
Mastering Machine Learning with R Second Edition

# 精通机器学习：基于R（第2版）

【美】Cory Lesmeister 著
陈光欣 译

- 利用R包轻松应用机器学习方法
- 展示各类机器学习方法的优势与潜在问题
- 技术与理论并重，通过丰富的商业案例实现机器学习高级概念

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

TURING 图灵程序设计丛书

Mastering Machine Learning with R Second Edition

# 精通机器学习：基于R（第2版）

【美】Cory Lesmeister 著
陈光欣 译

人民邮电出版社
北京

图书在版编目（CIP）数据

精通机器学习 : 基于R : 第2版 / (美) 考瑞·莱斯米斯特尔著 ; 陈光欣译. -- 北京 : 人民邮电出版社, 2018.3
(图灵程序设计丛书)
ISBN 978-7-115-47778-1

Ⅰ. ①精… Ⅱ. ①考… ②陈… Ⅲ. ①机器学习②程序语言－程序设计 Ⅳ. ①TP181②TP312

中国版本图书馆CIP数据核字(2018)第010916号

## 版 权 声 明



## 内 容 提 要

机器学习是近年来的热门技术话题，R 语言是处理其中大量数据的有力工具。本书为读者提供机器学习和 R 语言的坚实算法基础和业务基础，内容包括机器学习基本概念、线性回归、逻辑斯蒂回归和判别分析、线性模型的高级特征选择、K 最近邻和支持向量机等，力图平衡实践中的技术和理论两方面。

本书适合想理解和表述机器学习算法的 IT 人士、想在分析中发挥 R 强大威力的统计学专家。即使是同时精通 IT 技术和统计学的读者，在本书中仍然可以发现一些有用的窍门和技巧。

◆ 著　[美] Cory Lesmeister
译　陈光欣
责任编辑　陈　曦
责任印制　周昇亮

◆ 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
三河市君旺印务有限公司印刷

◆ 开本：800×1000　1/16
印张：19.5
字数：461千字　2018年3月第1版
印数：1－4 000册　2018年3月河北第1次印刷
著作权合同登记号　图字：01-2017-5046号

定价：69.00元

读者服务热线：(010)51095186转600　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

# 前　言

“应该给人第二次机会，但一定要留个心眼儿。”

——约翰·韦恩

人生中，能得到第二次机会可不常见。我还记得完成本书第1版的编辑工作之后，我不停地问自己：“为什么不……？”或者“我都写了些什么东西啊？”实际上，本书第1版出版之后，我做的第一个项目没有使用书中的任何一种方法。我暗下决心，如果还有机会，一定要在第2版中介绍这些方法。

当我开始写作第1版时，目标是做出点不一样的东西，在介绍各种机器学习方法的同时，还要使内容喜闻乐见。收到所有反馈之后，我认为自己实现了这个目标。但事物总是不完美的，而且，如果你想满足所有人的需要，那最终谁都满足不了。我想起了自己最喜欢的那句腓特烈大帝的名言：“诸事皆殚精竭虑者，终将一无所成。”所以，我并非一味求全，而是提供足够的技能和工具，来使读者尽量轻松愉快地学习R语言和机器学习。在第1版的基础之上，我又添加了一些非常有趣的新技术。总会有一些批评者抱怨这本书没有提供足够的数学知识，或是缺少某些方面的内容。我对这些意见的回答是：它们已经存在！为什么因为有人抱怨就要重复那些已经有人做了，并且做得非常好的事情呢？再次声明，我要写出一些与众不同的东西，一些能够抓住读者眼球并能使他们在这个充满竞争的领域取得成功的东西。

给出第2版每章内容的修改（或改进）之前，我先解释一下第2版总体上的变化。第一个总体变化就是，我放弃了一直使用 = 作为赋值操作符（而不是使用 <-）的努力。当我越来越多地与他人分享代码时，我意识到再也不能使用=，而应该使用 <-了。签下第2版合约之后，我做的第一件事就是逐行检查代码，将 = 修改为 <-。第2版更重要的一个改变是，代码更加整洁和标准化，这对于与合作者和管理者（恕我直言）分享代码也非常重要。使用版本较新的RStudio可以非常方便地实现代码标准化，写出的代码真是太标准了！嗯，首先就是要为代码加上合适的空格。举例来说，以前我会不假思索地写出`c(1,2,3,4,5,6)`这样的代码，连一个空格都懒得加。现在，我会写成`c(1, 2, 3, 4, 5, 6)`，每个逗号后面都加一个空格，这样代码就会更加易读。如果你还想了解更多代码标准，可以参见谷歌的R代码风格指南https://google.github.io/styleguide/Rguide.xml/。

