数据分析与决策技术丛书

HZ BOOKS 华章IT

[PACKT PUBLISHING

R Machine Learning By Example

R语言机器学习

实用案例分析

[印度] 拉格哈夫·巴利（Raghav Bali）
迪潘简·撒卡尔（Dipanjan Sarkar） 著

李洪成 潘文捷 译

本书将带你踏上数据驱动的旅程，帮助你理解R语言和机器学习的基础知识，建立你自己的动态算法来成功地处理复杂的现实世界问题

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

R Machine Learning By Example

R语言机器学习
实用案例分析

[印度] 拉格哈夫·巴利（Raghav Bali）
迪潘简·撒卡尔（Dipanjan Sarkar） 著

李洪成 潘文捷 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 语言机器学习：实用案例分析 /（印度）拉格哈夫·巴利（Raghav Bali），（印度）迪潘简·撒卡尔（Dipanjan Sarkar）著；李洪成，潘文捷译 . —北京：机械工业出版社，2017.4
（数据分析与决策技术丛书）
书名原文：R Machine Learning By Example

ISBN 978-7-111-56590-1

I. R… II. ①拉… ②迪… ③李… ④潘… 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 077888 号

本书版权登记号：图字：01-2016-8650

R 语言机器学习：实用案例分析

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：盛思源 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2017 年 6 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：15
书 号：ISBN 978-7-111-56590-1 定 价：59.00 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

The Translator's Words 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分析领域的基本技能之一。作为探索和分析数据的基本理论与工具，机器学习和数据挖掘成为时下非常热门的技术。R 作为功能强大并且免费的数据分析工具，在机器学习领域获得了越来越多用户的青睐。本书介绍了如何用 R 来进行实际应用中的机器学习，以及如何从数据中获取信息以帮助决策。

本书的作者 Raghav Bali 在机器学习领域具有丰富的实践经验。他在本书中介绍了多种机器学习算法，并且给出了机器学习最热门的 3 个领域（涵盖电子商务、金融和社交媒体领域）中的案例。对于每一个实际案例，从对案例数据的探索、整理，到模型的建立和评估，每一步都给出了详尽的步骤和 R 代码。读者从中可以掌握机器学习和 R 语言的应用与技巧，同时也可以学习相关的领域知识。

本书共分 8 章。第 1 章介绍了 R 语言和机器学习的基本概念与理论。第 2 章介绍了机器学习的核心概念和各种类型的机器学习算法与应用。第 3 章到第 8 章以现实世界中的 3 个典型机器学习案例为线索，介绍了应用 R 进行机器学习和数据分析的整个过程。它们分别是：市场购物篮分析和推荐系统、信用风险检测和预测的描述性分析与预测性分析、社交媒体数据分析。

R 本身是一款十分优秀的数据分析和数据可视化软件，其中包含大量用于机器学习的添加包（package）。本书以实际的案例为主线，通过机器学习算法的学习来组织内容，脉络清晰。读者只需要具有 R 的一些基本知识即可，不需要具备机器学习的深厚基础。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。

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。

在本书的翻译过程中，得到了王春华编辑的大力支持和帮助。本书的编辑盛思源老师具有丰富的经验，为本书的出版付出了大量的劳动，这里对她们的支持和帮助表示衷心的感谢。本书的翻译工作得到了许多机器学习和 R 软件专家的帮助与支持，在此表示感谢。

由于时间和水平所限，难免会有不当之处，希望同行和读者多加指正。

李洪成

Preface 前　言

数据科学和机器学习是当今技术领域中的顶级流行语。从零售商店到世界500强企业，每个人都在努力使用机器学习从庞大的数据中获得有价值的信息，以发展其业务。借助强大的数据处理功能、丰富的机器学习包和活跃的开发者社区，R使用户能够构建复杂的机器学习系统，解决现实世界中的数据问题。

本书将带你踏上数据驱动的旅程，从最基础的R和机器学习开始，逐步学习如何解决实际问题。

本书内容

第1章概述本书的内容，帮助你熟悉R及其基础知识。该章还简短地介绍机器学习。

第2章通过解释机器学习的基本概念，深入研究机器学习。同时，还呈现各种类型的学习算法，以及现实世界中的一些示例。

第3章开始介绍第一个项目的第一部分，使用各种机器学习技术进行电子商务产品推荐、预测和模式分析。该章针对市场购物篮分析和关联规则挖掘，检测客户的购物模式和趋势，使用这些技术进行产品预测和推荐。这些技术在零售企业和电子商务商店（例如Target、Macy's、Flipkart和Amazon）中广泛使用，用来进行产品推荐。

