Broadview
www.broadview.com.cn

Data Science

数据科学

方匡南 编著

配套案例源码和教学PPT课件

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据科学

方匡南　编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书是一本数据科学的入门书籍。每个知识点尽量从实际的应用案例出发，从数据出发，以问题为导向，在解决问题中学习数据挖掘、机器学习等数据科学相关方法。

本书将数据读/写、数据清洗和预处理作为开端，逐渐深入介绍和数据科学相关的决策树、支持向量机、神经网络、无监督学习等知识。此外，结合数据科学的实际应用，书中还讲解了推荐算法、文本挖掘和社交网络分析等热门实用技术。

本书在写作过程中尽量删去太过抽样的理论，让具有一定高等数学和概率论基础的读者就能看得懂。当然，如果读者对方法原理确实不感兴趣，只是为了用 R 语言程序实现某种方法，可以跳过方法只看案例和程序。

本书适合作为高校数据科学、机器学习、数据挖掘、大数据分析等相关专业的研究生和高年级本科的教科书，也适合作为相关企业的数据科学家、数据挖掘工程师、数据分析师及数据科学的爱好者等的工具书。

图书在版编目（CIP）数据

数据科学 / 方匡南编著. —北京：电子工业出版社，2018.6
ISBN 978-7-121-34244-8

Ⅰ. ①数… Ⅱ. ①方… Ⅲ. ①数据处理—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 106121 号

策划编辑：张月萍
责任编辑：张　慧　　文字编辑：苏颖杰
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编　100036
开　　本：787×980　1/16　印张：20.75　　字数：531 千字
版　　次：2018 年 6 月第 1 版
印　　次：2018 年 6 月第 1 次印刷
印　　数：1～3000 册　　定价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn

前　言

数据科学（Data Science）是一门交叉学科，是一门分析和挖掘数据并从中提取规律和利用数据学习知识的学科，包含了统计、机器学习、数据可视化、高性能计算等。近几年，大数据的发展如火如荼，与此同时，“数据科学家”这个词也跟着火起来，成为职场中的香饽饽。正如谷歌首席经济学家哈尔·瓦里恩（Hal Varian）于 2009 年在纽约时报撰文所说，“未来十年最性感的工作将是统计学家”，这里的统计学家是广义的统计学家，包括数据科学家。数据科学家职业被招聘网站 Glassdoor 在 2016 年评选为美国最佳工作。德勤（Deloitte）公司预测 2018 年全球企业将至少需要 100 万名数据科学家，大学培养的数据科学家数量远远不能满足市场需求，按照目前数据科学家的培养数量来看，这个缺口是很大的。我国真正的数据科学家人才是比较短缺的。数据科学家需要有较好的统计学、机器学习功底，能够理解模型背后的原理和算法，具备熟练的编程能力并熟悉业务知识。

数据科学主要由两拨人在做：一拨人在计算机圈子里，主要关注处理海量数据的能力、速度和算法；另一拨人在统计圈子里，更多地关注模型本身的精度和可解释性。市面上有各种各样讲解大数据、数据科学的书籍，但多数是讲解一些理念，或者只讲解一些抽象原理和算法，很少从数据到模型的角度去讲解，缺少真正能够将数据科学与实务操作结合起来的书籍。我觉得自己有责任写一本关于数据科学方面的教材，来帮助数据科学的初学者更快地掌握模型原理和实务操作。

我每年都在厦门大学开设数据挖掘的课程，在课程资料的基础上慢慢整理出本书稿，总体框架借鉴斯坦福大学统计系几位学者出版的两本经典统计学教材，即 Jamnes、Witten、Hastie 和 Tibshirani 写的 *An Introduction to Statistical Learning* 和 Hastie、Tibshirani 和 Friedman 写的 *The Elements of Statistical Learning*。后来，我受邀在北京、上海等地开设暑期数据挖掘现场公开课，前来听课的学生有国外著名高校的教师、研究生，国内高校的教师、研究生，医药、金融等公司的数据分析人员、数据挖掘分析师等。他们对我的讲义提出了很多有用的建议，经过不断地完善，最终形成了此书。

通过在很多地方上公开课，并与很多不同领域的学者交流，我深刻地体会到统计或数据挖掘方法的应用范围越来越广，借用马克思的话，“一种科学只有在成功地运用数学时，才算达到了真正完善的地步”，也可以说“一个学科使用、分析数据的程度可以反映出这个学科的发展程度”。

