

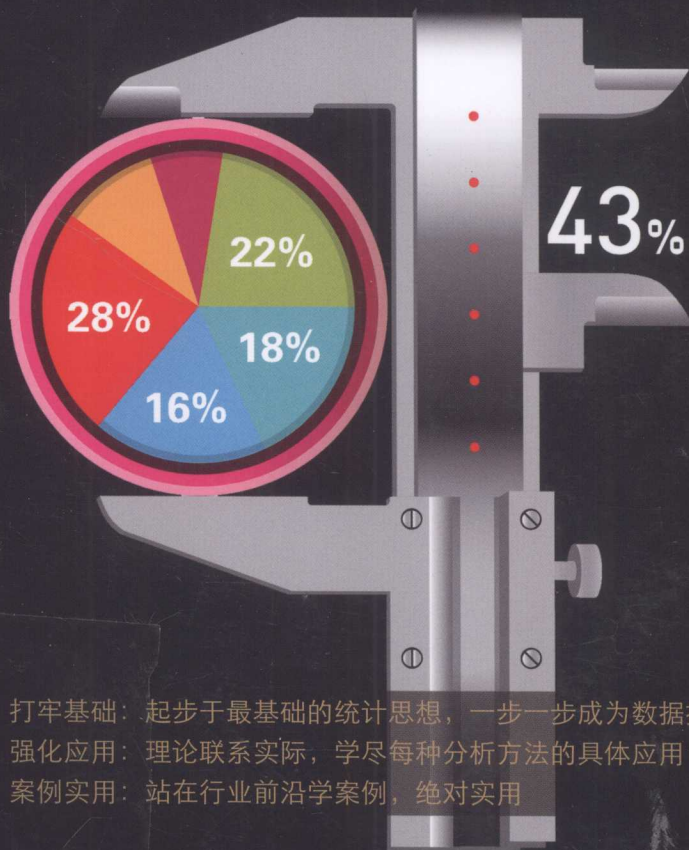
互联网+

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

电子工业出版社
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

电子工业出版社

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。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

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 . ① R… II . ①李… III . ①程序语言—程序设计 IV . ① TP312

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zltz@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计算法放在一起。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	1
1.1 强大的 R	1
1.2 R 的安装与启动	2
1.2.1 安装并启动 R	3
1.2.2 安装并启动一个 IDE	5
1.3 R 的向量、矩阵和数组	6
1.3.1 向量的操作方法和固有属性	6
1.3.2 矩阵的操作和运算	9
1.3.3 数组中的维度函数	12
1.4 R 的列表和数据框	14
1.4.1 列表的特性和编辑方法	14
1.4.2 数据框的创建和基本操作	18
1.5 R 数据文件的载入和载出	20
1.5.1 结构化纯文本文件的读取和输出	21
1.5.2 其他文件的读取和输出	23
1.6 向 R 中安装包	25
第 2 章 原始数据的探索与预处理	29
2.1 度量数据集的集中程度	29

2.2 度量数据集的分散程度	31
2.2.1 极值、方差和标准差	31
2.2.2 标准误和偏度系数、峰度系数	33
2.3 创建一个数值摘要表	35
2.4 异常值的观测与说明	37
2.4.1 利用箱线图观测异常值并处理	38
2.4.2 异常值检测的其他情况和说明	40
2.5 缺失值的填补与处理	42
2.5.1 删除缺失值或对其进行简单填补	42
2.5.2 按照相关性对空缺值进行填补	45
第 3 章 R 的数据可视化	47
3.1 plot() 函数和常用的图形参数	47
3.1.1 设置 plot() 函数中的参数	47
3.1.2 修改散点图的坐标并加入标注	51
3.2 经典的基础图形及用途	54
3.2.1 线图	54
3.2.2 直方图	59
3.2.3 箱线图和茎叶图	63
3.3 将图形组合起来	66
3.4 更多的高水平作图函数	69
3.5 更多的常用作图命令	72



第 4 章 R 中参数的估计和检验	75
4.1 使用 R 进行点估计和区间估计	75
4.1.1 简单的点估计和区间估计	75
4.1.2 估计单侧置信区间	79
4.2 与正态总体有关的参数检验	83
4.3 列联表与独立性检验	87
4.4 几种检验数据分布的函数	89
4.5 对非正态总体的区间估计和检验	92
4.5.1 非正态总体的区间估计	92
4.5.2 非参数检验中的符号检验	94
4.5.3 非参数检验中的秩检验	96
第 5 章 R 中的方差分析	99
5.1 方差分析模型的建立	99
5.2 单因素方差分析	100
5.2.1 单因素方差分析的数学思想与模型	101
5.2.2 检验样本是否满足方差分析的假设条件	102
5.2.3 构建单因素方差分析模型	105
5.3 多因素方差分析	108
5.3.1 多因素方差分析的数学思想与模型	108
5.3.2 不考虑交互作用的双因素方差分析	110
5.3.3 考虑交互作用的双因素方差分析	112