我还收到了一些电子邮件，说我在网上获取的部分数据已经不存在了。国家冰球联盟已经决定使用一套全新的统计方法，所以我必须从头开始，重新做一遍那个例子。为了解决类似的问题，我把数据放到了GitHub上。

总而言之，为了给大家提供最好的工具，我尽了相当大的努力。另外，企业家马克·库班此前的一些评论在网络上引起了非常大的反响：

- “人工智能、深度学习、机器学习——如果你还不懂这些知识，那么一定要学习一下，不管你是做什么的。否则在3年之内，你就跟恐龙差不多了。”
- “我个人认为，在未来10年内，对文科专业人才的需求要超过对编程专业甚至工程专业人才的需求。因为当所有数据都呈现在面前时，我们就面临多种选择，这就需要以不同的视角来看待数据，以便得到各种不同的数据视图。所以需要更多思维更加开放的人才。”

这两条评论除了在博客圈内有一些交集之外，乍看上去彼此之间没有什么联系。但是仔细想一下，我认为他触到了我觉得自己应该写这本书的痛点。我坚信机器学习在某种程度上应该造福于大众。随着计算能力和信息可用性的不断提高，机器学习对于所有人来说都将是一种司空见惯的事情。但从另一方面看，机器学习还有一个问题，这个问题在现在和将来都会存在，那就是对结果的解释。如果你努力描述真阳性率和假阳性率时，对方一脸茫然，你应该怎么办？你怎样才能通过讲故事迅速启发听众？如果你做不到，请通知我，我非常愿意与你一起分享我的故事。

必须有人带头来做这些事情，并以此影响自己所在的组织。如果一个具有历史学或音乐鉴赏学位的人想做这些事，那就让他做吧。我每天都学习历史，它对我帮助巨大。库班的评论从多个方面使我更加确信，本书第1章最重要。如果你还没有向商业伙伴提出这个问题：“你想做些什么不一样的事情？”那么最好明天就去问。有太多人将太多努力花费在那些和组织及其决策完全无关的分析上。

## 本书内容

下面按章节给出本书对第1版做出的修改。

第1章重新制作了流程图，更正了一个无意中的输入错误，并新增了一些方法。

第2章改进了代码，并给出了更美观的图表，此外基本与第1版一致。

第3章改善并精简了代码。增加了多元自适应回归样条模型，这是我最喜欢的技术之一，它的效果非常好，可以处理非线性问题，而且易于解释。我将它作为基础模型，将其他模型作为“挑战者”，看看其他模型能否在性能上超过样条模型。

第4章不但介绍了回归模型中的特征选择技术，还包括了分类模型中的特征选择技术。

第5章梳理并精简了代码。

第6章增加了XGBOOST扩展包提供的流行技术，还增加了使用随机森林作为特征选择工具的技术。

第7章更新了一些深度学习方法的信息，并改进了使用H2O软件包的代码，包括超参数搜索技术。

第8章新增了使用随机森林进行无监督学习的方法。

第9章使用了新的数据集，新增了样本外预测的方法。

第10章新增了序列分析方法，我发现这种方法越来越重要，特别是在营销领域。

第11章属于全新内容，使用了若干个非常棒的软件包。

第12章添加了另外几年的气候数据，以及对几种不同因果关系测试方法的演示。

第13章增加了数据，改进了代码。

第14章也是新内容，帮助你在云上简单而又快速地获取R。

附录增加了新的数据处理方法。

## 准备工作

R是免费的开源软件，你只需从https://www.r-project.org/下载并安装即可。我强烈建议你从https://www.rstudio.com/products/RStudio/下载IDE和RStudio，当然，这一步不是必需的。

## 目标读者

本书的目标读者是数据科学家、数据分析师等专业人员。如果你具有使用R进行机器学习的工作经验，又想提高能力以成为机器学习领域的专家，那么本书也非常适合你。

## 排版约定

本书以不同文本样式区分不同种类的信息，下面列出并解释几种样式示例。

文本中的代码、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用户定位都表示为：“可以在R的MASS包中找到该数据框，名为biopsy。”