本书是一本数据科学的入门教材，内容循序渐进、深入浅出，每个知识点都根据实际的应用案例从数据出发，以问题为导向，使读者在解决问题的过程中学习数据挖掘、机器学习

等数据科学相关方法。本书既可作为高校数据科学、机器学习、数据挖掘、大数据分析等相关专业的研究生和高年级本科的教科书，也可作为相关企业的数据科学家、数据挖掘工程师、数据分析师及数据科学爱好者等的工具书。本书为读者提供方法和程序上的参考，在写作过程中尽量删除过于抽象的理论原理，让具有一定高等数学和概率论基础的读者都能看得懂。当然，如果读者对方法原理确实不感兴趣，只是为了用 R 语言程序实现某种方法，或者分析某些有意义的数据，则可以跳过方法，只看案例和程序。

我的博士和硕士研究生陈子岚、王小燕、赵梦峦、范新妍、张晓晨、林颖、赵雪、张喆参与了资料收集、案例编写等工作，陈子岚参与了全书的校对、修改、排版等工作，在此一并感谢！感谢成都道然科技有限责任公司的专业意见和建议。再次感谢为本书提供直接或者间接帮助的各位朋友，没有他们的帮助，本书的出版没有这么顺利。

为了方便读者使用，我的团队为本书开发了一个 R 语言包 RDS。RDS 包和本书案例相应的代码可以从网址 http://www.kuangnanfang.com/?id=7 或 https://github.com/ruiqwy 下载。另外，由于篇幅限制，团队制作的一些经典案例无法在本书中展示，在以上网址也提供了部分经典案例。

在本书编写过程中，我深刻地体会到写书是一件“苦差事”，仔细较真，总能发现有很多值得完善的地方，这也是本书拖了 3 年才得以出版的原因。我希望此书尽可能以“完美”的形象与读者见面，但由于本人水平和精力有限，书中难免有错误或不足之处，恳请广大读者批评指正！

方匡南

2018 年 3 月于厦门大学

目　录

第1章 导 论

1.1 数据科学的发展历史

统计学作为一门学科已有三百多年的历史。按照统计方法及历史的演变顺序，通常可以将统计学的发展史分为三个阶段，分别是古典统计学时期、近代统计学时期和现代统计学时期。古典统计学的萌芽最早可以追溯到17世纪中叶，此时的欧洲正处于封建社会解体和资本主义兴起的阶段，工业、手工业快速增长，社会经历着重大变革。政治改革家们急需辅助国家经营和管理的数据证据以适应经济发展需要，此时，一系列统计学的奠基工作在欧洲各国相继展开。在这一时期，以威廉·配第和约翰·格朗特为代表的政治算术学派与海尔曼·康令（Hermann Conring）创立的国势学派相互渗透和借鉴，服务与指导了国家管理和社会福利改善。

18世纪末至19世纪末为近代统计学发展时期。在这一百年间，欧洲各国先后完成了工业革命，科学技术开始进入全面繁荣时期，天文、气象、社会人口等领域的数据资料达到一定规模的积累，对统计的需求已从国家层面扩展至社会科学各个领域。对事物现象静态性的描述已不能满足社会需求，数理统计学派创始人凯特勒（A.J.Quetelet）率先将概率论引入古典统计学，提出了大数定律思想，使统计学逐步成为揭示事物内在规律，可用于任何科学的一般性研究方法。一些重要的统计概念也在这一时期提出，误差测定、正态分布曲线、最小二乘法、大数定律等理论方法的大量运用为社会、经济、人口、法律等领域的研究提供了大量宝贵的指导。

20世纪科学技术的发展速度远远超过之前的时代，以描述性方法为核心的近代统计学已无法满足需求，统计学的重心转为推断性统计，进入了现代统计学阶段。随着20世纪初细胞学的发展，农业育种工作全面展开。1923年，英国著名统计学家费雪（R.A.Fisher）为满足农作物育种的研究需求，提出了基于概率论和数理统计的随机试验设计技术，以及方差分析等一系列推断统计理论和方法。推断性统计方法的进步对工农业生产和科学研究起到了极大的促进作用。

自 20 世纪 30 年代，随着社会经济的发展和医学先进理念的吸收融合，人们对于医疗保险和健康管理的需求日益增长，统计思想渗透到医学领域形成了现代医学统计方法。例如，在生存质量（Quality of life）研究领域，通过分析横向和纵向资料，逐步形成了重复测量资料的方差分析、质量调整生存年（QALYs）法等统计方法。这一阶段，统计学在毒理学、分子生物学、临床试验等生物医学领域获得了大量应用，这些领域的发展又带动统计方法不断创新，主成分估计、非参数估计、MME 算法等方法应运而生。

