HZ BOOKS 华章IT

[PACKT] PUBLISHING

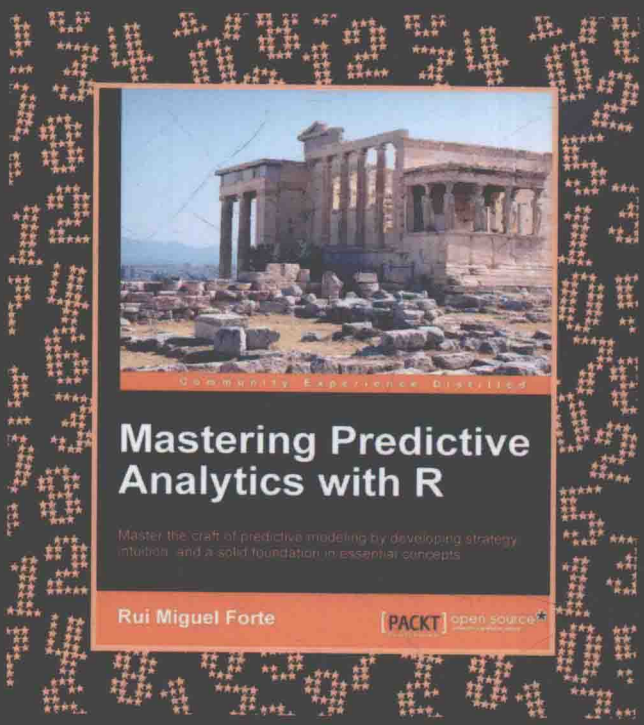
数据科学与工程技术丛书

预测分析

R语言实现

[希] 鲁伊·米格尔·福特（Rui Miguel Forte） 著

吴今朝 译



MASTERING PREDICTIVE ANALYTICS WITH R

机械工业出版社
China Machine Press

数据科学与工程技术丛书

MASTERING PREDICTIVE ANALYTICS WITH R

预测分析

R语言实现

[希] 鲁伊·米格尔·福特（Rui Miguel Forte）著
吴今朝 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

预测分析：R 语言实现 /（希）鲁伊·米格尔·福特（Rui Miguel Forte）著；吴今朝译 . —北京：机械工业出版社，2016.10
（数据科学与工程技术丛书）
书名原文：Mastering Predictive Analytics with R

ISBN 978-7-111-55354-0

I. 预… II. ① 鲁… ② 吴… III. 程序语言－程序设计 IV. TP312

中国版本图书馆 CIP 数据核字（2016）第 274893 号

本书版权登记号：图字：01-2015-5215

预测分析：R 语言实现

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：何欣阳

印　刷：藁城市京瑞印刷有限公司		版　次：2017 年 1 月第 1 版第 1 次印刷	
开　本：185mm × 260mm　1/16		印　张：16.25	
书　号：ISBN 978-7-111-55354-0		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

译 者 序

本书是一本比较全面的预测建模教材，覆盖了最常见的一些技术，例如逻辑回归、神经网络、支持向量机、隐马尔可夫模型、时间序列分析、推荐系统等。本书属于 Packt 出版社系列图书中的 Mastering 级别，是有一定难度和深度的高级教程。

作为一位兼具科研和产业经验的专家，作者很巧妙地把握了理论和实践之间的平衡。他的做法是，先以比较通俗的方式讲解理论背景，再通过一些实际案例的直观示范来帮助读者理解相关的理论和方法。这样就让读者既能对各种预测分析方法的理论基础有更深入的认识，又能掌握在实际工作中运用这些技术的方法。此外，作者还提供了大量的参考资料和在线资源，供学有余力的读者进一步提高。

在翻译完成《基于 R 语言的自动数据收集：网络抓取和文本挖掘实用指南》[⊖]一书后，我就在机械工业出版社编辑推荐的后续书单中挑出了这本书。我之所以会对这本书感兴趣，是因为之前那本书的核心内容是如何在线抓取数据，而现在这本书的核心内容则是如何对数据进行预测建模，两本书结合起来，就构成了一个完整的技术体系。

