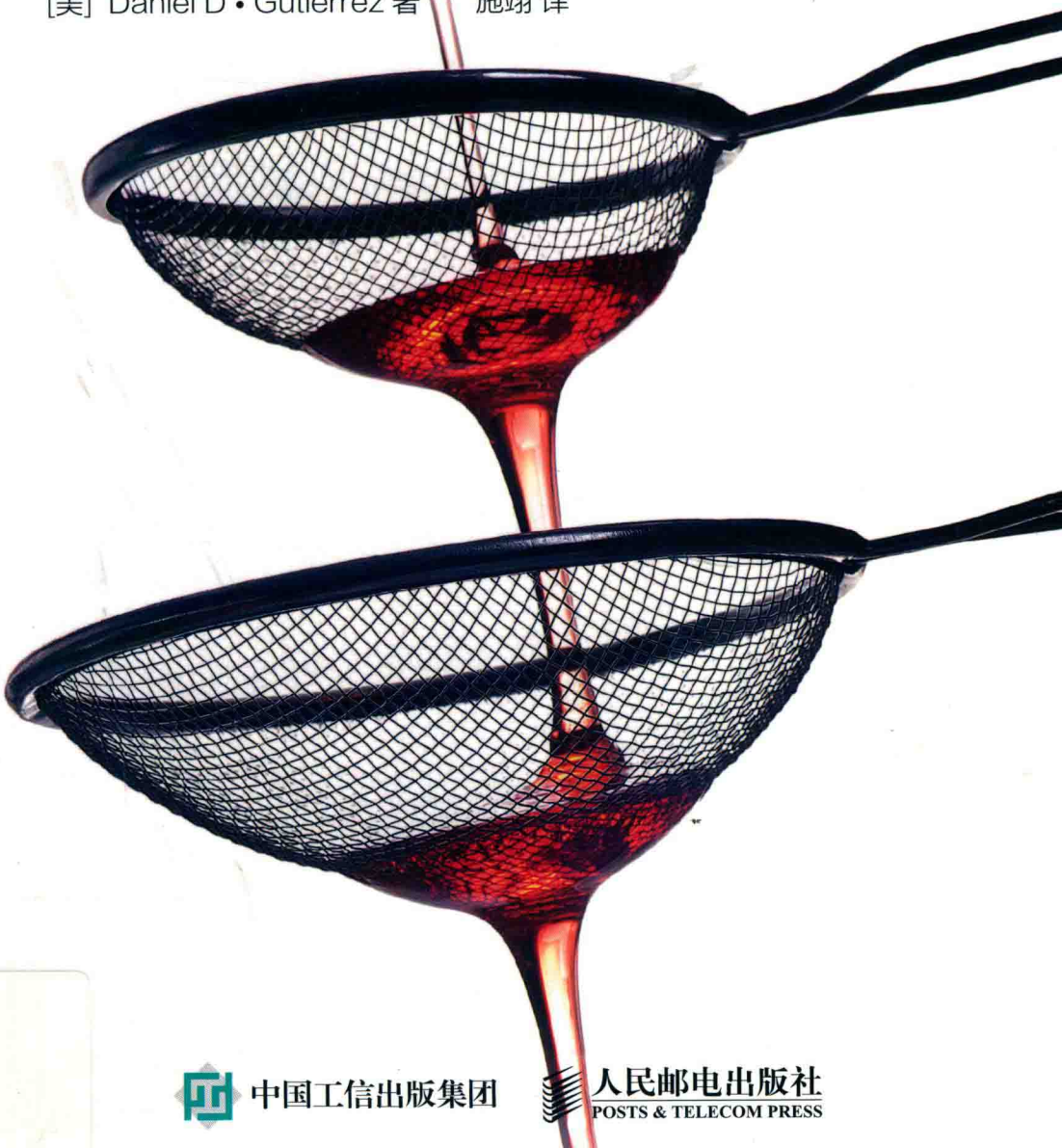
机器学习与数据科学
(基于R的统计学习方法)

[美] Daniel D • Gutierrez 著　　施翊 译



中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

机器学习与数据科学
(基于R的统计学习方法)

[美] Daniel D • Gutierrez 著　　施翊 译

人 民 邮 电 出 版 社
北　京

图书在版编目（CIP）数据

机器学习与数据科学 : 基于R的统计学习方法 /
(美) 古铁雷斯 (Daniel D. Gutierrez) 著 ; 施翊译
. -- 北京 : 人民邮电出版社, 2017.6
ISBN 978-7-115-45240-5