所有命令行输入和输出都表示为：

```
> bestglm(Xy = biopsy.cv, IC="CV",
  CVArgs=list(Method="HTF", K=10,
  REP=1), family=binomial)
```

**新名词**和**重点词**会以楷体表示。显示器屏幕（比如菜单或对话框）上的词在文本中表示为："如果想下载新模块，我们可以使用**Files|Settings|Project Name|Project Interpreter**。"

警告或重要的注意事项。

提示或小技巧。

## 读者反馈

欢迎各位提出宝贵意见，请让我们知道你对本书的看法——喜欢什么或者不喜欢什么。读者反馈对我们非常重要，因为这可以帮助我们发现对大家最有帮助的主题。

要想提供反馈，只需登录"图灵社区"本书页面（http://www.ituring.com.cn/book/1989）并留言。

## 客户支持

如果您购买了我们出版的图书，我们将提供一系列服务来使您获得最大收益。

### 下载示例代码

你可以从"图灵社区"本书页面（http://www.ituring.com.cn/book/1989）下载书中示例代码。

文件下载结束之后，请确定使用以下软件的最新版本解压或提取文件：

- Windows系统：使用WinRAR或7-Zip
- Mac系统：使用Zipeg、iZip或UnRarX
- Linux系统：使用7-Zip或PeaZip

本书的代码包也保存在GitHub上：https://github.com/PacktPublishing/Mastering-Machine-learning-with-R-Second-Edition。https://github.com/PacktPublishing/，这个地址还提供了其他种类丰富的图

书和视频资料相关代码包，好好看一下吧！

## 勘误

尽管我们做了各种努力来保证内容的准确性，依然无法避免出现错误。如果你在书中发现文字或代码错误并告知我们，我们将非常感谢。通过勘误，有助提高其他读者的阅读体验，并帮助我们在本书的后续版本中做出改进。不管您发现什么错误，都可以通过“图灵社区”本书页面（http://www.ituring.com.cn/book/1989）告诉我们。一旦勘误通过确认，将显示在页面上的勘误表中。

## 反盗版

互联网上针对有版权资料的盗版行为一直存在，并逐步扩展到所有媒体。出版社非常重视对自己版权和许可的保护，如果您在互联网上发现对于我们工作的任何形式的非法复制行为，请立即将地址或网站名通知我们，我们会采取对策。

请联系ebook@turingbook.com并提供有盗版嫌疑的链接。

如果我们在作者保护和造福读者方面得到您的帮助，我们将非常感谢。

## 问题

对本书有任何疑问，都可以登录“图灵社区”本书页面（http://www.ituring.com.cn/book/1989），我们会尽最大努力解决问题。

# 电子书

如需购买本书电子版，请扫描以下二维码。

# 第1版前言

“诸事皆殚精竭虑者，终将一无所成。”

——腓特烈大帝

机器学习领域浩瀚无边，下面的引言对此进行了很好的概括：你面临的第一个问题就是令人眼花缭乱的学习算法，到底要用哪一个？现在已经有几千种算法，每年还会发布几百种新的算法（Pedro Domingo，2012）。如果要在正文中尝试涵盖所有算法，那就是不负责任了。因为按照腓特烈大帝的意思，我们将一事无成。

请牢记这个信条。本书目的就是为你在算法和业务方面打下坚实的基础，这会消除你的困惑，最重要的是，要使你充满信心地面对每一项机器学习任务，并且理解其他算法和主题。如果这本书对你的自我提升有明显帮助，那我认为这就是胜利。不要把本书当作一个目标，而要把它当作自我发现的途径。

R的世界同机器学习一样令人不知所措，R支持社区提供了多如牛毛的R包、博客、网站、讨论组以及水平各异的论文。这是很好的信息积累，并可能是R的最大优势。但我一直坚信，一个实体的最大优势也是其最大劣势。R庞大的知识社区可以轻易地使人无所适从或步入歧途。给我一个问题和10名R程序员，会得到解决这个问题的10种不同的代码编写方式。我在每一章都会尽力找到使用R进行数据理解、数据准备和数据建模的关键要素。我绝对算不上是R编程专家，但要再次强调，我会从打好地基开始。

