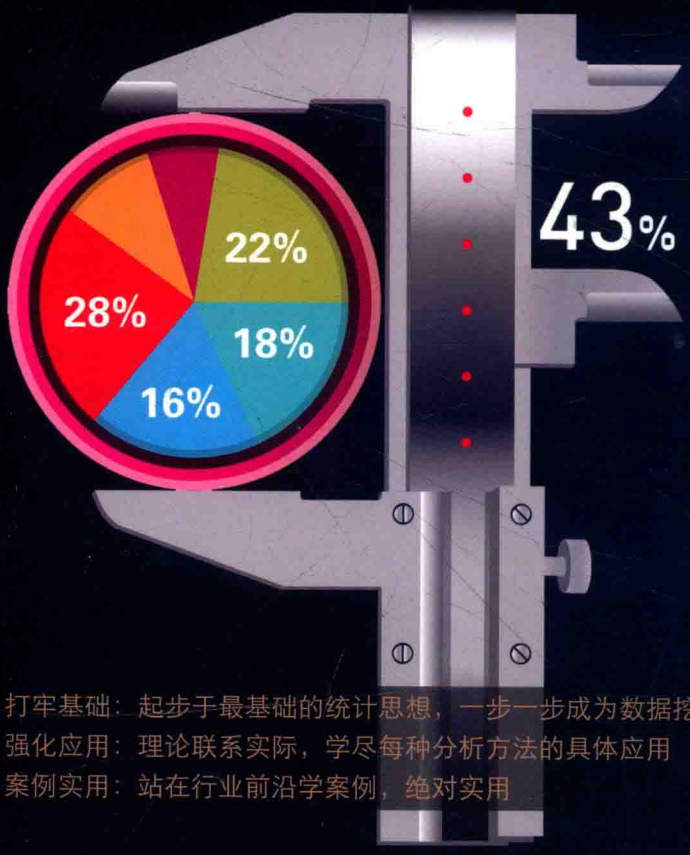
互联网+

大数据时代，
R语言助你从数据中发掘金矿

R语言实战

编程基础、统计分析与数据挖掘宝典

李倩星◎编著



打牢基础：起步于最基础的统计思想，一步一步成为数据挖掘大师
强化应用：理论联系实际，学尽每种分析方法的具体应用
案例实用：站在行业前沿学案例，绝对实用

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

R语言实战：编程基础、统计分析与数据挖掘宝典

李倩星　编著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书是一本优秀的R语言入门读物，它旨在帮助读者迅速构建起与数据分析相关的知识体系，并学习如何使用R软件实现数据分析方法。无论有无编程基础或数学基础，本书都能帮助读者成长为一名合格的数据分析师。

本书全面介绍了来自统计分析、机器学习、人工智能等领域的多种数据分析算法，在讲解与之相关的R代码时，还讨论了这些算法的原理、优缺点与适用背景。本书按照由易到难的原则组织章节主题，读者将获得最好的阅读体验。通过阅读本书，读者将对R语言在数据分析领域的应用有一个全面的认识。这种认识不被特定行业所局限，任何行业的读者都能利用本书介绍的数据分析方法解决本行业的数据分析问题。

图书在版编目（CIP）数据

R语言实战：编程基础、统计分析与数据挖掘宝典 / 李倩星编著 .—北京：电子工业出版社，2016.3

ISBN 978-7-121-28115-0

Ⅰ.①R… Ⅱ.①李… Ⅲ.①程序语言—程序设计Ⅳ.①TP312

中国版本图书馆CIP数据核字（2016）第022757号

策划编辑：李　冰

责任编辑：李　冰

特约编辑：彭　瑛　赵海军

印　　刷：三河市双峰印刷装订有限公司

装　　订：三河市双峰印刷装订有限公司

出版发行：电子工业出版社

北京市海淀区万寿路173信箱　　邮编：100036

开　　本：787×980　1/16　　印张：26.5　　字数：467千字

版　　次：2016年3月第1版

印　　次：2016年3月第1次印刷

定　　价：75.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线：(010) 88258888。

前　言

R 语言是如今最热门的编程语言之一，它由统计学家开发，在解决数据分析问题时具有先天优势。R 语言是一门新兴的语言，掌握它，就是掌握了一门高效的数据分析软件。随着大数据概念的普及，R 语言能够实现的功能越来越丰富，越来越多的数据分析从业人员产生了对学习 R 的需求。本书迎合时代潮流，讲解了大数据时代下 R 语言渗透最广泛的几个领域，全面介绍了如何使用 R 完成数据挖掘工作。对 R 语言编程人员来说，本书是一本不可或缺的工具书。