在多年的应用开发经历中，我一直更喜欢这种个人能掌控完整技术链条的工作风格，相信很多科研工作者和小团队的技术带头人也会有同感。其实这种风格和团队合作并不矛盾。实际上，只有具备了掌控全局的能力，才能提高团队合作的效率，降低沟通成本。

在之前那本书的读者评论里，有一条我想分享出来："这本书值得一读，作者很诚实，确实一本书不能解决你所有的问题，但是可以给你一些思路，顺着这个思路不断扩展自己的知识，最后娴熟运用。阅读纸质书最大的好处就是系统性，书中很多知识都通过网络资源零散地学过，但始终不成体系，本书能给你一个很好的网络数据获取的系统框架。"

确实，学习技术的过程是无止境的，但一套科学的体系能让读者把握全局，少走弯路。本书就比较系统地讲解了有监督学习的预测建模技术。

作为译者，我想给读者的一条建议就是多动手。在技术领域，看懂了书不等于掌握了技术。读者如果没有经过实际的应用，对书中内容的理解不但达不到足够的深度，而且很容易遗忘。针对自己感兴趣的某个问题，先把模型设计出来，把代码调通，再对模型进行优化，最后得到理想的结果，这个过程是非常关键的。

比如在之前那本书的 GitHub 讨论区里就有一个很有意思的话题，有一位读者在尝试抓取某个新闻下的所有评论时遇到了问题，后来才发现是 iFrame 元素的原因，进而引出了一些原书中没有讲到的技术。在和我一起分析讨论并调通了代码后，这位读者觉得收获很大。

⊖ 书号是 978-7-111-52750-3。——编辑注

照例，我也给本书开通了一个 GitHub 讨论区，链接是：https://github.com/coderLMN/MasteringPredictiveAnalyticsWithR/issues，欢迎读者去提出问题、解答和建议，参与讨论。

另外需要说明的一点是，我在译稿中加入了不少译者注，目的是帮助读者理解某些比较晦涩的概念、公式和代码。但是因为个人水平有限，译者注里的解释和说明难免会有不严谨甚至错误的地方，请读者多多指正。

最后，我要感谢我的家人在本书翻译过程中的付出和耐心。因为有他们的支持，我才能以一种精耕细作的方式完成本书的翻译工作。

吴今朝

前　言

预测分析以及更一般意义上的数据科学当前正处于被追捧的热潮中，因为像垃圾邮件过滤、单词补全和推荐引擎这样的预测性技术已经被广泛运用于日常生活。这些技术现在不仅越来越被我们所熟悉，还赢得了我们的信任。在计算机处理能力和软件方面（例如 R 语言及其大量专用的扩展包）的发展产生了这样的局面：用户经过培训就可以使用这些工具，而无需具备统计学的高级学位，也不需要使用公司或大学实验室专用的硬件。技术的成熟度和基础软硬件的可用性结合起来，让很多该领域的从业者倍感兴奋，他们感到可以为自己的领域和业务设计一些能产生重要影响的工具，事实也确实如此。

与此同时，很多新进入该领域的人士很快发现其中有很多陷阱需要克服。实际上，没有哪个学位足以把一位学生或从业者训练为成功的预测建模者。该领域依赖于很多学科，例如计算机科学、数学和统计学。当前，进入该领域的人们不仅只在其中的一门学科有比较强的背景，还往往会比较专精于其他学科。在给研究生和从业者们讲授了有关本书材料的几次课程之后，我发现学员们反复表达的两个最大担忧是对编程和数学的恐惧。有意思的是，对这两者的表达几乎总是互斥的。预测分析学实际上是一种实践性的学科，但同时也是一种具备较强理论基础的学科，这些理论基础的知识对于从业者是很关键的。因此，掌握预测分析需要一系列不同的技能，从编写良好的软件到实现一种新技术或对数据进行预处理，再到理解某个模型的假设条件，如何有效地训练该模型，如何对该模型出现的问题进行诊断，以及如何调整模型的参数以获得更好的结果。

