整合R语言深藏不露的强大威力，决胜数据分析之巅

Broadview
www.broadview.com.cn

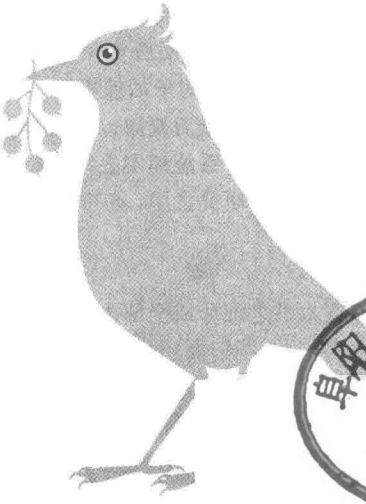
R语言实战

机器学习与数据分析

左飞 著

且听我将统计学之精髓娓娓道来
助你砥砺大数据时代的掘金技法
探寻数据挖掘之术，拨开机器学习迷雾，点破公式背后的层层玄机

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



R语言实战

机器学习与数据分析

左飞 著

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

内 容 简 介

经典统计理论和机器学习方法为数据挖掘提供了必要的分析技术。本书系统地介绍统计分析和机器学习领域中最为重要和流行的多种技术及其基本原理，在详解有关算法的基础上，结合大量R语言实例演示了这些理论在实践中的使用方法。具体内容被分成三个部分，即R语言编程基础、基于统计的数据分析方法以及机器学习理论。统计分析与机器学习部分又具体介绍了参数估计、假设检验、极大似然估计、非参数检验方法（包括列联分析、符号检验、符号秩检验等）、方差分析、线性回归（包括岭回归和Lasso方法）、逻辑回归、支持向量机、聚类分析（包括K均值算法和EM算法）和人工神经网络等内容。同时，统计理论的介绍也为深化读者对于后续机器学习部分的理解提供了很大助益。知识结构和阅读进度的安排上既兼顾了循序渐进的学习规律，亦统筹考虑了夯实基础的必要性。本书内容与实际应用结合紧密，又力求突出深入浅出、系统翔实之特色，对算法原理的解释更是细致入微。

本书非常适合大专院校相关专业师生自学研究之用，亦可作为数据分析和数据挖掘相关领域从业人员的参考指导用书。

图书在版编目（CIP）数据

R 语言实战：机器学习与数据分析 / 左飞著. —北京：电子工业出版社，2016.5
ISBN 978-7-121-28669-8

Ⅰ. ①R… Ⅱ. ①左… Ⅲ. ①程序语言—程序设计 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字(2016)第 089328 号

策划编辑：付　睿
责任编辑：李云静
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
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：24.5　字数：560 千字
版　　次：2016 年 5 月第 1 版
印　　次：2016 年 5 月第 1 次印刷
印　　数：3000 册　　定价：79.00 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——蕴藏巨大财富的宝藏

19 世纪中叶，英国伦敦曾经爆发过一场规模很大的霍乱。由于彼时人们对霍乱的致病机理还不甚了解，因此疫情在很长一段时间内都无法得到有效的控制。英国医师约翰·斯诺用标点地图的方法研究了当地水井分布和霍乱患者分布之间的关系，发现有一口水井周围，霍乱患病率明显较高，借此找到了霍乱暴发的原因：一口被污染的水井。关闭这口水井之后，霍乱的发病率明显下降。这便是数据分析在历史上展示其威力的一次成功案例。

毋庸置疑，数据是一座巨大的宝藏，而我们要做的恰恰就是挖掘这座宝藏。特别是进入信息时代以来，“大数据”这个概念更是越来越多地被人们提及。很多国家甚至把大数据提升到国家战略的高度。例如，我国的“十三五”规划建议中就提出：“实施国家大数据战略，推进数据资源开放共享。”

尽管“大数据”这个名词听起来很时髦，但是由此反映出来的对于数据本身的重视却并不是一个多么新鲜的现象。中国古代的施政治国观念中就非常强调掌握数据的重要性。例如商鞅变法中就提出，“强国知十三数……欲强国，不知国十三数，地虽利，民虽众，国愈弱至削”。

随着时代的进步，人们对于数据的重视程度更是有增无减，世界各国，概莫能外。列宁就曾经说过：“有许多问题，而且是涉及现代国家经济制度和这种制度之发展的最根本问题……如果不根据某个一定的纲要收集并经统计专家综合的关于某一国家全国情况的浩繁材料，就无法加以比较并认真地研究。”毛主席也曾指出：“胸中有‘数’。就是说，对情况和问题一定要注意到它们的数量方面，要有基本的数量分析。任何质量都表现为一定的数量，没有数量也就没有质量。”