Ⅰ. ①机… Ⅱ. ①古… ②施… Ⅲ. ①机器学习②数据处理 Ⅳ. ①TP181②TP274

中国版本图书馆CIP数据核字(2017)第080646号

版权声明

◆ 著 [美] Daniel D. Gutierrez
译 施 翊
责任编辑 陈冀康
责任印制 焦志炜

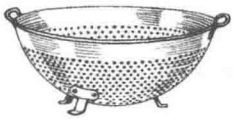
◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：720×960 1/16
印张：16.5
字数：210 千字 2017 年 6 月第 1 版
印数：1 – 3 000 册 2017 年 6 月河北第 1 次印刷

著作权合同登记号 图字：01-2016-3948 号

定价：59.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第 8052 号

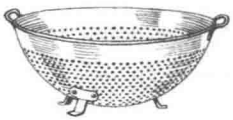


内容提要

当前，机器学习和数据科学都是很重要和热门的相关学科，需要深入地研究学习才能精通。

本书试图指导读者掌握如何完成涉及机器学习的数据科学项目。本书将为数据科学家提供一些在统计学习领域会用到的工具和技巧，涉及数据连接、数据处理、探索性数据分析、监督机器学习、非监督机器学习和模型评估。本书选用的是 R 统计环境，书中所有代码示例都是用 R 语言编写的，涉及众多流行的 R 包和数据集。

本书适合数据科学家、数据分析师、软件开发者以及需要了解数据科学和机器学习方法的科研人员阅读参考。



前言

在我的童年时代，我十分喜爱著名科幻作家、教授艾萨克·阿西莫夫（Asimov Isaac）的《基地三部曲》。故事的主角叫作 Hari Seldon，他是一位开创了“心理历史学”的数学教授，这门学科涉及历史学、社会学和数理统计，可以用来预测未来事件发生的概率。因此，我从小就迷上了预测这一概念。很自然地，我长大后成为了一名数据科学家。我把机器学习类比为 Seldon 的素数辐射法（Prime Radiant），是一个存储“心理历史学方程”的工具，可以用来展示人类未来发展前景。

远在“数据科学”这一概念问世之前，我就已经成为（或者假装成为）一名数据科学家（data scientist）很多年了。“数据科学家”这一头衔经历了数十年职业演化才建立，对此我表示十分欣喜。最近在业内论坛上，人们针对“数据科学”是否能恰当描述这一领域展开了激烈的辩论。我认为这个术语确实做出了了不起的贡献，因为数据科学家所做的事情实际上大多都是实验，这对我日复一日的基础工作毫无疑问是有效的科学方法（scientific method）。我个人认为“数据科学”比“数据挖掘”（data mining）或者“商业智能”（business intelligence）更精确、描述得更好。随着时间的流逝，后两种表述经历了严格的技术成熟度曲线[1]。我对“数据科学家”这一称呼很满意，因为我真切地感受到自己是一个用数据进行实验的科学家。

这里讲讲我是如何理解数据科学用科学方法来解决问题的。

- **提出一个问题**：问题可以是针对一个具体观察结果的解释。例如，是不是给目标顾客打的电话越多，销售团队能成交的单量也越多？这一阶段牵涉到寻找能够为解决问题提供线索的数据集。当在数据科学中使用科学研究方法时，确定一个好的问题相当不易，并且问题的好坏会直接影响到研究的最终结果。
- **作出一个假设（hypothesis）**：假设就是一个可能解释观测结果的猜想。这个假设是在提出问题时，基于现有的知识做出的。一个典型的假设表述形式是：是否批准一个房屋贷款，决定因素可能是房主的收入水平预期和信用评分。
- **预测（prediction）**：这一步骤牵涉到确定假设的逻辑结论，使用数据科学意味着选择一个合适的机器学习算法来解决这个问题。在理想状态下，预测必须把假设和其他的可能原因区分出来；如果两个假设做出了一样的预测，观测到预料中的结果就并不能说明其中一个假设是正确的。这就是为什么某些领域的机器学习需要用相同数据集、不同算法来做实验，来看最终的结果如何。这一步也需要用有限数据集来“训练”算法。
- **测试（testing）**：这一步骤是考察实际结果是否像假设预测的那样。作为一名数据科学家，你需要在训练过程中保留一份数据集，来评估预测的准确性。这一实验的目的是判断基于真实世界的观测与基于假设的预测是否一致。如果一致，该假设的置信度提高；否则，置信度降低。然而，一致性并不能确保假说的正确，更深入的实验可能会揭示其他问题。
- **分析（analysis）**：这一阶段需要确定实验得出的结论是什么，并决定下一步需要做什么。通过数据可视化，你可能发现之前在机器学习中使用的数据不足以预测得出你需要的结果。所以你回退到前面，重新审视提出问题那一步。你可能希望用不同的数据集重复实