讨论到这里，很自然地会反向思考预测分析学作为一个领域实际会覆盖的内容。事实上，该领域和机器学习、数据挖掘、商业分析学、数据科学等其他相关领域的边界是比较模糊的。本书中会用到的定义非常宽泛。对于本书的主题而言，预测分析学是一个领域，它利用数据建立模型来预测未来我们感兴趣问题的结果。当然，它和机器学习领域会有很大的重叠，机器学习更多地研究从数据中学习的程序和算法。这种重叠的情况对于数据挖掘（以从数据中提取知识和模式为目标）也同样成立。数据科学正在迅速成为覆盖所有这些领域的综合术语，它还包括了其他主题，例如呈现数据分析结果的信息可视化，围绕在实际环境中部署模型的业务概念，以及数据管理。本书会着重于机器学习，但我们不会覆盖学习可行性的理论探索，也不会讲解着眼于从无特定预测目标的数据中寻找模式和聚类的无监督学习方法。取而代之，我们会探索像时间序列这样的一些主题，通常在机器学习的教材里不会讲解它们。

无论对于学习预测分析学还是解决实际环境中的问题，R 语言都是一个优秀的平台。它是一个开源项目，有一个持续快速增长的用户社区。在编写本书时，它和 Python 是全世界数据科学家最常用的两种语言。它有很多适用于不同建模技术和应用领域的扩展包，

其中很多可以通过连接到 Comprehensive R Archive Network (CRAN) 从 R 语言平台本身直接获取。该语言还有很多在线资源，从教程到在线课程都包含在内。我们尤其要提到优秀的交叉验证式论坛（http://stats.stackexchange.com/）以及 R-bloggers 网站（http://www.r-bloggers.com/），该网站包含了大量来自不同博客的关于 R 语言应用的文章。对于那些对 R 语言有点生疏的读者，我们提供了一个免费在线教程章节，它是从我们在 AUEB 学生的课程讲义演化而来的。

本书的主要任务是在（强调直觉及实践而不是理论的）低端入门教程和（专注于数学、细节和严谨性的）高端学术教材之间的鸿沟上架起桥梁。另一个同等重要的目标是给读者灌输一些良好的实践经验，比如学习如何适当地测试和评估一个模型。我们还要强调一些重要的概念，例如偏误 – 方差权衡[⊖]和过拟合，这些概念在预测建模中是普遍存在的，并会在不同模型中以多种形式反复出现。

从编程的角度来说，虽然我们假定你已经熟悉 R 语言，不过还是会详细解释并讨论每个代码示例，以便读者提高他们的自信心，循序渐进。尽管如此，在学习的过程中，或者至少在转到下一章之前，实际运行代码的重要性是如何强调都不为过的。为了尽可能让这个过程顺利进行，我们已经为教材中的所有章节提供了代码文件，其中包含了教材中所有的代码示例[⊜]。此外，我们还在很多地方编写了自己对于特定技术的简单实现方法。典型的两个示例是第 4 章里的口袋感知器算法和第 7 章的 AdaBoost 自适应增强方法。在某种程度上，这么做是为了鼓励用户学习如何编写他们自己的函数，而不是完全依赖于已有的实现方法，因为并不是所有方法都有现成的函数可用。

重现能力是数据分析的一项关键技能，而且它并不限于教育领域。因此，我们大量使用了可自由获取的数据集并尽力在需要随机数生成器的地方运用特定的种子值。最后，我们尽可能尝试利用相对小规模的数据集，以确保读者在阅读本书时运行代码不需要等待太长的时间或被迫寻求更好的硬件。我们要提醒你，在真实世界里，耐心是一种非常有益的美德，因为你感兴趣的大部分数据集会比我们学习本书时用到的更大。

每章的结尾是两个或多个实际的建模案例，每章的开始则是一些理解新模型或技术所必需的理论和背景知识。虽然不避讳用数学解释重要的细节，但是我们在这方面很慎重，相关的介绍适可而止，以确保读者能理解相关的基本概念就可以了。这样做符合本书的理念，即弥补入门教程和涉及更多细节的学术教材之间的差距。具备高中数学背景知识的读者可以确信，他们能够借助基本的数学知识完整地学习本书的所有内容。学习所需的关键技能是简单微积分（例如简单微分）、概率论的关键概念（例如均值、方差、相关系数），以及重要的概率分布（例如二项分布和正态分布）。虽然我们不提供这方面的教程，但在前面几章我们的确是循序渐进的。为了照顾那些数学爱好者的需求，我们经常会以提示的形式提供额外的技术细节，并给出一些参考资料作为所讨论内容的自然延伸。

