O'REILLY®

基于R语言的机器学习

Introduction to Machine Learning with R

Scott V. Burger 著
马晶慧 译

中国电力出版社

基于R语言的机器学习

Scott V. Burger 著

马晶慧 译

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

O'Reilly Media, Inc. 授权中国电力出版社出版

中国电力出版社

图书在版编目（CIP）数据

基于R语言的机器学习 / (美) 斯科特（Scott V. Burger）著；马晶慧 译. — 北京：中国电力出版社，2018.11

书名原文：Introduction to Machine Learning with R

ISBN 978-7-5198-2585-0

I. ①基… II. ①斯… ②马… III. ①程序语言－程序设计 IV. ①TP312

中国版本图书馆CIP数据核字(2018)第245163号

北京市版权局著作权合同登记 图字：01-2018-3667号

出版发行：中国电力出版社
地　　址：北京市东城区北京站西街19号（邮政编码100005）
网　　址：http://www.cepp.sgcc.com.cn
责任编辑：刘 炽（liuchi1030@163.com）
责任校对：黄 蓓，郝军燕
装帧设计：Karen Montgomery，张 健
责任印制：杨晓东

印　　刷：北京天宇星印刷厂
版　　次：2018年11月第一版
印　　次：2018年11月北京第一次印刷
开　　本：750毫米×980毫米　16开本
印　　张：14.25
字　　数：270千字
印　　数：0001—3000册
定　　价：58.00元

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了《Make》杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

目录

前言

在这篇简短的介绍中，我将讨论几个关键点。

本书面向的读者对象

本书非常适合对 R 编程语言有一定了解的人。如果不了解 R 语言，那么也不用担心，R 是一种很容易学习的语言，并且代码可读性很强，相信你可以掌握代码示例中的要点。

本书的范围

本书是入门级的图书，所以我们不会深入研究每种算法涉及的数学知识。书中展示的内容可以帮助你大致掌握一些基本概念，比如神经网络与随机森林之间的区别等。

排版约定

本书使用了下述排版约定。

斜体（Italic）

表示新术语、URL、示例电子邮件地址、文件名、扩展名、路径名和目录。

等宽字体（`Constant Width`）

表示代码，在段内用以表示与代码相关的元素，如变量或函数名、数据库、数据类型、环境变量、声明和关键字。

等宽粗体字（**`Constant width bold`**）

表示命令或其他用户输入的文本。

斜体等宽字体（*`Constant Width Italic`*）

表示该文本应当由用户提供的值或由用户根据上下文决定的值替换。

表示提示或建议。

表示一般性说明。

表示警告或提醒。

O'Reilly Safari

Safari（以前的 Safari Books Online）是面向企业、政府、教育和个人的会员制培训与参考平台。

Safari 的会员可以访问成千上万的书籍、培训视频、学习路径、交互式教程和推荐的书单。这些内容由 250 多家出版社提供，其中包括：O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett 和 Course Technology 等。

更多关于 Safari 的信息，请访问我们的网站：*http://oreilly.com/safari*。

联系我们

请把你对本书的意见和疑问发给出版社：

美国：

O'Reilly Media，Inc.
1005 Gravenstein Highway North
Sebastopol，CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

这本书有专属网页，你可以在那里找到本书的勘误、示例和其他信息。这个网页的地址是 *http://bit.ly/intro_ML_withR*。

如果你对本书有一些评论或技术上的建议，请发送电子邮件到：*bookquestions@oreilly.com*。

要了解O'Reilly 图书、培训课程、会议和新闻的更多信息，请访问我们的网站：

http://www.oreilly.com
http://www.oreilly.com.cn

请在 Facebook 上联系我们，地址是：*http://facebook.com/oreilly*。

请在 Twitter 上关注我们，地址是：*http://twitter.com/oreillymedia*。

请观看我们的 Youtube 视频，地址是：*http://www.youtube.com/oreillymedia*。

致谢

写书一直是我的梦想。小学三四年级的时候，我想写的书是关于我收集的毛绒动物玩具的脱口秀。我从来没想过有一天，我的技术能力足够强到可以为大家讲解复杂的机器学习。一路走到今天，发生了很多事情，我想在这里感谢所有帮助过我写这本书的人：Allison Randal、Amanda Harris、Cristiano Sabiu、Dorothy Duffy、Elayne Britain、Filipe Abdalla、Heather Scherer、Ian Furniss、Kristen Brown、Kristen Larson、Marie Beaugureau、Max Winderbaum、Myrna Fant、Richard Fant、Robert Lippens、Will Wright 和 Woody Ciskowski。