验，来观察是否能得到相同的结果。一旦一个假说得到了数据的强烈支持，可以在同一主题下提出一个新的问题，来寻求更深入的了解。在这种情况下，科学方法是一个迭代的过程，它不断重复，直至发展出一个能继续前进的强大“理论”。

机器学习是数据科学家用来做预测和检验假设有效性的基本工具集。让我们继续简要地了解一下机器学习是什么、数据科学家用它来做什么。“机器学习”这一表述代表了多学科的融合：计算机科学（computer science）、数理统计、概率论和数据可视化。在接下去的章节中，我们将会看到机器学习有两大基本类型：监督学习（supervised learning）用于预测，非监督学习（unsupervised learning）用于发现。如果你真的想深入理解各种机器学习算法的奥秘，必须明白多个数学领域的原理，例如数理统计、概率论（probability theory）、计算学、线性代数（linear algebra）、偏微分方程（partial differential equations）和组合数学（combinatorics）。好在，在本书中我们使用了 R 语言，所以无需钻研算法的基本原理。我们只需要学习如何使用它们。

本书是如何编排的

本书希望能带领读者走进一个涉及机器学习的数据科学项目。并不是说我在这里提供的是学习机器学习的唯一方法，而是我认为这是数据科学家工作的典型方式。这一方法多年来对我十分受用，我希望通过本书把我的经验传授给大家。以下是本书的分章介绍。

- **第 1 章：机器学习综述**。这一章包含数据科学概论和企业对这一领域日益关注的原因。我们也会对机器学习做个简要介绍，包括它是如何在数据科学中扮演一个不可或缺的角色的。然后我们将回顾机器学习的不同类型，每种类型都提供示例，并提取机器学习过程的大纲。最后，我们将讨论在实验性机器学习中，R 环境如何通过使用众多的 R 包（R package）发挥重要的作用。

- **第 2 章：连接数据**。机器学习的第一步是连接到一个合适的数据集，在 R 环境下得到数据内容，然后开始对其进行分析。在这一章中，我们使用 R 来连接数据，使用不同数据源（逗号分割文件格式 CSV、Excel、JSON、Twitter 和谷歌分析）用多种方式连接。我们也会铺设一条在 SQL 数据库中连接数据的通路。一旦数据连接到 R 环境中，我们就能开始学习如何搭建一个用于数据分析和机器学习的开发环境了。
- **第 3 章：数据处理**。在开始一个机器学习项目的初期，一个冗长乏味但又不可或缺的步骤是“数据处理”，也称为“数据清洗”或是“数据转化”。换句话说，检查并精炼数据集以便进行更深入的分析。在这一章中，我们将着眼于创造一个数据处理工具箱，其中包括多种技术：修正变量名、创造新变量、数值离散化、日期处理、变量二分法、合并\按顺序排列\重塑数据集、使用 dplyr 进行数据整理以及处理缺漏数据和特征缩放。其他主题包括特征工程、数据采样和数据管道。最后，我们会一起学习主成分分析是如何做到有效降维的。
- **第 4 章：探索性数据分析**。一旦数据整理成合适的格式，下一步要做的就是熟悉数据，以便想出如何在机器学习中使用它们。在这一章中，我们会使用探索性和解释性数据可视化来理解数据的属性，寻找数据的特征，推荐建模策略。我们会从使用 R 的统计功能开始，包括数字摘要、因子变量水平、平均数\中位数\众数、分位数、标准差和变化率。我们也会使用 R 的绘图功能：直方图、箱线图、条形图、密度图、散点图、分位数图和热图。
- **第 5 章：回归**。在本章中，我们将介绍机器学习最常见的形式：监督学习。我们会仔细检视用于预测分析的主力工具：线性回归。也会学习如何在 R 环境下建立一个线性模型，并计算出一条用于预测