20 世纪 80 年代开始，随着现代生物医学的发展以及计算机技术的进步，人类对健康的管理和疾病的治疗已进入基因领域，对基因数据分析产生了大量需求。高维海量的基因数据具有全新的数据特征，变量维度远远大于样本数，传统的统计方法失效了。因此，一系列面向高维数据的统计分析方法相继产生，如著名的 Lasso 方法。

20 世纪 90 年代以来，随着互联网的发展，数据库中积累了海量的数据，如何从海量的数据中挖掘有用的信息就变得越来越重要了，数据挖掘（Data Mining）也就应运而生了。数据挖掘又称数据库中的知识发现（Knowledge Discover in Database，KDD），是目前人工智能（Artificial Intelligence）和数据库领域研究的热点问题。所谓数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。与数据挖掘比较接近的名词是机器学习（Machine Learning），机器学习被看作人工智能的一个分支，主要研究一些让计算机可以自动“学习”的算法，是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为机器学习算法中涉及了很多统计学理论，与统计学的关系密切，也称为统计学习（Statistical Learning）。

随着计算机技术、互联网等的普及与飞速发展，人类社会被呈爆炸性增长的信息所包围。据 IBM 公司资料显示，目前数据的生成每日以千万亿字节来计算，如 2017 年全球每秒发出的 E-mail 超过 1000 万封，淘宝每天产生的数据超过 50TB，Facebook 每个月更新的照片超过 10 亿张，全球数据量每年以 40%左右的速度增长。麦肯锡全球研究院（MGI）在 2011 年首次提出了大数据时代（Age of Big Data）概念。依照美国麦肯锡（McKinsey）咨询公司的定义，大数据是指那些规模超出了典型数据库软件工具能力的进行捕获、存储、管理和分析的数据集。与传统数据相比，大数据的大不仅是指体量上的扩充，而且是指数据的结构、形式、粒度、组织等各方面都更加复杂。不过，我们认为大数据并不是从方法论角度提出的，研究大数据的方法主要是数据挖掘和机器学习方法。

近几年，数据科学（Data Science）的概念被提出，这是一门分析和挖掘数据并从中提取规律和利用数据学习知识的学科，因此其概念也更广，包含了统计、机器学习、数据可视化、高性能计算等。近几年，数据科学家这个词也跟着火起来，成为职场中的“香饽饽”。德勤公司预测 2018 年全球企业将至少需要 100 万名数据科学家，大学培养的数据科学家数量远远不能满足市场需求，按照目前数据科学家的培养数量来看，这个缺口是很大的。我国真正的数据科学家人才是比较短缺的。数据科学家需要有较好的统计学、机器学习功底，能够理解模型背后的原理和算法，具备熟练的编程能力并熟悉业务知识。

1.2 数据科学研究的主要问题

数据科学研究的问题比较广泛，甚至可以说，只要和数据收集、清洗整理、分析和挖掘有关的问题都是数据科学要研究的问题。数据科学研究的问题，应该是从实际业务需求中提炼出来的问题。下面通过举几个例子来讲解数据科学研究的主要问题。

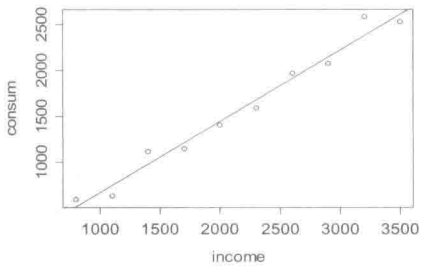

图 1-1 消费与收入的散点图

例 1.1 家庭收入与消费支出。为了研究某社区家庭月消费支出与家庭月可支配收入之间的关系，随机抽取并调查了 12 户家庭的相关数据，如图 1-1 所示。通过调查所得的样本数据能否发现家庭消费支出与家庭可支配收入之间的数量关系，以及如果知道了家庭的月可支配收入，能否预测家庭的月消费支出水平呢？

例 1.2 消费贷公司对客户的信用评分。客户在申请消费贷时，公司收到客户的收入、工作年限、职业等数据，以及从其他渠道获取的数据，消费贷公司需要评估客户的信用评分，以便决定是否给予核准贷款。那么，该如何预测客户借钱后是否会违约？该如何给每位客户评分？

