异步图书
www.epubit.com.cn

机器学习实践指南
—— 基于R语言

[英] 尼格尔·刘易斯（N.D. Lewis）著　高蓉 李茂 译

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

机器学习实践指南
—— 基于R语言

[英] 尼格尔·刘易斯（N.D. Lewis）著　高蓉 李茂 译

人 民 邮 电 出 版 社
北　京

图书在版编目（CIP）数据

机器学习实践指南 : 基于R语言 / (英) 尼格尔·刘易斯 (N.D. Lewis) 著 ; 高蓉, 李茂译. -- 北京 : 人民邮电出版社, 2018.4
ISBN 978-7-115-47817-7

Ⅰ. ①机… Ⅱ. ①尼… ②高… ③李… Ⅲ. ①机器学习－指南 Ⅳ. ①TP181-62

中国版本图书馆CIP数据核字(2018)第016536号

版权声明

◆ 著　[英] 尼格尔·刘易斯（N.D. Lewis）
译　高　蓉　李　茂
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
北京东方宝隆印刷有限公司印刷

◆ 开本：720×960　1/16
印张：10
字数：120 千字　2018 年 4 月第 1 版
印数：1－2 400 册　2018 年 4 月北京第 1 次印刷

著作权合同登记号　图字：01-2016-5099 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：京东工商广登字 20170147 号

内容提要

随着R语言的流行，从数据中学习比过去更加轻松。本书是通过R语言掌握机器学习和数据科学技能的快速入门指南，书中一步一步地介绍如何在免费和流行的R统计包中建立每一种类型的模型。书中的案例描述得很清楚，几乎所有的代码都可以使用。读完本书，读者将可以在自己专注的某个领域把书中所介绍的技术付诸实践。

本书适合机器学习和数据科学入门的读者阅读，尤其适合通过R语言实现数据建模和分析方法的读者学习。

献给安吉拉，她是杰出的妻子、朋友、母亲

致谢

特别感谢：

我的妻子安吉拉，感谢她的耐心和不断鼓励。

我的女儿戴安娜，感谢她为我的书和网站拍摄了数百张照片。

感谢在我更早的书中向我提问和提出建议的读者。

译者简介

高蓉，博士，任教于杭州电子科技大学，毕业于南开大学；研究领域包括资产定价、实证金融、数据科学应用；已出版教材和译著多部，发表学术论文数篇。感谢杭州电子科技大学 2016 年高等教育研究资助项目 YB201631“投资学教学与 R 软件应用”对本书翻译工作的支持。

李茂，任教于天津理工大学，毕业于北京师范大学，热爱数据科学，从事与统计和数据分析相关的教学和研究工作。

前言

感谢你阅读本书。我希望书中的这些想法能够加快你的数据科学实践，正如它们帮助了我和其他成千上万的人。事实上，我希望这本书能够让你和像你一样的成千上万的人接受数据科学工具。

人生中从来就没有足够的时间可以学习所有的知识。你差不多淹没在工作和个人责任、项目、最后期限以及一系列五花八门能消耗你一整天的任务中。本书的目的在于指导动手实践，并成为学习成功的思想、一流的技术以及数据科学家可用的从数据中学习解决方案的实用指南。

本书所介绍的内容，适用于那些为小型广告公司工作的数据科学家、由决策科学家组成的小组、为了完成数据科学项目课程作业的学生或者进行预测项目的个人顾问。即使你不是天才的统计学家或编程专家，也可以很好地理解本书讨论的实用思想和直接的解决方案。

重点在于“如何做到”，正如本杰明·富兰克林所说：“告诉我，我忘了；交给我，我记得；让我参与，我学会了。”本书的实践知识将为你提供新方法和切实可行的解决方案。

本书希望能把强大实用的机器学习技术传授给日常工作者。因此，本书的材料为重点关注数据分析和建模的个人设计。重点内容仅仅是那些已证明可行、能够迅速理解并能在最短的时间内部署的技术、思想和策略。

在许多场合，各行各业的个人都提过这样的问题：“在我关心的领域中，如何能够快速理解并应用从数据中学习要求的技术？”答案曾经是阅读复杂的数学教科书，然后使用诸如 C、C++ 和 Java 这样的语言对复杂的

公式进行编程。

随着R的兴起，从数据中学习比过去更加轻松。本书的目的在于带你快速入门。它一步步地向你展现如何在免费和流行的R统计包中建立每一种类型的模型。本书中的案例描述得很清楚，几乎可以把印在书页上的代码直接键入到R中。

对于实践者来说，这个主题最不那么激动人心的地方在于计算机制。尽管理论家必须面对这个主题的许多“可怕”之处，但是从业者并不需要重视，甚至可以通过使用R包而几乎完全忽略。本书按照惯例保留了一些算法并进行充分的讨论。但是，因为这是一本实践导向的书，指导你亲自动手，在现实数据中实现想法，所以我没有在处理算法细节、证明定理、讨论引理上花费太多的时间。

R的新用户可以轻松使用这本书，不需要任何预备知识。键入书中的实例并阅读实例下面的注意，将是你最大的收获。R的副本和免费的入门教程指南可以从 https://www.r-project.org/ 下载。如果你对R完全陌生，那么可以到 http://cran.r-project.org/other-docs.html 阅读精彩的教程。该教程向新手很好地介绍了R。

最后要注意的是，数据科学的主题并不是数学，不关注定理的证明。在根本上，它在为真实的生活、真实的人、机器学习算法应用的真实问题提供有用的解决方案。无论你是谁，无论你来自哪里，无论你的背景或教育经历如何，你将有能力理解本书概述的思想。我个人认为，结合适合的软件工具，具备一点点恒心和正确的引导，任何真正有兴趣的人都可以成功运用数据科学技术。