燃起我写这本书的热情的另一个原因，是几年前发生的一件事。我的团队中有一个负责数据库管理的IT合同工。当时，我们一边走一边聊着大数据之类的话题，他说他买了两本书，一本关于使用R进行机器学习，另一本则使用的是Python。他称自己可以完成所有编程工作，但是完全不懂其中的统计学知识。写作本书的过程中，这场谈话一直萦绕在我的脑海里。技术、理论与实践的平衡一直是一项有挑战性的工作。肯定有人可以将每章中的理论单独写一本书，或许已经有人做到了。我用一种勉强称得上是启发式的方法来判断一个公式或一项技术是否有用，比如对我或读者在与团队成员或公司老板讨论时有所帮助。如果我认为有用，就会尽力提供必要的细节。

我特意对实际使用的数据集做了处理，使它们的规模既大到足以使用，又小到足以获取知识而不至于迷失。本书不是关于大数据的，但没关系，书中讲到的方法和概念完全可以扩展到大数

据方面。

简言之，本书对很多人群都具有意义，不论是试图理解和表述机器学习算法的IT精英，还是想在分析中发挥R的强大威力的统计学大师。尽管有些人同时精通IT技术和统计学，但他们在本书中仍然可以发现一些有用的窍门和技巧。

## 定义机器学习

机器学习已经无处不在！它可以用于网页搜索、垃圾邮件过滤、推荐引擎、医疗诊断、广告投放、欺诈检测、信用评分，甚至会用在自动驾驶汽车上。公路已经相当危险了，人工智能汽车每跑100英里（约160千米）就要用CTRL+ALT+DEL来重启，它们在高速公路和辅路上漫无目的地行驶，想想就令人害怕。好吧，我跑题了。

恰当地定义我们正在讨论的事物一直都很重要，机器学习也不例外。Machinelearningmastery.com这个网站用了一整页来讨论这个问题，并提供了很好的背景资料。它提供了一个简洁的、可接受的、可操作的定义，只有一句话："机器学习是使用数据对模型进行的训练，它针对某种性能指标形成决策。"

请记住这个定义。为进行机器学习，我们会有几个要求：第一，需要数据；第二，确实存在一个模式，也就是说，通过训练数据中的已知输入值，可以基于没有用于训练模型的数据做预测或决策，这就是机器学习中的泛化；第三，需要某种性能指标，以衡量学习/泛化的结果，比如均方误差、精确度或其他指标。本书会介绍几种性能指标。

在机器学习的世界中，我发现的趣事之一就是描述数据和流程的语言的变化。说到这儿，我忍不住要引用哲学家乔治·卡林的一段话：

> "没人告诉过我这件事，也没人问我是否同意，它就这么发生了。厕纸变成了卫生纸，胶鞋变成了跑步鞋，假牙变成了牙齿矫正器，吃药变成了药物治疗，问讯处变成了查号服务，垃圾场变成了填埋地，撞车变成了交通事故，局部多云变成了局部晴朗，汽车旅馆变成了汽车客栈，房车变成了活动房屋，二手车变成了曾被拥有的运输工具，客房服务变成了客房餐饮，便秘变成了偶发性不规律。"
>
> ——乔治·卡林，哲学家、喜剧演员

当我初入机器学习殿堂时，使用的是有因变量和自变量的数据集，建立模型的目标是找到最佳拟合。现在呢？我要对实例和输入特征做标记，然后进一步选择加工，最后生成用以学习模型的特征空间。当这一切都完成之后，以前我要做的是查看模型参数，而现在要检查权重。

我要告诉你，我会交替使用这些新旧名词，以后也会一直这样。机器学习的纯粹主义者可能会因此而诅咒我，但我不认为这样会造成什么严重的问题。

## 机器学习注意事项

在我们开启香槟，从此高枕无忧地认为机器学习会解决所有社会问题之前，还有一些相当重要的事情——注意事项。在实际工作中，要时刻将其记在心间，前事不忘，后事之师。

### 失败的特征工程

仅靠堆砌数据来解决问题是不够的，不管数据量有多大。这显而易见，我有过亲身经历，也见过其他人步入这个误区。商业领袖们天真地认为，只要提供巨量原始数据，再加上机器学习应有的魔力，就能解决一切问题。我之所以在第1章重点阐述如何限定业务问题和领导期望，以上就是原因之一。