有时候，我们需要给出某个概念的直观解释，以节省篇幅，避免另辟一章专门讨论不必要的纯理论。在这么做的时候（例如对第 4 章里的反向传播算法），我们会确保前后的

⊖ 即 bias-variance trade-off。在本书中，bias 都翻译为“偏误”而不是“偏差”，这是为了避免和第 3 章中出现的 deviance 一词产生混淆，本书中 deviance 一词会翻译为“偏差”，请读者注意这两个词的区分。——译者注

⊜ 本书配套代码的下载链接是：https://www.packtpub.com/code_download/17457/other。——译者注

衔接性，让读者能具备坚实的基础知识来进一步掌握更详细的内容。同时，还会给出精心挑选的参考文献，其中很多都是可读性好而且可以免费获取的文章、论文或在线教材。当然，我们会在任何必要的地方引用重要的教材。

本书没有练习题，但是鼓励你把好奇心发挥到极致。好奇心对于预测建模者来说是一种巨大的天赋。我们从中获取了分析数据的很多网站上都有我们没有研究到的其他大量数据集。我们偶尔还会讲解如何创建人工数据来演示某个特定技术背后的概念验证过程。很多用来创建和训练模型的 R 语言函数都有一些其他调优参数是本书中没有时间讲解的。用到的扩展包也往往会包含和我们讲解的函数相关的其他函数，正如用到的扩展包本身也往往会有其他替代包可用。所有这些都是进一步研究和实验的途径。要掌握预测分析学，认真学习和个人的探索及练习都是同等重要的。

学生在该领域的一个普遍诉求，是用额外的实例来模拟有经验的建模者针对数据集所遵循的实际过程。在现实中，可信的模拟过程从分析开始后可能会持续很多小时。这是因为花在预测建模上的大部分时间都用来研究数据、尝试新特征和预处理步骤，以及对结果试验不同的模型。简而言之，正如我们在第 1 章中将要看到的，探索、试验和误差是有效分析的关键组成部分。编写一本讲解关于每个数据集的错误或不成功方案的书是完全不现实的。相反，强烈推荐读者将本书中的所有数据分析过程视为改进的起点，并自己延续这个过程。好的思路是尝试把其他章节讲解的技术运用于特定数据集，以便观察其他方法是否有效。从简单地给某个输入特征运用不同的变换方式到采用另一章里讲解的完全不同的模型，任何尝试都是可以的。

作为最后一个提示，我们要指出，创建美观规范的图来呈现数据分析结果是一项重要的技能，尤其是在职场中。虽然 R 语言的基础绘图能力覆盖了基本的需求，但它往往缺乏美观性。因此，除了用分析代码中的某些函数产生的特定图形之外，我们会用 ggplot2 包绘图。虽然我们不提供这方面的教程，但是本书中包括的所有产生绘图的代码都在配套的代码文件里，希望用户可以从中受益。ggplot2 包的一个很有用的在线参考资料是“the Cookbook for R”（http://www.cookbook-r.com/Graphs）网站上有关图形的章节。

本书内容

第 1 章会讲解统计模型的通用语言和在对这些模型进行分类时所依据的一些重要差别，由此开启我们的学习之旅。本章的亮点是对预测建模过程的探索，我们会通过它展示第一个模型，即 k 近邻（k Nearest Neighbor，kNN）模型。

第 2 章会介绍预测数量值最简单且最著名的方法。本章的重点是理解线性回归的假设，以及一些可以用来评估训练模型质量的诊断工具。此外，本章还会涉及正则化的重要概念，它可以用于避免预测模型常见的一种瑕疵——过拟合（over fitting）。

第 3 章会对前一章里线性模型的思想进行扩展，方法是引入广义线性模型的概念。虽然这类模型有很多示例，但本章的重点是逻辑回归这样一个针对分类问题的流行方法。我们还会探讨该模型扩展到针对多类别的情况，发现该方法对于二元分类（binary classification）的效果最好。