第 1 章

什么是模型？

大学时代读物理学研究的时候，有一段时间我对模型十分感兴趣。当时的情况我记得很清楚。我们在上恒星与星系的课程，正准备学习大气层的模型，这个模型不仅适用地球，而且适用太阳系的任何行星。我很清楚气候模型非常复杂，所以突击学习了本应花费数周时间学习的数学课程。当终于学到这个课题的时候，我有点失望，因为原来我已经接触过数据模型了，只是当时自己没有意识到而已。

因为模型是机器学习的基础之一，所以无疑这个故事反映了我进行机器学习的过程。在研究生学习期间，我曾经犹豫过是否该进入金融业。我听说机器学习在金融业的应用非常广泛，作为物理专业的学生，我感觉我需要多积累计算机工程方面的知识。我又一次意识到机器学习不仅没有想象中那么可怕，而且我之前就用过机器学习，甚至在高中就接触过这方面的知识！

仪表盘可以静态地显示数据当前（或特定时间点）的状况，而模型与之不同，它可以深入数据并帮助你掌握未来的情况，因此模型非常有帮助。例如，某个从事销售工作的人可能仅熟悉静态的报告，也许他们熟知每日的销售情况。我曾经见过也创建过不计其数的仪表盘，但这些仪表盘只显示了“目前有多少资产”或者“这是我们当前的关键业绩指标。”报告是静态的，无法直观地表述事态将会怎样发展。

图 1-1 给出了报告示例：

```
op <- par(mar = c(10, 4, 4, 2) + 0.1)  #margin formatting

barplot(mtcars$mpg, names.arg = row.names(mtcars), las = 2, ylab = "Fuel
Efficiency in Miles per Gallon")
```

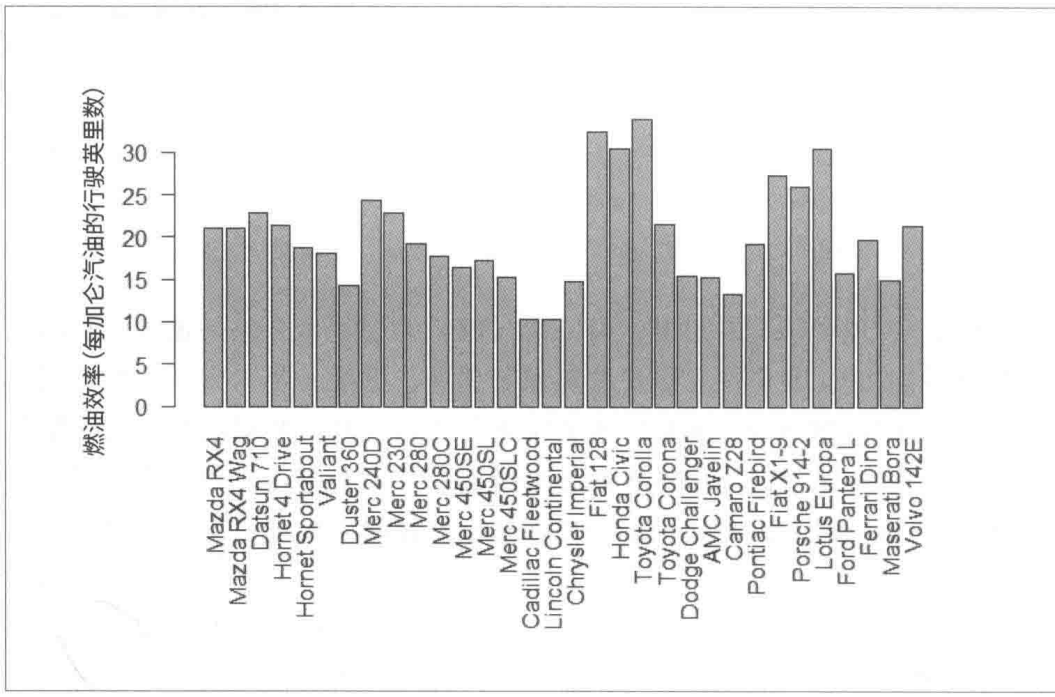

图 1-1：基于 R 语言自带的 mtcars 数据集的车辆燃油效率分布图

图 1-1 展示了 R 语言自带的 mtcars 数据集的图表。该图显示了每加仑汽油车辆行驶的英里数。这个报告没什么特别之处，它没有给出任何预测。能够显示车辆的燃油效率固然很好，但是我们怎样才能将它与数据中的其他事物联系起来，并且从中做出预测呢？

模式是具有预测能力的一种函数。

那么我们怎样才能将这个索然无味的报告转换成更加有用的东西呢？我们怎样才能将报告和机器学习结合到一起呢？通常，这个问题正确的答案是“更多数据！”我们可以更进一步地观察这组相同的数据，或者收集可以用作对比的新型数据。