第4章介绍第一个项目（电子商务产品推荐、预测和模式分析）的第二部分。该章分析不同用户对电子商务产品的评论和评级，使用算法和技术（例如，用户协同过滤）设计一个推荐系统。

第5章开始介绍第二个项目，将机器学习应用到一个复杂的金融场景中，即处理信用风险检测和预测。该章介绍新的主题，研究1000名向银行申请贷款的用户的金融信用数据集。

我们将使用机器学习技术检测具有潜在信用风险以及贷款后可能无法偿还的用户，同时对未来进行预测。该章还详细介绍数据集及其主要特征，讨论处理数据时将面临的主要挑战。最后总结适合解决这一问题的最佳机器学习技术。

第 6 章基于上一章的描述分析继续进行预测分析。这里，我们将使用几种机器学习算法来检测和预测哪些客户具有潜在信用风险，即贷款后可能无法偿还的用户。这最终将帮助银行做出数据驱动的决策，决定是否批准贷款申请。我们将涵盖几种有监督学习算法，并比较它们的性能。我们将讨论评估各种机器学习算法的性能和准确度的不同指标。

第 7 章介绍社交媒体分析。首先，我们将介绍社交媒体和通过 Twitter 的 API 收集数据的过程。该章将引导你从推文（tweet）中挖掘有用的信息（包括可视化实际案例的 Twitter 数据），推文的聚类和主题建模，解决这些问题面临的挑战、复杂度和策略。我们通过例子展示如何使用 Twitter 数据计算一些强大的度量指标。

第 8 章根据 Twitter API 的知识建立一个项目，基于该项目分析推文中的情感。这个项目呈现了多种机器学习算法，用于根据推文的情感进行分类。该章还对这些结果进行比较，帮助你理解这些算法的工作原理和运行结果之间的差异。

本书需要的软 / 硬件支持

以下软件适用于本书的所有章节：

- Windows/Mac OS X/Linux
- R 3.2.0（或以上）
- RStudio Desktop 0.99（或以上）

对于硬件，没有特定的要求，因为 R 能在任何 Mac、Linux 或 Windows 系统的个人计算机上运行，但是物理内存最好不低于 4GB，这样一些迭代算法可以更快地运行。

本书适用对象

如果你对使用先进的技术从数据中挖掘有用信息来进行数据驱动决策感兴趣，那么本书将指导你如何实现。虽然 R 的基本知识非常有用，但是在阅读本书时，不需要掌握数据科学的先验经验。掌握机器学习的先验知识十分有用，但这不是必要的。

本书约定

正文中的码字、数据库表名、文件夹名、文件名、文件扩展名、路径名、虚拟 URL、用户输入和 Twitter 句柄如下所示："我们可以使用 `include` 命令包括其他上下文。"

命令行的输入或输出如下所示：

```
# comparing cluster labels with actual iris  species labels.
table(iris$Species, clusters$cluster)
```

新的术语（new term）**和重要词**（important word）以粗体显示。


警告或者重要注释。


提示和技巧。

下载示例代码

你可以在网站 http://www.packtpub.com 上从你的账户中下载本书的示例代码文件。如果你在其他地方购买了这本书，你可以访问 http://www.packtpub.com/support 网站并注册，就可以通过电子邮件方式获得相关的文件。

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。

下载本书的彩图

我们还在一个 PDF 文件中向你提供了本书中屏幕截图和图表的彩色版本。彩色图片可以帮助你更好地理解输出中的变化关系。可以从 http://www.packtpub.com/ sites/default/files/downloads/Machine_Learning_With_R_Second_Edition_ ColoredImages.pdf 下载这个文件。

关于作者 *About the Authors*

拉格哈夫·巴利（Raghav Bali）拥有印度班加罗尔国际信息技术学院（International Institute of Information Technology）信息技术硕士学位（金牌得主）。他是世界上最大的芯片公司 Intel 的 IT 工程师，在该公司主要负责分析、商务智能和应用程序开发。他曾在 ERP、金融、商务智能等领域的一些世界顶级公司从事分析和开发工作。Raghav 是一位摄影爱好者，当他不忙于解决问题时，他会捕捉生活中的瞬间。