除非数据来自于精心设计的实验，或者已经进行了预处理，否则原始观测数据几乎不可能直接用来建模。在任何一个项目中，实际花在建模上的时间都非常少。最需要花费时间的环节是特征工程：数据收集、数据集成、数据清洗和数据理解。在本书的练习中，我估计与建模相比，90%的时间要花在上述环节的编码工作上，这还是在大多数数据集都非常小且易于获取的情况下。在我的实际工作中，使用SAS时，99%的时间用在了PROC SQL上，只有1%的时间用在PROC GENMOD、PROC LOGISTIC或Enterprise Miner上。

对于特征工程，人们有两种观点。一种认为专业知识必不可少（我认同这一观点），另一种认为机器学习算法可以自动完成特征选择/构建的大部分工作。一些创业公司声称这是完全可行的。（我曾经和几个家伙聊过，谈理论时他们滔滔不绝，一旦涉及具体细节就闭口不言了。）假设你有几百个候选特征（自变量），进行自动特征选择的方法是计算单变量信息值。一个特征在孤立状态下会表现为完全不相关，但与另一个特征组合起来就可能变得非常重要。为解决这个问题，需要生成无数的特征组合。这必然会带来一个潜在的问题——计算时间和成本将大幅提高，并且可能造成模型的过拟合。说到过拟合，我们会在下一个注意事项中继续讨论。

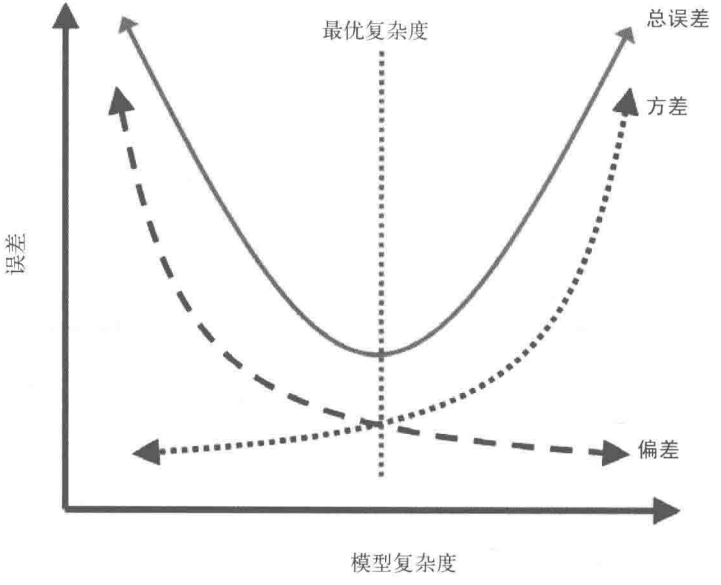
### 过拟合与欠拟合

当模型的泛化效果不佳时，就会表现出过拟合。如果你使用训练数据达到了95%的分类精确度，但使用另一组数据测试时，精确度却下降到了50%，那么模型的方差就太高了。如果训练数据的精确度为60%，测试数据的精确度为59%，那么模型的方差很小，但是偏差太大。这种偏差与方差之间的权衡是机器学习和模型复杂度的一个基本问题。

让我们先看看定义。偏差是模型的预测值或预测水平与训练数据中的实际值或实际水平之间的差别。方差是训练数据集的预测值或预测水平相对于其他数据集的预测值或预测水平的离散程度。当然，我们的目标是使总体误差（偏差+方差）最小，但这与模型复杂度又有什么关系呢？

为了说明其中的关系，假设要进行一项预测，并用训练数据建立一个简单的线性模型。这个

模型非常简单，所以具有高偏差；另一方面，训练数据和测试数据之间的方差却很小。如果我们在线性模型中加入多项式或者建立决策树，模型会变得更复杂，偏差会减小。但偏差减小的同时，模型的方差会扩大，泛化能力会降低。你可以在下图中看到这个现象。所有机器学习项目都应该尽力达到偏差和方差之间的最佳平衡点，这说起来容易，做起来很难。