5.4 秩检验和协方差分析	114
5.4.1 对控制变量应用秩检验方法	114
5.4.2 协方差分析的假设与应用	116
第 6 章 R 中的相关分析和回归分析	118
6.1 多种相关系数的度量和分析	118
6.1.1 简单相关系数的计算和检验	118
6.1.2 散布矩阵图和偏相关系数	121
6.1.3 典型相关分析	123
6.2 线性回归分析及其常规参数	125
6.2.1 对数据进行预处理	126
6.2.2 构建第一个回归模型	127
6.2.3 修正方程并检验残差	129
6.3 使用逐步回归筛选自变量	132
6.3.1 逐步回归的思想与分类	132
6.3.2 构建逐步回归模型	133
6.4 哑变量和逻辑回归	135
6.4.1 哑变量和逻辑回归的思想	135
6.4.2 向线性回归模型中纳入哑变量	137
第 7 章 更高级的数据可视化	140
7.1 基础图形的拓展与延伸	140
7.1.1 绘制分类散点图并添加图标	140



7.1.2	绘制含多种类别的密度分布图	143
7.1.3	复合条形图和堆栈条形图	146
7.2	有关多元分布函数的特殊图形	149
7.2.1	星图和脸谱图	150
7.2.2	轮廓图	153
7.2.3	调和曲线图	155
7.3	建立最简单的 3D 图形	157
7.4	如何让图形更美观	160
7.5	更多的绘图包和系统	162
第 8 章	R 中的聚类分析和判别分析	164
8.1	几种聚类分析的异同	164
8.2	使用 R 实现 KNN 聚类	165
8.2.1	KNN 算法的思想和模型	165
8.2.2	使用 R 实现 KNN 聚类	167
8.3	使用 R 实现系统聚类	170
8.3.1	系统聚类的思想和模型	170
8.3.2	使用 R 实现系统聚类	171
8.4	使用 R 实现快速聚类	174
8.4.1	快速聚类的思想和模型	174
8.4.2	使用 R 实现快速聚类	176
8.5	几种判别分析模型综述	178
8.5.1	距离判别模型	179

8.5.2	Fisher 判别模型	182
第 9 章	R 中的主成分分析和因子分析	186
9.1	主成分分析的实现与应用	186
9.1.1	主成分分析的模型假设和数据处理	186
9.1.2	构造一个主成分分析模型	189
9.1.3	计算主成分的综合得分	191
9.2	因子分析的初次构建与完善	193
9.2.1	构造一个简单的因子分析模型	194
9.2.2	计算因子得分并分析	196
9.3	对因子分析模型进行修正	198
9.3.1	修改因子分析模型中的因子个数	198
9.3.2	基于主成分法和主轴因子法进行因子分析	200
9.4	在降维分析的基础上进行回归分析和聚类分析	202
9.4.1	在降维分析的基础上进行回归分析	202
9.4.2	在降维分析的基础上进行聚类分析	206
第 10 章	R 中的广义线性回归模型	209
10.1	一般的广义线性回归模型	209
10.1.1	使用二次函数拟合线性回归模型	209
10.1.2	拟合更多的广义线性模型	212
10.1.3	比较线性模型的优劣	214
10.2	Logistic 线性回归模型	217
10.2.1	Logistic 模型的原理与构建方法	217



10.2.2	Logistic 模型的显著性检验和优势比	220
10.2.3	修正被警告的 Logistic 模型	221
10.3	泊松回归分析模型	224
10.3.1	拟合第一个泊松回归模型	224
10.3.2	泊松回归模型的过散布检验	228
10.4	广义线性模型的交叉验证	230
第 11 章	R 中的时间序列模型	233
11.1	将数据转换为时间序列格式	233
11.1.1	使用 ts() 函数转换数据格式并绘制时间序列曲线	233
11.1.2	使用 zoo() 函数转换数据格式并绘制时间序列曲线	236
11.2	分解时间序列并检验时间序列的自相关性	239
11.2.1	使用经典方法分解时间序列	239
11.2.2	使用 STL 方法分解时间序列	241
11.3	探究时间序列的自相关性	243
11.3.1	使用月图和季度图探究自相关性	243
11.3.2	使用散点图探究自相关性	245
11.4	构建时间序列并预测	248
11.4.1	均值预测、单纯预测和漂移	248
11.4.2	不考虑长期趋势和季节波动的简单指数平滑	251
11.4.3	在指数平滑中加入长期趋势和季节波动	253
11.4.4	自回归移动平均模型	256