第 4 章会讲解能够处理回归及分类两种任务的一种仿生模型。神经网络有很多种，而本章会重点关注多层感知网络（multilayer perceptron network）。神经网络是复杂的模型，

本章的主要关注点是理解在训练过程中起作用的一组不同的配置和优化参数。

第 5 章会通过学习支持向量机来掌握非线性模型的问题。在这部分，我们会通过利用最大边缘分离（maximum margin separation）来尝试以几何方式拟合我们的训练数据，以探索对分类问题进行思考的另一种方法。本章还会介绍交叉验证（cross-validation）这一评估和优化模型的基本技术。

第 6 章讲解决策树（decision tree）这样另一类已经成功运用于回归和分类等问题的模型。决策树的类型有很多种，本章会介绍一批不同的训练算法，例如 CART 和 C5.0。我们还可以看到，树形方法具有独特的优点，例如内建的特征选择、支持缺失数据和类别变量，以及非常易于解释的输出。

第 7 章不同于以往章节，本章会采用一种迂回方式，不会讲解新类型的模型，而是尝试解答如何把不同模型有效地结合起来的问题。我们会讲解装袋（bagging）和增强（boosting）这两种著名的技术，并会把随机森林（random forest）作为树的装袋的一种特例进行介绍。

第 8 章讲解的是机器学习研究领域的一个活跃领域，即概率图模型。这些模型通过一个图结构把变量之间的条件性独立关系（conditional independence relation）进行编码，已经成功运用于从计算机视觉到医疗诊断等很多领域的问题中。本章会学习它的两种主要表现形式，即朴素贝叶斯（Naïve Bayes）模型和隐马尔可夫（hidden Markov）模型。尤其是，后一种模型已经成功运用于序列预测（sequence prediction）问题，例如基因测序的预测和通过词性标注对句子进行标记。

第 9 章研究的是对特定的时间过程建模的问题。它的一个典型应用是根据一段时间内的原油价格历史数据来预测原油的远期价格。对时间序列建模有很多不同的方式，而本章的重点是 ARIMA 模型，也会讨论一些替代方案。

第 10 章在本书中的独特之处是它会讲解主题建模，这是一种源于聚类和无监督学习的方法。不过，我们会在一个预测建模场景下学习如何运用这种重要的方法。本章会强调主题建模里最著名的方法，即隐含狄式分布（Latent Dirichlet Allocation，LDA）。

第 11 章会通过讨论推荐系统对全书的内容进行总结。推荐系统会分析和商品进行交互的用户的喜好，从而做出推荐。它的一个著名示例是 Netflix，它利用一个用户对所观看电影进行评分的数据库来进行电影的推荐。本章会重点讲解协同过滤（collaborative filtering），这是一种产生推荐内容的纯数据驱动方法。

《Introduction to R》给出了对 R 语言的介绍和概述。它是为了让用户能快速入门以便理解本书中的代码样例而提供的。它作为本书的在线章节，可以从 https://www.packtpub.com/sites/default/files/downloads/Mastering_Predictive_Analytics_with_R_Chapter.pdf）访问。

阅读准备

运行本书代码的唯一硬性要求就是安装 R。它可以从 http://www.r-project.org/ 自由获取，并可以在所有主流操作系统上运行。本书中的代码已经在 R 的 3.1.3 版本上测试过。

所有的章节都至少会引入一个 R 基础安装包里没有预装的新扩展包。我们不会在教材中专门讲解安装 R 包的过程，但是如果某个包当前没有安装在你的系统里，或它需要更

新，你就可以利用 install.packages() 函数来安装它。例如，下面的命令会安装 tm 包：

```
> install.packages("tm")
```

我们使用的所有包都可以在 CRAN 上获取。下载和安装它们以及获取我们在实例中使用的开源数据集都需要 Internet 连接。最后，虽然不是绝对必需，但我们要推荐你养成利用集成开发环境（Integrated Development Environment，IDE）进行 R 语言编程的习惯。有个很棒的 IDE 是 RStudio (http://www.rstudio.com/)，它是开源的。