我们会在其他章节讨论解决这个问题和优化模型复杂度的方法，包括交叉验证（第2~7章）和正则化（第4章）。

## 因果关系

相关性不等于因果关系，这一点应该已经广为人知，不需多费唇舌。但果真如此吗？现实世界中，明显有人依然搞不清相关性和因果关系的区别。所以，我们必须牢记并且坚定地告诉他人，算法基于观测数据而不是实验数据，不管从机器学习中得到多么完美的相关性，都不能胜过从正确的实验中得到的结论。正如佩德罗·多明戈斯教授所说：

> “如果发现人们在超市经常同时购买啤酒和纸尿裤，那么把啤酒放在纸尿裤旁边可能会提高销量。但如果没有经过实验验证，这个结论就站不住脚。”
>
> ——佩德罗·多明戈斯，2012

第11章会使用一种来自计量经济学的方法在时间序列中探索因果关系，讨论一个情感和政治敏感性的问题。

我就不再啰唆了，下面开始用R玩转机器学习吧！如果你对于R编程完全是个门外汉，我建议你跳过前面的内容，直接学习附录中的R使用方法。不管你从哪里开始阅读，请记住本书探讨的是掌握机器学习的过程，而不是要达到某个目标。只要我们在这个领域内辛勤耕耘，就会一直有令人惊喜的新事物值得我们探索。所以，我非常希望收到你的评论、想法、建议、抱怨和牢骚。就像印第安苏族勇士的口号一样："共同前进！"

## 本书内容

第1章说明机器学习不仅仅是写代码。为了使你的工作在业界具有持久的影响，我们介绍一个经过考验的流程，使你有个好的开始并走向成功。

第2章为学习支持向量机和梯度推进等高级方法打下坚实基础。没有比最小二乘线性回归更基础的方法了。

第3章讨论如何使用逻辑斯蒂回归与判别分析来预测分类结果。

第4章介绍正则化技术，帮助提高模型的预测能力和可理解性。特征选择是机器学习中最关键、最有挑战性的部分。

第5章开始研究更高级的非线性技术。机器学习的真正威力将揭开面纱。

第6章介绍几种机器学习领域内具有最强预测能力的技术，特别是对于分类问题而言。单决策树将与更高级的随机森林和提升树一起讨论。

第7章介绍一些当前应用中的最激动人心的机器学习技术。神经网络的灵感来自于大脑的工作原理，它将与其最近的高级分支——深度学习一同接受检验。

第8章开始涉及无监督学习，它的目标不是做出预测，而是把重点放在发现观测数据中的隐含结构上。我们将讨论3种聚类方法：层次聚类、K均值和围绕中心的划分（PAM）。

第9章继续研究无监督学习方法。主成分分析用来发现特征中的隐含结构，一旦发现其中的结构，新的特征将用于监督学习。

第10章介绍用来提高销量、检查欺诈和增进健康的技术。你将学习食品杂货店对于购买习惯的购物篮分析，然后研究如何在网站评估的基础上建立推荐引擎。

第11章讨论单变量预测模型、二元回归模型和格兰杰因果关系模型，还包括一个关于碳排放和气候变化的分析。

第12章展示一个定量文本挖掘框架以及如何建立主题模型。伴随着时间序列，数据世界包含着文本形式的海量数据。既然如此多的数据都是文本形式，那么懂得如何对文本数据进行处理、

编程和分析就显得特别重要。

附录介绍R的语法及其强大功能。R有一个陡峭的学习曲线，一旦你熟练掌握，就会发现对于数据准备和机器学习来说，R的威力有多么强大。

## 准备工作

R是免费的开源软件，你只需从https://www.r-project.org/下载并安装即可。我强烈建议你从https://www.rstudio.com/products/RStudio/下载IDE和RStudio，当然，这一步不是必需的。

## 目标读者

本书的目标读者是数据科学家、数据分析师等专业人员。如果你具有使用R进行机器学习的工作经验，又想提高能力以成为机器学习领域的专家，那么本书也非常适合你。

## 排版约定

本书以不同文本样式区分不同种类的信息，下面列出并解释几种样式示例。

文本中的代码、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用户定位都表示为："可以在R的MASS包中找到该数据框，名为biopsy。"

所有命令行输入和输出都表示为：

```
cor(x1, y1) #correlation of x1 and y1
[1] 0.8164205

> cor(x2, y1) #correlation of x2 and y2

[1] 0.8164205
```