例 1.3 员工离职预测。一定的员工流动率能够为企业注入新鲜的活力，增强组织的创新能力，但过多的员工离职，特别是核心员工的离职则会导致企业人力资本投资的损失，员工士气低落，破坏企业建立的竞争优势等消极影响，甚至对社会稳定也会造成一定的威胁。因此，通过对离职影响因素的分析，企业管理者可以有效地对员工的离职行为进行管理。例如，通过收集员工满意度、绩效评估、完成的项目数量、每月工作时数、工作年数等因素（如图 1-2 所示），如何预测员工是否离职，以便提前做好准备？

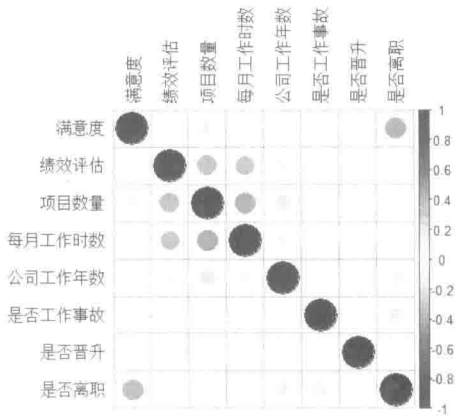

图 1-2 是否离职与其他因素的相关系数

例 1.4　前列腺影响因素分析。研究对象是前列腺根治手术的病人，分析他们的前列腺特殊抗原水平（lpsa）与 8 个临床指标之间的相关性（如图 1-3 所示）。这 8 个临床指标包含肿瘤体积（lcavol）、前列腺重量（lweight）、年龄（age）、良性前列腺增生量（lbph）、精囊浸润（svi）、包膜穿透（lcp）、格里森评分（gleason）和格里森评分 4 或 5 百分比（pgg45）。其中，svi 是二元变量，gleason 是分类变量。如何从 8 个临床指标中筛选出与前列腺特殊抗原水平有关的影响因素？

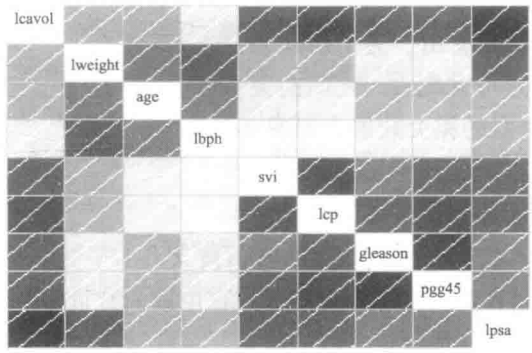

图 1-3　前列腺特殊抗原水平（lpsa）与影响因素的相关系数

例 1.5　购物篮分析。表 1-1 是某超市顾客购买记录的数据库 D，包含 6 个事务 $T_k(k=1,2,\cdots,6)$，其中项集 I ={面包，牛奶，果酱，麦片}。现在要分析已购买面包的顾客，有多大概率会买牛奶？如何根据顾客过去的购买记录，推荐其感兴趣的商品？

表 1-1　某超市顾客购物记录数据库 D

TID	Date	Items
T100	6/6/2010	{面包，麦片}
T200	6/8/2010	{面包，牛奶，果酱}
T300	6/10/2010	{面包，牛奶，麦片}
T400	6/13/2010	{面包，牛奶}
T500	6/14/2010	{牛奶，麦片}
T600	6/15/2010	{面包，牛奶，果酱，麦片}

例 1.6　犯罪率水平分析。收集美国 50 个州的犯罪率相关数据，包含 50 个观测值和 4 个变量，其中 Murder、Assault、Rape 3 个变量分别为每 10 万居民中被逮捕的谋杀、暴力和强奸犯罪人数，UrbanPop 表示各州城市人口比例。我们想研究的问题是如何用一两个综合的变量来总结这些信息，并对各州犯罪率水平进行评价？

例 1.7　花卉细分。测量 18 种花卉的 8 个指标，这 8 个指标包括是否能过冬、是否生长在阴暗的地方、是否有块茎、花卉颜色、所生长泥土、某人对这 18 种花卉的偏好选择、花卉

高度、花卉之间所需的距离间隔等。如何根据这 8 个指标对 18 种花卉进行细分？该分为几类比较合适？

例 1.8 文本挖掘。从网上收集 20000 多篇关于房地产的相关新闻，如何分析这 20000 多篇新闻里都在讨论哪几个主要话题？如何有效地把这些新闻聚为几类？如何提取新闻的情感倾向并编制成指数？