读者人群

本书是为预测建模相关领域的从业者中的新秀和老手编写的。本书大部分的内容已经在研究生、专业人士和 R 语言培训的授课中使用过，因此它在策划的时候就已经把这些学员的情况都考虑到了。读者必须熟悉 R 语言，不过即便是那些从来没有接触过这种语言的人，也能够通过阅读在线教程的章节掌握必要的背景知识。不熟悉 R 语言的读者至少要接触过某些编程语言，例如 Python。那些具备 MATLAB 背景的人会发现切换到 R 语言相当容易。正如之前提到的，本书对数学的要求是非常适度的，只需要中学数学的某些元素，例如均值和方差的概念及基础的微分。

本书约定

在本书中，你会看到用来区分不同类型信息的多种文本样式。这里是一些关于这些样式的示例及其含义的解释。

代码段的样式设置如下：

```
> iris_cor <- cor(iris_numeric)
> findCorrelation(iris_cor)
[1] 3
> findCorrelation(iris_cor, cutoff = 0.99)
integer(0)
> findCorrelation(iris_cor, cutoff = 0.80)
[1] 3 4
```

新术语和重要的关键字会用**粗体**显示。

警告或重要的注解会出现在像这样的一个方框内部。

提示和小技巧的样式是这样的。

下载样例代码

你可以从 http://www.packtpub.com 通过个人账号下载你所购买的所有 Packt 书籍的样例代码文件。如果你从其他地方购买了本书，你可以访问 http://www.packtpub.com/support 并完成账号注册，以便直接通过电子邮件获得相关文件。

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，下载本书中的源代码。

致谢

每段伟大的探险背后都有一个精彩的故事，本书的写作也不例外。正是因为有了很多人的贡献本书才得以面世。我要感谢我在 AUEB 教过的很多学生，他们的投入和支持简直是铺天盖地的。可以确信的一点是，我从他们那里学到的东西和他们从我这里学到的一样多，甚至更多。我还要感谢 Damianos Chatziantoniou 策划在希腊开设领先的研究生数据科学课程。Workable 公司是一个大熔炉，在这里我能和才华横溢且激情澎湃的工程师们并肩工作，从事有益于全球商业的激动人心的数据科学项目。为此，我要感谢我的同事们，特别是两位点石成金的公司创始人 Nick 和 Spyros。

我要感谢 Subho、Govindan、Edwin 和 Packt 出版社所有同仁的专业和耐心。我要对很多给予我鼓励和激励的朋友表达我永恒的感谢。我的家人和亲友们在本项目中给予了我不可思议的支持。特别地，我要感谢我的父亲 Libanio，他鼓励我从事科学事业，还有我的母亲 Marianthi，她对我的信心一直远远超过其他任何人。感谢我的妻子 Despoina 耐心而坚定地站在我一边，即使这本书让我在她首次怀孕的时候难以陪伴在旁。最后也同样重要的是，在本书写作的收尾阶段，我的小女儿在我身边安睡并天真无邪地守望着我。她帮助我的方式是无法用语言描述的。

目　录

第 1 章
准备预测建模

在第 1 章，我们首先要学习的内容是掌握模型的通用语言，并深入了解预测建模的过程。预测建模的很多方面会涉及统计学和机器学习的关键概念，本章会对这些领域的核心特征进行简短的介绍，它是预测建模者所需的基础知识。我们会重点强调的一个重要知识点是如何对适用于我们要解决的问题类型的模型进行评估。最后，我们会讲解第一种模型，即 k 近邻（k-nearest neighbor）模型，以及对预测建模者非常有用的一个 R 语言包 caret。

1.1 模型

模型是预测分析学的核心，因此，本书一开始会讨论各种模型及其形式。简而言之，模型是我们要理解和分析的状态、流程或系统的一种表现形式。我们创建模型的目的是根据它得出推论以及（在本书中对我们更为重要的一点）对世界进行预测。模型的格式和风格有很多种，我们在本书中会探讨这种多样性中的一部分。模型可以是和我们能够观察或测量的数量值相关的一些方程，也可以是一套规则。我们大部分人在学校都熟悉的一个简单模型是牛顿第二运动定律。该定律表明，一个物体受到的合力会使之在合力作用的方向加速，加速度和合力大小成正比，和物体的质量成反比。