本书特色

1. 通俗易懂，实用性强，适合各层次读者学习

本书对读者的数学基础或编程基础不做任何要求。在讲解知识点时，本书采用了通俗易懂的语言，对每个疑难点都加以详细解释。此外，本书以实用为主旨，秉承“看得懂、学得会、用得上”的编写原则，精心选取了流行于行业前沿的 18 个主题，不仅通俗易懂，还确保读者所学的知识具有实际应用价值。通过阅读本书，任何读者都能迅速掌握 R 语言的编程技巧及相关的数据分析知识，并在实际工作中立刻应用它们。

2. 条理清晰，结构巧妙，全面盘点数据分析常用算法

数据分析是一个涉及多领域的交叉学科，R 软件的触角同样也能伸展到多个领域。本书选取了统计分析、机器学习、人工智能等多个学科的流行算法作为主题，讲解了如何使用 R 语言实现它们。这些算法有些偏重数学思维，有些偏重编程技巧，本书主要遵循由易到难的顺序排列主题，并尽量把起源于同一学科的算法放在一起。读者可以按照顺序阅读本书，也可以优先选择感兴趣的部分。此外，本书还穿插介绍了与 R 软件相关

的一些其他编程主题，这些主题共同形成知识网络，帮助读者迅速成长为能够独当一面的数据科学家。

3. 知识点丰富，可拓展性强，满足读者的多重需求

本书涉及多个学科，全面介绍了R软件能够实现的多种算法，满足了读者的三大需求：首先，使用通俗易懂的语言介绍R软件，帮助读者实现零基础入门；其次，囊括多种数据分析算法，带领读者全面认识R软件的强大之处，帮助读者成长为合格的数据科学家；最后，本书具备较强的可拓展性，从事任何行业的读者都能够从本书中获取适合其行业的知识。本书还给出了R语言进阶的线索，无论想向哪一方面进阶，本书都能为读者打造最坚实的基础。

本书内容及体系结构

本书分为18章，分别为R的基本介绍、原始数据的探索与预处理、R的数据可视化、R中参数的估计和检验、R中的方差分析、R中的相关分析和回归分析、更高级的数据可视化、R中的聚类分析和判别分析、R中的主成分分析和因子分析、R中的广义线性回归模型、R中的时间序列模型、R中的最优化问题、使用R绘制地理信息图形、使用R构建支持向量机、实现更高效的流程控制和高级循环、R代码的调试与优化、构建电影评分预测模型、贝叶斯垃圾邮件过滤器模型。这18章进一步又分为五部分。

第一部分为本书的第1~6章。其中前3章展示了R软件的一些入门功能，如数据预处理和数据可视化等，后3章则介绍了三种基础的统计分析方法，即参数的估计和检验、方差分析、相关分析、回归分析。这6个章节围绕初级的统计方法展开，是数据分析师必备的基本知识。

第二部分为本书的第7~11章，这5个章节介绍了更高级的统计方法。其中，第7章为第3章的延伸，介绍了数据可视化的高级方法，第8~11章则介绍了6种高级统计分析方法，这部分的内容与第一部分互为补充。

第三部分为本书的第 12~14 章，这部分内容围绕机器学习展开。第 12 章的主题为最优化，是机器学习的基本理论。第 13 章介绍了如何使用矢量化的思想绘制地图。第 14 章则介绍了支持向量机，它是最典型的机器学习算法之一。这部分讲解了更高深的 R 语言编程技巧，讨论了一些 R 软件能够解决的最高难度问题。