第 12 章 R 中的最优化问题	260
12.1 最优化问题简述	260
12.2 黄金分割法	261
12.2.1 黄金分割法和局部最优解	262
12.2.2 使用 R 实现黄金分割法	264
12.3 牛顿最优化方法	266
12.3.1 牛顿法的算法原理	266
12.3.2 在一维情形下实现牛顿迭代法	268
12.3.3 在多维情形下实现牛顿迭代法	271
12.4 最快上升法	273
12.4.1 利用梯度求解上升最快的相邻点	273
12.4.2 构建最快上升法函数并检验	275
12.5 R 中的最优化函数.....	277
第 13 章 使用 R 绘制地理信息图形	280
13.1 绘制世界、国家、省市地图	280
13.1.1 使用 map() 函数绘制地图	280
13.1.2 另一种绘制地图的方法	282
13.1.3 分省市绘制地图	285
13.2 向地图中添加颜色	289
13.2.1 向地图中添加颜色前的准备工作	289
13.2.2 在地图上添加颜色	291



13.3	向地图中添加标签和线条	293
13.3.1	向地图中添加标签前的准备工作	294
13.3.2	在地图上添加标签	296
13.3.3	在地图上添加线条	299
13.4	使用其他格式的文件优化地图	301
第 14 章	使用 R 构建支持向量机	306
14.1	构建一个简单的支持向量机	306
14.1.1	支持向量机的算法原理	306
14.1.2	构建一个简单的支持向量机	309
14.1.3	使用其他核函数构建支持向量机	312
14.2	优化支持向量机的参数	316
14.2.1	优化参数 degree	316
14.2.2	优化参数 cost	319
14.2.3	优化参数 gamma	322
14.3	比较支持向量机与 Logistic 回归的优劣	326
14.4	比较支持向量机和 KNN 聚类算法的优劣	330
第 15 章	实现更高效的流程控制和高级循环	333
15.1	R 中的流程控制	333
15.1.1	if 语句的多种实现方法	333
15.1.2	ifelse 语句与花括号的结合	335
15.1.3	适合多分支情况的 switch 语句	337
15.2	R 中的 for 循环、while 循环和 repeat 循环	340

15.2.1	R 中的 for 循环和 while 循环	340
15.2.2	R 中的 repeat 循环	342
15.3	apply 家族中的循环函数	345
15.3.1	R 中的 apply() 函数	345
15.3.2	R 中的 lapply() 函数和 sapply() 函数	349
15.3.3	R 中的 tapply() 函数	352
15.3.4	R 中的 mapply() 函数	355
15.4	更多的高级循环函数	358
15.4.1	R 中的 replicate() 函数和 sweep() 函数	358
15.4.2	R 中的 aggregate() 函数	361
第 16 章	R 代码的调试与优化	365
16.1	R 代码的常见信息与警告	365
16.1.1	R 代码的正常信息与警告	365
16.1.2	R 代码中的警告处理方法	367
16.2	R 代码中的错误与错误处理方法	370
16.2.1	使用 try() 函数处理错误信息	370
16.2.2	将 try() 函数与循环相结合	372
16.3	调试 R 代码	374
16.3.1	查看调用栈或暂停代码	374
16.3.2	修改 error 选项	376
16.4	向量化编程方法	378
16.4.1	向量化编程思想	378



16.4.2	比较循环和向量的运行速度	379
第 17 章	构建电影评分预测模型	382
17.1	获取数据并探索	382
17.2	利用 recommenderlab 包处理数据	385
17.3	建立模型并评估	388
17.3.1	模型的选择与建立	388
17.3.2	模型之间的比较和评估	390
第 18 章	贝叶斯垃圾邮件过滤器模型	394
18.1	贝叶斯模型中的条件概率	394
18.2	复杂的数据预处理过程	396
18.2.1	利用 for 循环读入多封邮件正文	396
18.2.2	利用 tm 包进一步转换数据格式	398
18.2.3	将 TDM 转换成真正有用的数据框	400
18.3	利用 occurrece 值构造分类器	403
18.3.1	完成理论准备并处理测试邮件和普通邮件	403
18.3.2	创建一个函数用于比较概率	406