的回归线。单变量和多变量回归以及多项式回归都会在本章中进行演示。

- **第 6 章：分类**。在本章中，我们会介绍监督学习的另一种常见形式：分类。我们将使用大量有用的 R 包来考察各种分类算法，包括逻辑回归、分类树、朴素贝叶斯分类器、K 最近邻、支持向量机和神经网络。本章也会考虑集成方法，例如流行的随机森林算法。最后，我们会学习梯度提升机，它在机器学习比赛中十分流行。
- **第 7 章：评估模型性能**。本章会讨论如何挑选模型，并且评估它的预测水平。我们还会讨论统计学习中影响表现的方面，比如过度拟合、偏差和方差的平衡、混杂因素和数据泄漏。同时，定义了衡量回归和分类模型准确度的标准。最后，我们将展示使模型泛化误差达到最小的交叉检验过程。
- **第 8 章：非监督学习**。本章将会介绍使用两种聚类技术的非监督机器学习：分级聚类和 K-均值聚类。在分级聚类算法的帮助下，用聚合法得到一个树状图或树形结构图，来展示元素之间的关系。然后运用 K-均值聚类，使用迭代分割法来估计聚类的中心，并把每个数据点分配到聚类节点中。最后，我们会快速地看一下另一个流行的非监督工具——主成分分析。

在介绍机器学习的过程中，为了让读者的学习过程尽可能简单和直截了当，我确定了几个基本原则。

- 我不会在代码示例中使用复杂的（或者容易混淆的）R 编程技术。当然，使用嵌入式的函数调用，一行语句就能解决一个程序问题，但是理解编程语句将会与我们的学习目的背道而驰，特别如果这是你第一次接触到 R，所以本书让一切都将保持简单。
- 在本书中，我将尽量不用到流行的 ggplot2 图形包。作为替代，我们将选择使用基础的 R 的图形函数。毫无疑问，使用基础的 R 函

数会更加直截了当。

- 我们会努力将 R 包和数据集的数目降到最低，针对每章的主题都会专注于最常用的程序包，加上一些能让过程更简便的支持包。

本书的目标读者

本书的目标受众相当广泛。如果你是一名分析师，不论在私人企业还是公众部门，需要通过从一些工具（如 Excel）中得到的特征集来扩展你的分析技巧，那么这本书适合你；如果你是一位软件开发者，需要在代码中实现机器学习，那么这本书适合你；如果你是一名学术科研人员，需要了解数据科学和机器学习方法的最新进展，那么这本书适合你。这些细分读者的共同点是：诚心诚意地想要学习这一领域基础知识，并想快速地做出一些成绩。我希望各行各业的读者都能在本书中有所收获，因此书里使用的案例涉及各个领域。

我假设你已经了解了 R 程序设计，或者通过本书给出的一些材料能快速的学会它。我们不教授 R 语言，而是把 R 作为一个快速上手机器学习的工具。好消息是本书只使用了一些很基本的 R 语言；坏消息是，众所周知，R 语言对初学者来说十分晦涩难懂。书中使用的大部分 R 代码脚本十分直白，在有必要的情况下，我会在代码中添加注释来解释。我不会浪费时间用复杂棘手的代码来介绍机器学习的概念。我希望你有足够的动力来面对快速了解机器学习这一挑战。本书会提供学习的大纲，同时下面也会给出很多附加的学习资源来帮助你完成这一过程。

你需要什么

本书不需要任何其他附加的硬件或者软件，很显然，你需要 R 统计编程环境。好在，它是开源的，可以免费使用。你可以通过访问 www.r-project.org 来获取 R 软件。它可以在各种 UNIX 平台、Windows 和 MacOS 环境下安装运

行。当你在访问 www.r-project.org 网站时，请尽可能利用上面所有的学习资源，包括 R 手册、R 期刊、图书和其他关于 R 的文档。

在学习本书的过程中，另一个强烈推荐使用的软件是 RStudio 集成开发环境（IDE）。访问 www.rstudio.com 来下载 RStudio。RStudio 是一个功能强大的 R 用户界面，免费开源，并且在 Windows、Mac 和 Linux 上都有很好的表现。在编写这本书的过程中，我频繁使用了 RStudio，也推荐你这样做。虽然你可以使用 R 自带的基本编程环境来工作，但是 RStudio 包含了很多对程序员来说很有吸引力的特色：