第 15、16 章可视为本书的第四部分。这两章围绕如何优化 R 代码展开，系统地讨论了如何写出错误较少的、运行速度较快的代码。这部分内容帮助读者建立良好的编程习惯，以及与其他 R 用户更好地协同工作。

第 17、18 章则为本书的最后一部分，这两章分别讨论了一个完整的数据挖掘项目。其中电影评分预测的案例着重于表现数据挖掘的完整流程，包括繁复的数据预处理与反复的模型比较等工作；垃圾邮件过滤的案例则引出 R 软件能够处理的另一个主题——文本分析。

上述划分方法仅为一个参考，本书的 18 个章节既互相联系又彼此独立，读者可按照上述划分方法阅读本书，也可优先阅读某些章节，如将第 3、7、13 章等与数据可视化相关的三个章节放在一起阅读。

本书读者对象

- 想要了解 R 语言的数据分析从业人员。
- 统计学、金融学、计算机技术与科学等专业的学生。
- 想要提高 R 语言编程能力的数据分析师。
- 希望系统学习统计分析方法的从业人员。
- 其他对 R 语言有兴趣的各类人员。

目　录

第1章 R的基本介绍

作为一门新兴的编程语言，R是如今值得学习的语言。由统计学家开发出的R语言具有许多奇特性质，本章将较为全面地介绍R的特性和用途，并讲解R的安装方法、变量类型、从其他数据源读取数据、程序包等基本知识。本章帮助读者对R形成整体印象，同时本章内容也是后续章节的基石。

1.1 强大的R

R语言脱胎于S语言，是一门专门用于处理数据探索、统计分析等任务的编程语言。它由统计学家开发完成，在数据分析方面具有天然的优势，运行R程序的R软件是如今最流行的统计软件之一。

与其他统计软件相比，R软件最特别的地方在于它是开源的。这同时意味着：第一，R是免费的；第二，R的用户能够自由地参与到R的开发中。R社区将它的忠实用户紧紧地黏在一起，这些用户主要由统计学家、计算机学家、数据分析师等组成，不同领域的用户在R社区中交流碰撞，协助R核心团队丰富和完善R的功能。

R的用户之间具有非常紧密的联系，他们最大的贡献是创建了形形色色的程序包，这些程序包分别封装了一些具有特定作用的函数。如今，R软件已经内置了非常丰富的函数种类，能够满足绝大多数统计人员的各类需求，它的制图功能也远超其他统计软件。

R的另一个特点在于它支持混合型的编程范式。R是一种解释型的语言，当用户在R软件中编写好一条代码后，R会立即执行它。这种做法的好处在于用户可以即时地看到程序的返回结果，在作图时尤其方便。R是一种面向对象的语言，同时它也支持函数式编程，即用户可以在R中调用现成的或自己编写的函数，这一点与C语言较为相似，但R要比C语言更加灵活。

尽管R的优点很突出，但它也同样具有局限之处。首先，R语言的编程原理较为传统，在处理数据时，R需要将数据全部载入内存，这一点极大地影响了R的运行效率，尽管如今的计算机内存做得越来越大，但在有些大规模数据集的处理工作中，R还是会显得不够得力。其次，R软件的保密性不如SAS等统计软件好，这限制了R在大型商业项目中的应用。最后，由于R软件是由统计学家开发的，因此其语法设计并不特别严谨，有时它会出现一些奇怪的错误。

随着大数据时代的到来，R语言正被越来越多的人关注，不仅是统计分析和数据挖掘，一些研究机器学习和模式识别的专家同样关心R的发展。根据TIOBE提供的编程语言排行榜，R语言的流行程度在近几年内已经飙升至前十名，其火爆程度只有python才能与其比肩，而同为统计软件的SAS和MATLAB则一直徘徊于二三十名的位置。

R的优点使它广泛地流行于统计人员和中小型商业公司中。Google、百度等互联网巨头则将R语言看作一个沙盘，使用软件验证各种数据模型的可行性，并最终使用其他语言实现。随着R的用户越来越多样化，其可扩展能力进一步强化，能够解决的问题也越来越丰富。如今，金融、医药、教育、社会科学等每一个需要数据分析的领域都需要精通R的人才。