我要感谢 Packt 出版社提供了这次机会，感谢 Kajal Thapar 和 Utkarsha S. Kadam 完美的支持和编辑，感谢让生活变得更简单、让数据科学变得更有趣的 R 社区的每一个人。

最后，我要感谢我的家人，特别是我的父母和兄弟对我的信任，本书将是一个惊喜。我还要感谢一直鼓励我的导师、老师和朋友。最后同样重要的是，特别要感谢我的同事 Dipanjan Sarkar，没有他这一切都没有可能。

迪潘简·撒卡尔（Dipanjan Sarkar）是世界上最大的芯片公司 Intel 的 IT 工程师，在该公司主要负责分析、商务智能和应用程序开发。他拥有印度班加罗尔国际信息技术学院信息技术硕士学位。他的专业领域包括软件工程、数据科学、机器学习和文本分析。Dipanjan 的兴趣包括学习新的技术、颠覆性的初创企业和数据科学。在闲暇时间，他喜欢阅读、玩游戏以及看流行的情景喜剧。他还审阅了 Packt 出版的《Data Analysis with R》《Learning R for Geospatial Analysis》和《R Data Analysis Cookbook》。

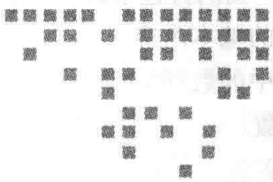
我要感谢我的好朋友和同事 Raghav Bali，谢谢他能够和我共同写作这本书。没有他的支持，这本书不可能完成。同时，我要感谢 Kajal Thapar 和 Utkarsha S. Kadam 及时向我提出修改建议，使整个写作过程充满了互动和愉快。非常感谢 Packt 出版社给我这个重要的机

会，感谢他们让我能够分享机器学习的知识。还要感谢 R 爱好者，他们每天都在做了不起的事情。

最后同样重要的是，我要感谢我的家人、朋友、老师和同事，他们一直陪伴在我的身边，支持我所有的工作。他们的支持让我每天都能迎接新的挑战！

关于审稿人 *About the Reviewer*

Alexey Grigorev 是一位熟练的数据科学家和软件工程师，拥有超过 5 年的专业经验。目前他正在 Searchmetrics 担任数据科学家。在日复一日的工作中，他使用 R 和 Python 进行数据清洗、分析和建模。在此之前，他已经是 Packt 出版的其他关于数据分析书籍的审稿人，例如《Test-Driven Machine Learning》《Mastering Data Analysis with R》。



第 1 章 Chapter 1

开始使用 R 语言和机器学习

本章是介绍性的，它将让你从基础部分学习 R 语言，包括 R 语言的各种元素、有用的数据结构、循环和向量化。如果你已经是一个 R 语言行家，你可以跳过这部分，直接进入下一章。下一章将介绍机器学习作为一个领域所代表的真正内容以及它所包含的主要方向。我们还将介绍每个领域所使用的不同机器学习技术和算法。最后，我们将通过介绍 R 语言中一些最常用的机器学习添加包结束本章，其中的一些添加包将在后续的章节中使用。

如果你是数据或机器学习的爱好者，想必一定听说过《哈佛商业评论》（Harvard Business Review）将数据科学家称作 21 世纪最热门的职业。

参考下面的链接：

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

主要由于数据科学家的主要工作是从结构化数据和非结构化数据中收集关键的洞察力和信息，以帮助他们的企业和组织战略性成长，所以对数据科学家有巨大的市场需求。

一部分人可能想知道机器学习和 R 语言如何与此相关。为了成为一名成功的数据科学家，在你的工具箱中，需要的一个主要工具是一门强大的语言，它帮助你进行复杂的统计计算，处理不同形式的数据，建立模型来获取以前不知道的信息。R 是一门能够完成这些任务的完美语言。机器学习提供了你成为一名数据分析师或数据科学家所需要的基本技能，包括使用不同的技术建立模型和从数据中获得洞察力。

本书不仅介绍 R 语言和机器学习的概念，而且还将这些概念运用在实际案例中，这些都为你熟练应用 R 和机器学习提供了必要的工具。现在，让我们开始使用 R 进行机器学习的旅程吧！

在本章中，我们将包括以下内容：

- ❑ 探究 R 的基本内容。
- ❑ 理解 R 中的数据结构。
- ❑ 应用函数。
- ❑ 控制代码流。
- ❑ 深入学习 R。
- ❑ 理解机器学习的基本内容。
- ❑ 熟悉 R 中常用的机器学习添加包。