古希腊哲学家伊壁鸠鲁曾经说过：“我不是为大多数人而写，我为你而写；我们每个人都是另一个人的听众。”尽管本书中的思想与成千上万的人有关，但我依然努力牢记伊壁鸠鲁的原则，让读到的每一页都完全对一个人有意义，那个人就是你。

其他资源

读完本书，你将可以在自己特别关心的某个领域实践我讨论过的一个或几个内容。你会惊奇地发现，这些技术结合 R 可以快速且轻松地使用和部署。只需要一些不同的应用，你很快就能训练有素。

因此，你务必要把书中学到的知识付诸实践。为了帮助你，我创建了免费的指南“快速提高 R 语言生产效率的 12 种资源”，可到 http://www.auscov.com 下载该指南。它将和你分享 12 种可以提高 R 语言生产效率的优秀资源。

好了，现在轮到你了！

阅读本书的建议

这是一本鼓励你亲自动手操作的书。通过输入案例代码、阅读参考材料并且动手做实验，你会最大程度地获益。通过完成大量案例和阅读参考资料，你将扩展知识面，深化直观理解和强化实践技能。

另外，至少还有其他两种阅读本书的方法。你可以把它作为有效的参考工具。翻到你需要的章节，迅速查看计算如何在R中执行。如果书中的案例给出了最佳的结果类型，那么检查这些结果，并把案例调整到自己的数据上。另一种方法是观察真实世界的例子、例证、案例研究、提示以及笔记，以激发你产生自己的想法。这样既有助于学习普遍的方法，又能搞清相关例子、案例研究和文献的线索来源。

专家提示

如果你正在使用Windows操作系统，那么使用installr包可以轻松地更新到R的最新版本。输入以下代码：

```
> install . packages (" installr ")
> installr :: updateR ()
```

如果你的计算机没有安装文中提到的某个包，可以键入install.packages("package_name")进行下载并安装。例如，要下载并安装class包，你需要在R控制台键入：

```
install . packages (" class ")
```

一旦包安装完成，你要调用它。为了实现这一点，在R控制台键入：

```
require (class)
```

class 包现在可以使用了。你只需要在 R 会话开始时键入这些代码，一次就可以。

R 函数通常有多个参数。在本书的例子中，我主要关注快速模型开发需要的关键参数。在 R 控制台中键入“? function_name”，可以获取函数中可用的附加参数的信息。例如，要找到 naiveBayes 函数的附加参数，就键入：

```
? naiveBayes
```

函数和附加参数的细节会出现在默认 Web 浏览器中。在拟合你关心的模型完毕之后，我强烈鼓励你对附加参数进行实验。

在本书始终展示的 R 代码例子中，我也引入了 set.seed 方法，帮助你精确重复页面上出现的结果。

目前，主要的操作系统的 R 包都可以获取。考虑到 Windows 操作系统广受欢迎，本书示例使用 R 的 Windows 版本。

专家提示

不要为记不住两个小时前输入的内容而焦虑！我也记不住！假如你在同一个 R 会话中登录，只需要键入：

```
history (Inf)
```

它将向你返回当前会话输入命令的完整历史。

无须等到读完整本书才在自己的分析中实践学到的方法。你几乎可以立刻体验到它们神奇的力量。你可以直接翻到有兴趣的部分，直接在自己的研究和分析中检验、创造并探索知识。

专家提示

在 32 位的 Windows 操作系统计算机上，无论你安装的内存容量

有多大，R只能使用最多3GB的内存。使用下列命令可以检查内存的可用性：

```
memory . limit ()
```

使用下列命令可以从内存中移除所有的对象：

```
rm(list=ls())
```

正如标题所示，本书与数据科学模型的理解和实践有关。更确切地说，它是一种尝试，为你提供必要的R工具来轻松快捷地建立分类器。本书的目标是为读者提供完成这项工作需要的工具，并提供足够的说明，使你在自己感兴趣的领域中思考真正的应用问题。我希望这个过程不仅有益而且充满欢乐。

运用本书中的知识将改变你的数据科学实践。哪怕你在每一章只运用一个例子，在面对日益泛滥的可用数据的挑战与机遇时，你将为优胜而非仅仅生存进行了更完善的准备。

当你在自己的专业领域中成功使用了这些模型，可写信让我知道，我非常想听一听你的意见。联系我info@NigelDLewis.com或者访问www.auscov.com。

目录

第 1 章
简明学习问题

我们只能向前看到很短的路，但却可以看到还有很多事情需要做。

——艾伦·图灵

从数据中学习究竟是什么？科学家从数据中学习，企业、政府和慈善机构也一样。事实上，无论是私人、公共的，还是慈善部门的领域，几乎没有哪个领域不在部署数据驱动的模型，以发掘和利用数据中的关系。

我们置身于数据之中，亚马逊网站每天发生 2.5 万次销售 / 交付，10 万个基因几乎同时测序，超过 100 亿张图片存储在网页上。而大约在几个月之内，英国的国家卫生局对 6000 万份健康记录进行了数字化处理。我们所有人每天都在使用数据，而且许多人在工作的付薪过程中都使用了数据。营销公司的分析师必须决定，他的受众 / 听众选择模型需要包含哪些因素。本地卫生部门的研究人员测量季节性流感的发病率。气象学家运行气候模型，计算降水的可能性、温度的变化以及云层覆盖的百分比。

公共部门和某些公司需要将海量信息转换为可操作的战略性公共 / 商业决策。从数据中学习提供了一系列实践性的技术和工具，来帮助开发稳健的归纳模型，用以从数据中提取可用的见解。归纳的简单含义是指观点