1.2 R的安装与启动

本节介绍了R软件的安装与启动，以及几个好用的IDE，同时也比较了使用IDE执行代码与直接使用R控制台执行代码的异同。

1.2.1 安装并启动 R

R 核心团队免费向用户提供 R 软件，无论用户使用的是 UNIX、Windows 还是 Mac OS 系统，都可以在 R 网站上很方便地下载最新的 R 软件。

R 网站的主页网址为 https://www.r-project.org/，这是一个无须翻墙即可登入的英文网页，在这一主页中使用蓝色字体标出了一些超链接，分别链向与 R 相关的其他网页，比如 FAQ（常见的问题及答案）网页，以及有关 R 的新闻等。

在 R 主页的 Getting Started 标题下方有一个蓝色加粗的 download R，它指向一个 CRAN 的镜像网址汇总页，汇总了分布在全球各个国家的许多镜像服务器，其中我国总共有 4 个，用户可以选择最近的镜像。在镜像页中，用户根据自己的计算机系统选择对应的 R 软件版本，即可点击下载。R 3.2.2 是目前最新的版本，Windows 版本只有 62.2MB 大小，此时下载的 R 仅包含最基础的函数，在后续的学习中还需要陆续添加其他的程序包。

执行下载好的 exe 文件，Windows 系统中将弹出一个安装向导，R 软件的安装目录默认为系统盘，但将其安装到其他盘也不影响使用。需要注意的是在第四步中，安装向导要求用户选择组件时，最好不要选择默认的选项。

如图 1.1 所示，当安装向导执行到第 4 步时，用户组件共有 4 个组件可供选择，默认选项是将这 4 个组件全部安装。在“用户安装”下拉框下还有“32 位用户安装”、“64 位用户安装”、“自定义安装”三个选项，用户只需根据自己的机型选择 32 位或 64 位即可，不需要将 4 个组件全部下载。

R 成功安装完毕后，桌面将出现一个宝蓝色的 R 图标，只需双击 R 图标，即可启动 R 软件。

Windows 系统下运行 R 的工具也称为 RGui，即 R 图形用户界面。图 1.2 所示的截图显示这是一个 64 位的 RGui。在 RGui 的菜单栏中有一些中文选项及一些快捷操作，其中最常用的菜单选项是文件菜单和程序包菜单。

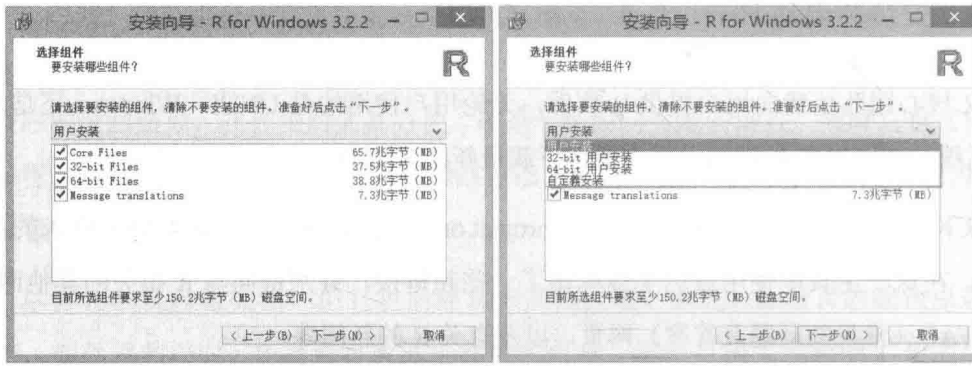


图 1.1　选择 R 软件的组件

在 RGui 中间还有一个更小的 R Console，也称为 R 控制台，控制台是用户执行代码的地方。蓝色的文字对 R 进行了一些基本声明，并给出了一些用以查看帮助文件的命令。在声明文字下方有一个醒目的红色“>”形提示符，其后跟随了一个闪烁的红色光标。“>”提示用户在此输入命令，只需回车，命令就会被 R 执行，同时 R 会在新的一行上再次生成“>”提示符。显然，退出 RGui 时，只需点击右上角红色小叉即可。