**新名词**和**重点词**会以楷体表示。显示器屏幕（比如菜单或对话框）上的词在文本中表示为："如果想下载新模块，我们可以使用**Files|Settings|Project Name|Project Interpreter**。"

警告或重要的注意事项。

提示或小技巧。

## 读者反馈

欢迎各位提出宝贵意见，请让我们知道你对本书的看法——喜欢什么或者不喜欢什么。读者反馈对我们非常重要，因为这可以帮助我们发现对大家最有帮助的主题。

要想提供反馈，只需登录http://www.packtpub.com本书页面并留言。

## 客户支持

如果您购买了我们出版的图书，我们将提供一系列服务来使您获得最大收益。

### 下载示例代码

你可以从http://www.packtpub.com本书页面下载书中示例代码。

文件下载结束之后，请确定使用以下软件的最新版本解压或提取文件：

- Windows系统：使用WinRAR或7-Zip
- Mac系统：使用Zipeg、iZip或UnRarX
- Linux系统：使用7-Zip或PeaZip

https://github.com/PacktPublishing/，这个地址还提供了其他种类丰富的图书和视频资料相关代码包，好好看一下吧！

### 勘误

尽管我们做了各种努力来保证内容的准确性，依然无法避免出现错误。如果你在书中发现文字或代码错误并告知我们，我们将非常感谢。通过勘误，有助提高其他读者的阅读体验，并帮助我们在本书的后续版本中做出改进。不管您发现什么错误，都可以通过http://www.packtpub.com/submit-errata ）告诉我们。一旦勘误通过确认，将显示在页面上的勘误表中。

### 反盗版

互联网上针对有版权资料的盗版行为一直存在，并逐步扩展到所有媒体。出版社非常重视对自己版权和许可的保护，如果您在互联网上发现对于我们工作的任何形式的非法复制行为，请立即将地址或网站名通知我们，我们会采取对策。

如果我们在作者保护和造福读者方面得到您的帮助，我们将非常感谢。

## 问题

对本书有任何疑问，都可以通过questions@packtpub.com联系我们，我们会尽最大努力解决问题。

# 目　录


第 1 章　成功之路 …… 1
1.1　流程 …… 1
1.2　业务理解 …… 2
1.2.1　确定业务目标 …… 3
1.2.2　现状评估 …… 4
1.2.3　确定分析目标 …… 4
1.2.4　建立项目计划 …… 4
1.3　数据理解 …… 4
1.4　数据准备 …… 5
1.5　建模 …… 5
1.6　评价 …… 6
1.7　部署 …… 6
1.8　算法流程图 …… 7
1.9　小结 …… 10
第 2 章　线性回归：机器学习基础技术 …… 11
2.1　单变量回归 …… 11
2.2　多变量线性回归 …… 18
2.2.1　业务理解 …… 18
2.2.2　数据理解和数据准备 …… 18
2.2.3　模型构建与模型评价 …… 21
2.3　线性模型中的其他问题 …… 30
2.3.1　定性特征 …… 30
2.3.2　交互项 …… 32
2.4　小结 …… 34
第 3 章　逻辑斯蒂回归与判别分析 …… 35
3.1　分类方法与线性回归 …… 35
3.2　逻辑斯蒂回归 …… 36
3.2.1　业务理解 …… 36
3.2.2　数据理解和数据准备 …… 37
3.2.3　模型构建与模型评价 …… 41
3.3　判别分析概述 …… 46
3.4　多元自适应回归样条方法 …… 50
3.5　模型选择 …… 54
3.6　小结 …… 57
第 4 章　线性模型中的高级特征选择技术 …… 58
4.1　正则化简介 …… 58
4.1.1　岭回归 …… 59
4.1.2　LASSO …… 59
4.1.3　弹性网络 …… 60
4.2　商业案例 …… 60
4.2.1　业务理解 …… 60
4.2.2　数据理解和数据准备 …… 60
4.3　模型构建与模型评价 …… 65
4.3.1　最优子集 …… 65
4.3.2　岭回归 …… 68
4.3.3　LASSO …… 71
4.3.4　弹性网络 …… 73
4.3.5　使用 glmnet 进行交叉验证 …… 76
4.4　模型选择 …… 78
4.5　正则化与分类问题 …… 78
4.6　小结 …… 81
第 5 章　更多分类技术：K 最近邻与支持向量机 …… 82
5.1　K 最近邻 …… 82
5.2　支持向量机 …… 84
5.3　商业案例 …… 86