1.3 数据科学的主要方法

从方法来看，数据科学的方法主要还是统计学或者机器学习里的方法。机器学习方法往往可分为有监督学习（supervised learning）、无监督学习（unsupervised learning）及半监督学习（semi-supervised learning），详见图 1-4。

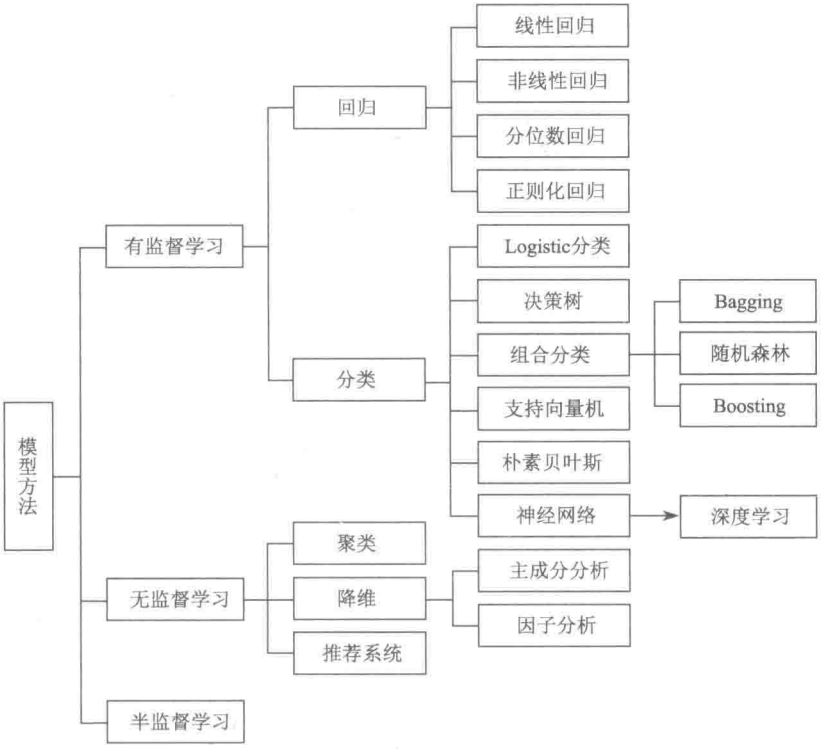

图 1-4 数据科学的模型方法

有监督学习是指在建模时，对每个（某些）自变量 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$（向量默认用列表示），$i=1,2,\cdots,n$，都有对应的因变量 $\boldsymbol{y}_i$。模型学习的好坏，可以由因变量的实际观察值评判，一个好的模型对因变量的预测值要尽可能接近其对应的真实观察值。

另外，根据因变量 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 取连续值或离散值，有监督学习又分为回归（regression）

和分类（classification）两大类问题。当因变量取连续值时，我们称之为回归。回归分析（regression analysis）是研究一个变量关于另一个（些）变量的具体依赖关系的计算方法和理论。通过后者的已知或设定值，去估计和（或）预测前者的（总体）均值。可以表示为

$$y = f(X) + \varepsilon$$

其中，y是因变量向量，$X = (X_1, X_2, \cdots, X_p)$是含有$p$个自变量的矩阵，$f$是关于$X$的函数，$f$的形式可以是已知的（如最简单的就是线性回归，即$y = X\beta + \varepsilon$），我们称这类方法为参数回归；$f$也可以是未知的，此时就需要根据数据去估计$f$，我们称这类方法为非参数回归。$\varepsilon$是随机误差项（error term），这部分是无法预测的。

建模时，我们要拟合一个比较合理的$\hat{f}$去估计f，当给定了X，就可以得到$\hat{y} = \hat{f}(X)$。由于变量间关系的随机性，回归分析关心的是根据X的给定值，考察y的总体均值$E(y|X)$，即当解释变量取某个确定值时，被解释变量所有可能出现的对应值的平均值。

不同的学科或者不同的教材对y和X有不同的术语，我们把这些术语整理归纳，以免读者产生混淆。通常，y称为被解释变量（Explained Variable）或因变量（Dependent Variable）或响应变量（Response）；X称为解释变量（Explanatory Variable）或自变量（Independent Variable）或者协变量（Covariate）。在数据挖掘和机器学习里，往往把模型f看作“机器（Machine）”或者“箱子（Box）”。因此，往往又更加形象地将y称为输出变量（Output Variable），将X称为输入变量（Input Variable）。即输入变量X丢入“机器”（“箱子”）里，输出变量y就被输出来。所以，当模型f比较简单且容易理解时，往往也称“白箱子（White Box）”，而当f比较复杂且难以理解时，也就相应地称为“黑箱子（Black Box）”。