让我们来仔细地看看 R 语言内置的 mtcars 数据集：

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

调出 R 语言自带的 mtcars 对象，我们可以看到数据中的各列，还可以从中选择一些来构建机器学习模型。在机器学习的世界里，数据列有时也被称为特征。现在我们了解了需要处理的数据，接下来可以试试看车辆燃油效率是否与其他特征有任何关系，如图 1-2 所示。

```
pairs(mtcars[1:7], lower.panel = NULL)
```

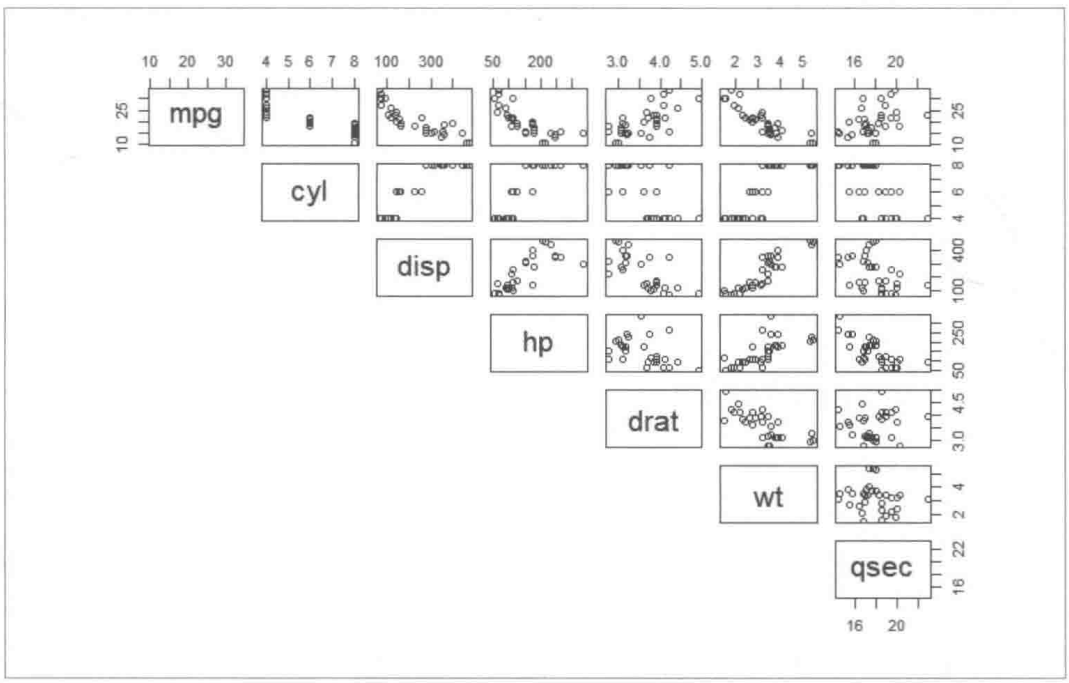

图 1-2：mtcars 数据集的对比图，重点观察前 7 行

每个框都是一张图，每列底部的文本框是因变量，而每行开头的文本框是自变量。从趋势的角度来看，一些图比另一些更加有趣。例如，看起来 `cyl` 一行中没有任何图容易做成简单的回归模型。

在这个例子中，我们绘制了一些特征与其他的对比。数据的列或特征定义如下：

mpg

每加仑汽油行驶的英里数。

cyl

车辆发动机的气缸数。

disp

发动机排量或容量（立方英寸）。

hp

发动机马力。

drat

车辆后轴比。

wt

车辆重量（千磅）。

qsec

车辆四分一英里的加速时间。

vs

车辆发动机气缸配置，“V”是指V形发动机，而“S”是指直列发动机。

am

车辆变速器类型，0是指自动挡，而1指手动挡。

gear

车辆变速器的齿轮数。

carb

车辆发动机使用的化油器数。

例如，图 1-2 的右上角表示“根据四分之一英里加速时间计算每加仑汽油行驶英里数的函数”。这里我们更加有兴趣看到一些具有量化关系的东西。这取决于调查人员选择哪种看起来更加有趣的模式。请注意“根据气缸数计算每加仑汽油行驶英里数的函数”看起来与“根据重量计算每加仑汽油行驶英里数的函数”有非常大的不同。这种情况下，我们更加关注后者，如图 1-3 所示。

```
plot(y = mtcars$mpg, x = mtcars$wt, xlab = "Vehicle Weight",
    ylab = "Vehicle Fuel Efficiency in Miles per Gallon")
```