1.1 探究 R 的基本内容

这里，假定你至少已经熟悉了 R 中的基础内容，或者以前已经使用过 R。因此，我们不会介绍太多有关下载和安装的内容。网上提供了这些部分的大量相关信息。推荐你使用 RStudio，这是一个**集成开发环境**（IDE），它比 R 自带的**图形用户界面**（GUI）更好用。可以访问 https://www.rstudio.com/ 获取更多信息。

更多关于R项目的详细内容，可以访问 https://www.r-project.org/ 获取R语言的概览。除此以外，在该网站有R语言的大量精彩的添加包，可以在网站 https://cran.r-project.org/ 浏览任何与R及其添加包的相关内容，该网站包含了大量的文档。

你必须已经熟悉 R 的交互式解释器，通常称作“**读入－求值－输出**”**循环**（Read-Evaluate-Print Loop, REPL）。这个解释器与任何等待输入的命令行界面类似，它以输入提示符>作为开始，表示 R 正在等待输入。如果输入需要多行，例如当编写函数时，在每个后续行中会有 + 提示符，这意味着你没有完成整个表达式的输入，R 要求你输入表达式的剩余部分。

R 也可以读取和执行以 .R 为扩展名的完整文件，该文件包括命令和函数。通常，任何一个大的应用程序都由多个 .R 文件组成，每个文件都在应用程序中扮演各自的角色，通常被称作一个模块。我们将在接下来的各节中探索 R 的主要特点和功能。

1.1.1 使用 R 作为科学计算器

R 中最基本的元素包括变量和算术运算符，算术运算符可以用来进行像计算器那样的数学运算，甚至复杂的统计计算。例如：

```
> 5 + 6
[1] 11
> 3 * 2
```

```
[1] 6
> 1 / 0
[1] Inf
```

记住，在 R 中的一切都是以向量形式存在的。即使在以上代码片段中的输出结果也是向量。它们都有一个先导符号 [1]，表示这是一个含有一个元素的向量。

也可以像任何其他程序设计语言一样，将值赋给变量。例如：

```
> num <- 6
> num ^ 2
[1] 36
> num
[1] 6     # a variable changes value only on re-assignment
> num <- num ^ 2 * 5 + 10 / 3
> num
[1] 183.3333
```

1.1.2　向量运算

R 中最基本的数据结构是向量。基本上，在 R 中的任何元素都是向量，即使是像上述例子中看到的一个数也是向量。向量本质上是一个序列或值的集合。可以使用 : 运算符或用于连接值的 c 函数来生成向量。例如：

```
> x <- 1:5
> x
[1] 1 2 3 4 5
> y <- c(6, 7, 8 ,9, 10)
> y
[1]  6  7  8  9 10
> z <- x + y
> z
[1]  7  9 11 13 15
```

在以上代码段中，你可以清楚地看到，我们仅仅使用 + 运算符把两个向量相加，而没有使用任何循环。这称为向量化，我们在后面将进行更多的讨论。接下来，介绍更多的向量运算，如下所示：

```
> c(1,3,5,7,9) * 2
[1]  2  6 10 14 18
> c(1,3,5,7,9) * c(2, 4)
[1]  2 12 10 28 18 # here the second vector gets recycled
```

输出：

```
Warning message:
In c(1, 3, 5, 7, 9) * c(2, 4) :
```

```
  longer object length is not a multiple of shorter object length

> factorial(1:5)
[1]   1   2   6  24 120
> exp(2:10)   # exponential function

[1]     7.389056    20.085537    54.598150   148.413159   403.428793
1096.633158
[7]  2980.957987  8103.083928 22026.465795
> cos(c(0, pi/4))   # cosine function
[1] 1.0000000 0.7071068
> sqrt(c(1, 4, 9, 16))
[1] 1 2 3 4
> sum(1:10)
[1] 55
```

你或许被上面的第2个运算搞糊涂了，这里尝试用一个较小的向量乘以一个较大的向量，但仍然得到了运算结果！如果仔细观察，就会发现R还返回了一个警告。在这个示例中，当两个向量在长度上不同时，本例中的小向量 `c(2,4)` 循环或者重复变为 `c(2,4,2,4,2)`，然后将它乘以第一个向量 `c(1,3,5,7,9)`，得到最终的结果向量 `c(2,12,10,28,18)`。这里使用的其他函数是R基础包中的标准函数。