- 语法高亮显示，代码补全和智能排版；
- 工作窗口和数据查看器；
- 历史曲线，缩放，灵活的图片、pdf 导出；
- 集成的 R 帮助和文档；
- 可搜索的命令历史；
- 直接从源编辑器执行 R 代码；
- 便于管理在用项目的多个工作目录。

在学习本书的过程中，你也需要一些额外的 R 包（拓展 R 的统计环境）。这些 R 包也是开源的，并且能在 R 内部进行下载和安装。当具体案例出现时，我会指导你如何下载、使用 R 包。

同时，我也有意地避免读者去寻找、下载和安装本书案例中用到的数据集。在大多数情况下，我尽量使用 R 自带的数据集；在一些情况下，我们可能使用特定的 R 包带有的数据集；在少数其他情况下，我使用了 R 之外的数据集，但是我会指导你如何连接这些数据集。

R 代码和图表

你会发现本书包含了很多 R 编程代码的示例，以及使用特定命令后 R 环境所返回的结果。为了在本书中展示代码，我将把熟悉的“>”符号放在

R 控制台输入的所有命令之前。我们也会使用一种特殊的“代码字体”（courier new）以便和本书正文区分。此外，R 的输出也会用同样的代码字体显示，但是前面没有“>”标识。在阅读本书的时候，我鼓励你们自行在 R 环境中输入所有的代码示例，以便获得使用这一环境的语感。改变不同的元素来实验每个代码示例，观察输出的不同结果，这同样也是一个不错的学习方法。

为了让你的学习经历尽可能愉快，我在出版社网站（https://technicspub.com/analytics/）上传了本书中使用的所有 R 源代码和注释。同时也收录了所有的图表（很多是彩色的），它们在正文中是以灰度图的形式存在的，这会使一些图表更易于理解。日后你可以访问这一资源库得到最新的代码。

超越本书

一旦你读完了本书，你将需要一些指导，关于如何更进一步学习机器学习。好在，机器学习在过去的几年里逐渐演化完整，该领域正日益受到关注。因此，有很多资源能帮助你扩充知识面。学无止境，你只需要考虑自己想钻研到多深入。为了帮助你开始学习，这里有一个简短的资源列表。

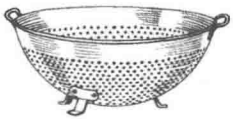
- **免费的在线课程**：在大规模在线开放课程（MOOC，massive open online course）时代，你可以找到大量关于机器学习和相关领域的优质免费课程。我最喜欢的两个 MOOC 平台是 Coursera（www.coursera.org）和 edX（www.edx.org）。相比之下，我更偏爱 Coursera 的产品，因为我试学过上面很多关于机器学习的课程，并且作为社区助教工作过一段时间。
- **博客**：很多优秀的博主提供的优质文章能帮助你更好地在这个领域学习。好在，有一个特别好的网站聚合了很多流行的 R 博客（r-bloggers）的内容：www.r-bloggers.com。常常造访这个网站或者订

阅它的每日邮件推送，你就能紧跟这个快速发展的领域的潮流了。

- **Meetup 群组**：寻找你们当地聚焦于数据科学的 Meetup 群组，并选择其中一个或者多个你感兴趣的群组加入，是一种学习的好方法。在我的家乡洛杉矶，就有许多优秀的机器学习群组和 R 群组。
- **Twitter 信息流**：就我自己而言，我从 Twitter 关注的人那里学习到了很多知识。一个人不可能监控整个业界的动向，所以我的 Twitter 朋友提供了非常有价值的服务。他们提醒我业界的新动向、文章、研究方法、产品、服务项目、会议等等。马上开始，搜索最相关的标签：#MachineLearning、#DataScience、#R 和#BigData 来找到一些你喜爱的用户。

联系作者

你可以通过多种方式找到我。访问我的咨询公司网站 www.amuletanalytics.com 或者在 LinkedIn 上都可以。然而可能最容易找到我的地方是在 Twitter 上（@AMULETAnalytics）。在那里你可以订阅我，我会发布关于数据科学、机器学习和大数据的一些感想。



目录