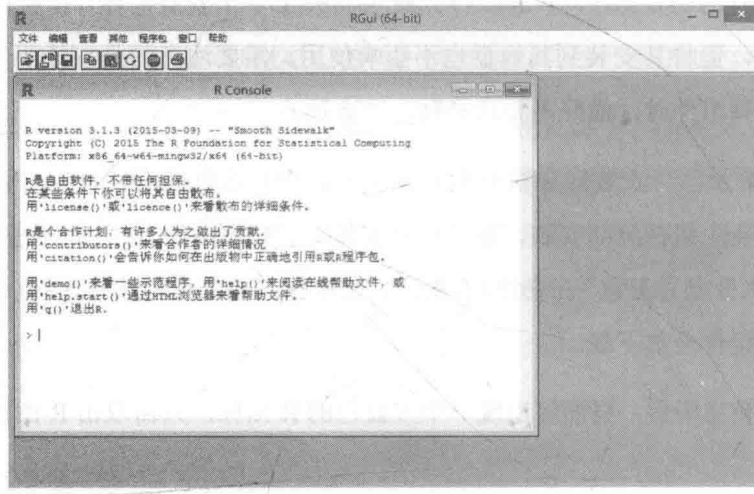


图 1.2　R 控制台

在 Mac OS 环境下运行 R 的工具是 R.app，在 UNIX 环境下启动 R 时需要在路径中包含 R 文件，然后输入命令 R 即可。

1.2.2 安装并启动一个 IDE

R 软件虽然提供了文本编辑器，但为了更方便地使用它，大多数用户都会选择额外安装一个 IDE（集成开发环境）用于辅助编程。IDE 提供一个图形开发环境，语法编辑功能也通常更为强大。

在此向统计人员强烈推荐 RStudio 软件。RStudio 是一个专门为 R 定制的免费 IDE，它将 R 中的特色功能体现得淋漓尽致。网址 https://www.rstudio.com/products/rstudio/download/ 提供了适用于不同计算机平台的 RStudio 版本，包括 Windows、Mac OS、Ubuntu、Fedora 等，用户只需在“Installers for Supported Platforms”标题下方选择合适的版本即可。

RStudio 软件的安装十分简单，仅需设定安装目录和文件夹名称即可。为了方便起见，通常把 R 和 RStudio 放在同一个目录下。安装完毕后，桌面上同样会生成一个宝蓝色的、与 R 稍有区别的 RStudio 图标，双击图标即可启动 RStudio 软件。

如图 1.3 所示，RStudio 软件被切分为 4 个小窗口。其中，左上角的窗口用于展示 R 脚本，用户可以在这里打开已经写好的 R 脚本，或者在这里编写一个脚本而不立刻执行它。左下角的窗口则是 R 软件的控制台，它其实就是 1.2.1 节中提到的 R 控制台，在用户打开 RStudio 时，RStudio 会在后台启动 R 软件。

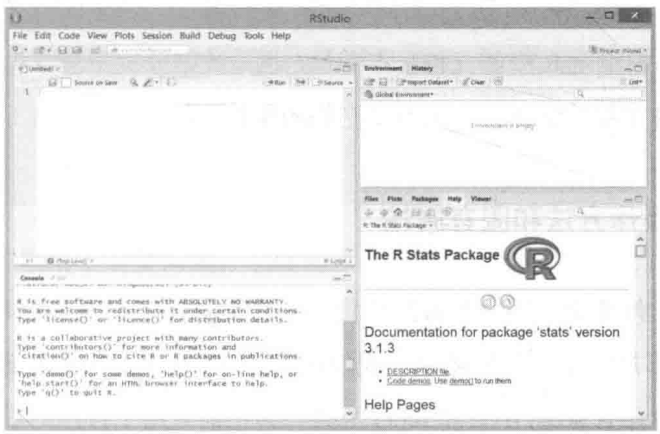


图 1.3 RStudio 软件启动界面