当因变量取离散值时，称为分类。例如，我们在信用卡违约预测时，因变量y取值是{违约，不违约}，这是一个二元（binary）的取值。模仿回归的表达式，我们可以将分类问题写作

$$y = C(X)$$

其中，C是关于X的函数，往往称为分类器（classifier），如Logistic分类、决策树、随机森林和支持向量机等都是经典的分类器。

无监督学习指的是只有X而没有y，对于这类数据，我们无法像监督学习方法那样拟合模型去预测y。所以，无监督学习往往用于理解数据的结构、数据降维等。无监督学习经典的方法有聚类分析（clustering）、主成分分析（principle component analysis）、因子分析（factor analysis）、关联规则（association rule）、社交网络（social network）等。

最后，在实际问题中，假设共有n个观测，其中有m（$m < n$）个既能观测到X，也能观测到y，而剩下的$n - m$个由于数据采集困难等原因，只能观测到X，而无法观测到y。例如，在信用评分研究中，假设公司数据库里有10000名顾客的资料，其中已经给6000名顾客发放了贷款，并且已知有500名发生了违约，5500名没发生违约。而剩下的4000名顾客由于还未发放贷款，故他们是否违约我们是不知道的。在建模时，如果综合利用了这两部分信息，则把这类问题称为半监督学习。

注意，图 1-4 展示的数据科学模型方法的划分并不是绝对的。例如，Logistic 分类主要是针对因变量取离散值的问题，但在很多书上，我们习惯地称为 Logistic 回归。再如，决策树、随机森林、支持向量机、神经网络等方法除可以针对取离散值的因变量建模（分类）外，还可以针对取连续值的因变量建模（回归），但在实际应用中，这些方法更多是应用到分类问题上，所以在本书里，我们主要将它们归到分类中。另外，图 1-4 罗列的这些方法并非是全部的方法，随着该领域的快速发展，每年都有很多新的方法提出，由于篇幅有限，本书主要讲解在实践中被反复检验的经典方法，这只是数据科学方法浩瀚的大海里取的一瓢水而已。

1.4 R 语言的优势

R 语言是由新西兰奥克兰大学的 Ross Ihaka 与 Robert Gentleman 一起开发的一个面向对象的编程语言，因两人的名都是以 R 开头，所以命名为 R 语言。R 语言是一个免费开源编程语言，它能够自由有效地用于统计计算和绘图的语言和环境，可以在 UNIX、Windows 和 Mac OS 系统运行，它提供了广泛的统计分析、机器学习和绘图技术。R 语言的前身是 S 语言，S 语言是贝尔实验室（Bell Laboratories）的 Rick Becker、John Chambers 和 Allan Wilks 开发的。最初，S 语言的实现版本主要是 S-PLUS，这是一个商业软件，提供了一系列统计和图形显示工具，一度是数据分析领域里面的标准语言，但是正在逐步被 R 语言取代。R 语言是一套完整的数据处理、计算和制图软件系统，是一套开源的数据分析解决方案，由一个庞大而活跃的全球性社区维护。R 语言不仅是一种统计软件，也是一种统计分析与计算的环境。

KDnuggets 网站每年都会做一些数据分析和数据挖掘软件使用的专题问卷调查。据 KDnuggets 网站 2016 年对 2895 名数据科学家、数据挖掘工程师等进行关于过去 12 个月数据挖掘和数据分析所使用的编程语言的调查显示（http://blog.revolutionanalytics.com/2016/06/r-holds- top-ranking-in-kdnuggets-software-poll.html），R 语言名列榜首（如图 1-5 所示），占接近半壁江山（49%），而紧随其后的 Python、SQL 则在某一领域具有各自独到的优势，而 SAS 和 MATLAB 等被甩出前 10 名。

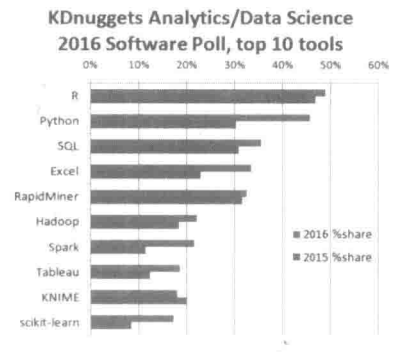

图 1-5 数据挖掘与数据分析编程语言使用调查结果