我们通常会用一个以字母 F、m 和 a 代表模型包含的量的方程总结这个规则。我们还利用大写希腊字母 sigma（Σ）来表示要对受力求和，用字母上的箭头表示它们是向量（就是既有大小也有方向的量）：

$$\Sigma \vec{F} = m\vec{a}$$

这个简单却强大的模型让我们能够对世界进行某些预测。例如，如果对已知质量的物体施加已知的作用力，我们就可以利用这个模型来预测它的加速度。如同大部分模型一样，该模型也会作出某些假设和概括。例如，它会假设物体的颜色、它所处环境的温度，以及它所处空间的准确坐标都和模型中列出的三个数值的相互影响无关。这样，模型就可以把待分析的流程或系统中的特定实例（在这个示例中，就是我们对其运动状况感兴趣的物体）的各种细节进行抽象化，从而把我们的关注点限制在重要的性质上。

牛顿第二运动定律并不是描述物体运动的唯一模型。物理学专业的学生很快会发现其他更复杂的模型，例如那些把相对论质量考虑进去的模型。总体而言，如果模型需要考虑更多数量的数值或具有更复杂的结构，它们往往就会更复杂。例如，非线性模型一般会比线性模型更复杂。在实际工作中，要决定采用哪个模型，并不仅仅是选择更复杂的模型而不是相对

简单的模型那么简单。实际上，我们在本书中逐步讲解很多不同模型的过程中，会把这个问题作为核心主题反复进行讨论。为了直观地理解其中的道理，可以考虑一下这个案例：我们用来测量物体质量和施加作用力的设备是非常粗糙的。在这样的情况下，寄希望于采用更复杂的模型也许就没有什么意义，因为我们知道，由于输入中的这种噪声，这样的预测所额外增加的精确度对结果产生不了什么差别。我们需要采用更简单模型的另一种情况是，在应用中我们根本不需要更高的精确度。第三种情况是更复杂的模型涉及我们无法测量的数值。最后，如果因为复杂度的关系，模型会需要太长时间进行训练或预测，我们也不会使用更复杂的模型。

1.1.1 从数据中学习

在本书中，我们要学习的模型具有两种重要和本质的特征。第一个特征是我们不会用数学推理或逻辑归纳的手段从已知事实产生模型，也不会根据技术规范或商业规则来构建模型；相反，预测分析学领域是根据数据来构建模型的。更具体地说，我们会假设，对于要完成的任何具体预测任务，我们会把该任务以某种方式关联或以衍生出来的某些数据作为起点。例如，如果要创建一个预测某国各地全年降雨量的模型，我们也许已经收集了（或具备了收集的手段）关于不同地点降雨量的数据，同时会测量一些我们会关心的数值（例如海拔高度、经度和纬度）。我们创建的这个模型之所以有能力执行预测任务，是因为我们可以利用一组有限地点的降雨量测量数据的样例来预测我们没有收集过任何数据的地点的降雨量。

我们建立模型时所针对的问题的第二个重要特征是，在根据某些数据创建模型来描述特定现象的过程中，我们必然会遇到某些随机性的来源。我们称之为模型的**随机成分**（stochastic）或**不确定成分**（nondeterministic component）。有时候，我们要尝试建模的系统本身并不具有任何内在的随机性，而是数据包含了随机成分。数据中的随机来源的一个范例是在测量时对温度之类的数量进行读取时所产生的误差。不包含内在随机成分的模型称为**确定型模型**（deterministic model），比如牛顿第二定律就是这类模型的一个范例。随机模型则是假设在建模的过程中含有内在的随机源的模型。有时候，这种随机性的来源是来自这样一个事实：对最有可能影响系统的所有变量都进行测量是不可能的，因而我们只能利用概率来对其进行建模。纯粹随机模型的一个著名示例是扔一个六面无偏误的骰子。回顾一下，在概率中，我们使用**随机变量**（random variable）这个术语来描述某个实验或随机过程的特定结果值。在扔骰子示例中，我们可以定义随机变量 Y 作为每次扔骰子之后朝上的那一面的点数，由此产生了下面的模型：