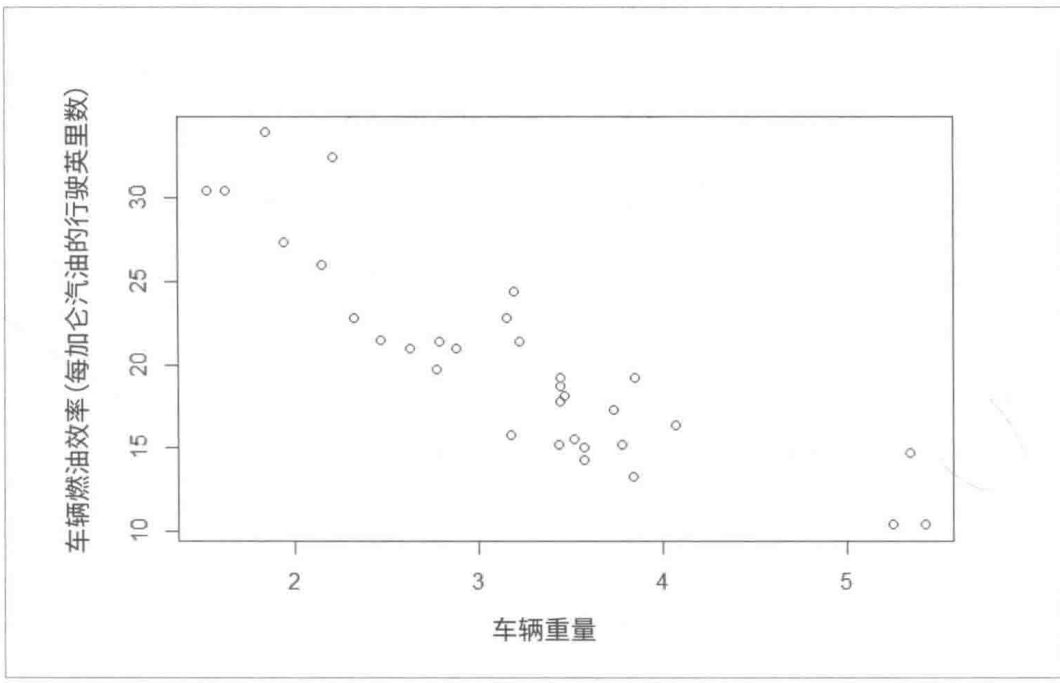

图 1-3：该图为在数据中绘制回归线的基础

现在这个数据集更有趣了。图 1-3 中仍然有燃油效率，但现在绘制的是它与车辆重量（单位为吨）之间的关系。我们可以从这种数据格式中提取所有数据点的最佳拟合，并将此图转换为方程。我们将在后面的章节中更详细地介绍这一点，这里我们利用 R 语言的函数对我们感兴趣的值（称为 response）与数据集的其他特征建模：

```
mt.model <- lm(formula = mpg ~ wt, data = mtcars)

coef(mt.model)[2]
##        wt
```

```
## -5.344472
coef(mt.model)[1]

## (Intercept)
##    37.28513
```

在这段代码中，我们建立了模型：根据车辆重量（wt）计算车辆燃油效率（mpg）的函数，并从该模型对象中提取值，用于如下方程：

燃油效率 = 5.344 × 车辆重量 + 37.285

现在如果我们想知道以上数据集以外的车辆的燃油效率，那么只需要输入车辆重量，就可以得到结果。这就是建模的好处。我们拥有预测的能力，给出某种输入（如重量），无论输入数据是什么，模型都可以给出相应的预测值。

这个模型可能有局限性，但是这种方式可以帮助我们将数据扩展到静态报告以外，建立具有更多灵活性和洞察力的东西。对于已知的车辆重量，上述方程可能无法给出实际的燃油效率。数据或观察结果可能存在一些错误。

在处理数据之前，你可能会遇到这种建模的过程。如果遇到了，那么恭喜，你已经在不知不觉中从事机器学习了！这种类型的机器学习模型称为线性回归。它比其他机器学习模型（比如神经网络）简单得多，但它使用的算法确实是机器学习的原理。

算法与模型有什么不同？

机器学习与算法密不可分。算法是另一个刚接触时让人望而生畏的主题，但其核心其实非常简单，而且你可能已经使用了很长时间，只是没有意识到而已。

算法是一套按照顺序执行的步骤。

这就是算法。算法就像穿鞋一样，首先把脚放进鞋里，然后脚尖往前蹬，脚后跟往下蹬，就穿进去了。当然，创建一套机器学习的算法远比设计穿鞋的算法要复杂得多，但是本书的目标之一就是通过简化算法的过程，解释最广泛使用的机器学习模型的内部工作原理。