下载本书示例代码

你可以在 http://www.packtpub.com 通过你的账户下载本书的示例代码文件。如果你在其他地方购买了本书，可以访问 http://www.packtpub.com/support 并进行注册，选择通过邮件把文件直接寄给你。

可以通过以下步骤下载代码文件：

- 使用你的电子邮件地址和密码进行登录或者注册。
- 将光标放在顶部的 SUPPORT 选项卡。
- 单击 Code Downloads & Errata。
- 在 Search 文本框中栏输入书名。
- 选择你要下载的代码文件的书。
- 在下拉菜单中选择你购买本书的地方。
- 单击 Code Download。

一旦文件下载，请确保使用以下软件的最新版本对文件夹进行解压缩：

- 用于 Windows 的 WinRAR/7-Zip
- 用于 Mac 的 Zipeg/iZip/UnRarX
- 用于 Linux 的 7-Zip/PeaZip

1.1.3　特殊值

由于在数据分析和机器学习的过程中，你将处理大量混乱和脏的数据，所以记住一些 R 中的特殊值是十分重要的，这样它们中的某一个在后面出现时，你不会太惊讶。

```
> 1 / 0
[1] Inf
> 0 / 0
[1] NaN
> Inf / NaN
[1] NaN
> Inf / Inf
[1] NaN
> log(Inf)
[1] Inf
> Inf + NA
[1] NA
```

这里你应该关心的主要值包括：Inf，代表**无穷大**（Infinity）；NaN，代表**非数值**（Not a Number）；NA 代表数值缺失或者**无效**（Not Available）。下面的代码片段展示了对这些特殊值的逻辑测试以及它们的结果。请记住，TRUE 和 FALSE 是逻辑数据类型值，类似于其他程序设计语言。

```
> vec <- c(0, Inf, NaN, NA)
> is.finite(vec)
[1]  TRUE FALSE FALSE FALSE
> is.nan(vec)
[1] FALSE FALSE  TRUE FALSE
> is.na(vec)
[1] FALSE FALSE  TRUE  TRUE
> is.infinite(vec)
[1] FALSE  TRUE FALSE FALSE
```

从这些函数的名字中，可以清晰地看出它们的作用。它们清楚地表明哪些值是有限的，哪些值是无限的，并分别检查 NaN 值和 NA 值。在清洗脏数据时这些函数十分有用。

1.2　R 的数据结构

这里将介绍 R 中最有用的数据结构，并在一些虚构的示例中使用它们，以便更好地掌握它们的语法和构造。这里将介绍的主要数据结构包括：

- 向量
- 数组和矩阵

- 列表
- 数据框

这些数据结构在 R 和 R 添加包以及函数（包括我们在后续章节中将要使用的机器学习函数和算法）中广泛地使用。因此知道如何有效地使用这些数据结构来处理数据是十分必要的。

1.2.1 向量

正如我们在上一节中简单提到的，向量是 R 中最基本的数据结构。我们使用向量来表示任何内容，包括输入和输出。我们以前知道如何生成向量以及对它们进行数学运算。这里，我们将看到更多的例子。

1.2.1.1 生成向量

这里，我们将看到初始化向量的方法，其中的一些方法我们之前已经使用过，例如：运算符和函数 c。在接下来的代码片段中，我们将使用 seq 系列的函数通过不同的方法来初始化向量。

```
> c(2.5:4.5, 6, 7, c(8, 9, 10), c(12:15))
 [1]  2.5  3.5  4.5  6.0  7.0  8.0  9.0 10.0 12.0 13.0 14.0 15.0
> vector("numeric", 5)
[1] 0 0 0 0 0
> vector("logical", 5)
[1] FALSE FALSE FALSE FALSE FALSE
> logical(5)
[1] FALSE FALSE FALSE FALSE FALSE
> # seq is a function which creates sequences
> seq.int(1,10)

 [1]  1  2  3  4  5  6  7  8  9 10
> seq.int(1,10,2)
[1] 1 3 5 7 9
> seq_len(10)
 [1]  1  2  3  4  5  6  7  8  9 10
```

1.2.1.2 索引和命名向量

选择向量子集和索引向量来访问向量的特定元素是最重要的向量运算之一，当我们仅仅想要在特定数据点上运行一些代码时，这些运算通常是很有用的。接下来的例子将介绍一些索引和选择向量子集的方法：

```
> vec <- c("R", "Python", "Julia", "Haskell", "Java", "Scala")
> vec[1]
```