$$P(Y=y)=\frac{1}{6}, y \in \{1, 2, 3, 4, 5, 6\}$$

该模型告诉我们，扔到一个特定数字（比如 3）的概率是六分之一。注意，我们没有对具体扔骰子的结果进行明确的预测，相反，我们表明的是，每种结果出现的可能性是均等的。

概率是在日常交流中经常会用到，但有时也导致其准确含义产生混淆的一个术语。其实，对概率有很多不同的理解方式。两种常常被援引的解释方式是**频率论概率**（Frequentist probability）和**贝叶斯概率**（Bayesian probability）。频率论概率是和重复试验相关的，例如扔单面骰子。在这种情况下，如果该实验重复了无穷次，那么看到数字 3 的概率就是数字 3 出现的相对比例。贝叶斯概率则是和在看到特定结果时

主观上置信或意外的程度相关的，因此可以用于对一次性事件赋予含义，例如某位总统候选人赢得大选的概率。在我们的扔骰子试验里，我们看到数字 3 时的意外程度和看到其他数字时是一样的。注意，在这两种情况下，我们讨论的仍然是同一个数字概率（1/6），而只是解释不同而已。

在扔骰子模型的情况下，我们没有任何变量需要测量。不过，在大部分情况下，我们会看到包含了一批需要测量的自变量的模型，而这些模型会用来对某个因变量进行预测。预测性建模依赖于很多不同的领域，因此，根据你所学习的具体学科，你会发现它关联了不同的一些领域。在展开讲解这个问题之前，让我们把一个数据集加载到 R 语言里。R 语言预装了很多常被引用的数据集，我们要挑出其中可能最著名的一个，即鸢尾花数据集（iris data set）。

要查看 R 还绑定了哪些其他的数据集，我们可以利用 data() 命令获取一个数据集清单及每个数据集的简短描述。如果要修改一个数据集里的数据，我们可以把待分析的数据集名作为输入参数提供给 data() 命令，例如，data(iris) 会重载 iris 数据集。

```
> head(iris, n=3)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
```

iris 数据集由鸢尾花的 3 个不同品种的共 150 个样本的测量数据组成。在前面的代码中，我们可以看到对每个样本产生了 4 种测量值，分别是花瓣（petal）和萼片（sepal）的长度和宽度。iris 数据集常被用来作为一个典型的标杆，用于评估根据前文中 4 种测量数据来预测鸢尾花样本品种的不同模型。花瓣长、花瓣宽、萼片长、萼片宽合在一起，在文献中会被称为**特征**（feature）、**属性**（attribute）、**预测因子**（predictor）、**维度**（dimension）或**自变量**（independent variable）。在本书中，我们会优先使用特征这个词，但其他术语也同样是可以用的。类似地，该数据框中的品种（species）列是我们要利用模型来进行预测的，因此它被称为**因变量**（dependent variable）、**输出**（output）或**目标**（target）。同理，为了保持一致性，在本书中我们会优先使用其中之一，即输出。该数据框的每一行对应单个数据点，被称为一条**观测数据**（observation），不过它通常会包括对一组特征的观测值。

由于我们将要用到一些数据集（比如前面讲解的 iris 数据集）来构建预测模型，因此设定一些符号惯例会大有帮助。这里要讲解的惯例在大部分文献里很常见。我们会用大写字母 Y 来代表输出变量，带下标的大写字母 X_i 来表示第 i 个特征。例如，在我们的 iris 数据集里，我们有 4 个特征，分别表示为 X_1 到 X_4。我们会用小写字母表示单条观测数据，因此 x_1 就对应第一条观测数据。注意，x_1 本身是由特征分量 x_{ij} 组成的一个向量，因此 x_{12} 就代表第一条观测数据中的第二个特征值。为简单起见，我们会尽量谨慎使用双后缀，也不会使用箭头或其他向量标记的形式。通常情况下，我们会讨论观测数据或特征两者之一，因此变量的大小写就能让读者明确到底它指的是这两种情况中的哪一个。

在考虑对应某个数据集的预测模型时，我们通常会假设，对于具有 n 个特征的模型，有一个真实或理想的函数 f 可以把特征映射到输出：

$$Y=f(X_1, X_2, \cdots, X_n)$$