“大数据时代，统计学依然是数据分析灵魂。”

人民网在 2015 年 7 月曾经以《大数据时代，统计学依然是数据分析灵魂》为题刊发了一篇对某位知名专家的访谈。其间，这位专家就形象地说道：“大数据是‘原油’而不是‘汽油’，不能被直接拿来使用。就像股票市场，即使把所有的数据都公布出来，不懂的人依然不知道数据代表的信息。”同时该篇文章也引用了美国加州大学伯克利分校迈克尔·乔丹教授的观点：“没有系统的数据科学作为指导的大数据研究，就如同不利用工程科学的知识来建造桥梁，很多桥梁可能会坍塌，并带来严重的后果。”

面对大数据，现在很多人可能会时常把数据挖掘这样时髦又深奥的词汇挂在嘴边，而认为或许传统的统计学此时已经不合时宜。这种观点在我看来至少有两个致命的问题。首先，传统的统计学方法仍然在各个领域扮演着不可取代的重要作用。包括生命科学、经济学、管理学等在内的诸多学科都涉及大量的数据分析工作，并从中汲取推进各自领域进步的动力。这里所谓的数据分析工作，更多的是基于传统统计分析方法来完成的。其次，很多数据挖掘的技术又是建立在传统的统计理论基础之上的。例如，期望最大化算法中就用到了极大似然估计。不仅如此，像计量经济中常常用到的“回归”，它既是一种数据挖掘方法，同时又是传统的统计学中必不可少的重要组成部分。

机器学习 VS 数据挖掘

在大量数据背后很可能隐藏了某些有用的信息或知识，而数据挖掘就是指通过一定方法探寻这些信息或知识的过程。另一方面，数据挖掘同时受到很多学科和领域的影响，大体上看，数据挖掘可以被视为数据库、机器学习和统计学三者的交叉。简单来说，对数据挖掘而言，数据库提供了数据管理技术，而机器学习和统计学则提供了数据分析技术。而本书所关注的重点，恰恰在于以机器学习和统计学为基础的数据分析方法。

从名字中就不难看出，机器学习最初的研究动机是为了让计算机具有人类一样的学习能力以便实现人工智能。显然，没有学习能力的系统很难被认为是智能的。而这个所谓的学习，就是指基于一定的“经验”而构筑起属于自己之“知识”的过程。小蝌蚪找妈妈的故事很好地说明了这一过程。小蝌蚪们没有见过自己的妈妈，它们向鸭子请教。鸭子告诉它们：“你们的妈妈有两只大眼睛。”看到金鱼有两只大眼睛，小蝌蚪们便把金鱼误认为是自己的妈妈。于是金鱼告诉它们：“你们妈妈的肚皮是白色的。”小蝌蚪们看见螃蟹是白肚皮，又把螃蟹误认为是自己的妈妈。螃蟹便告诉它们：“你们的妈妈有四条腿。”小蝌蚪们看见一只乌龟摆动着四条腿在水里游，就把乌龟误认为是自己的妈妈。于是乌龟又说：“你们的妈妈披着绿衣裳，走起路来一蹦一跳。”在这个学习过程中，小蝌蚪们的“经验”包括鸭子、金鱼、螃蟹和乌龟的话，以及“长得像上述四种动物的都不是妈妈”这样一条隐含的结论。最终，它们学到的“知识”就是“两只

大眼睛、白肚皮、绿衣裳、四条腿，一蹦一跳的就是自己的妈妈”。当然，故事的结局，小蝌蚪们就是靠着学到的这些知识成功地找到了妈妈。反观机器学习，由于“经验”在计算机中主要是以“数据”的形式存在的，所以机器学习需要设法对数据进行分析，然后以此为基础构建一个“模型”，这个模型就是机器最终学到的“知识”。可见，小蝌蚪学习的过程是从“经验”学到“知识”的过程。相对应地，机器学习的过程则是从“数据”学到“模型”的过程。正是因为机器学习能够从数据中学到“模型”，而数据挖掘的目的恰恰是找出数据背后的“信息或知识”，二者不谋而合，所以机器学习才逐渐成为数据挖掘最为重要的智能技术供应者而备受重视。

正如前面所说的，机器学习和统计学为数据挖掘提供了数据分析技术。而另一方面，统计学也是机器学习得以建立的一个重要基础。所以，统计学本身就是一种数据分析技术的同时，它也为以机器学习为主要手段的智能数据分析提供了理论基础。可见统计学、机器学习和数据挖掘之间是紧密联系的。基于这样的认识，我们可以说本书的副标题“机器学习与数据分析”主要包含了下面几层意思。首先，如果把数据分析看作狭义上的以数理统计为基础的统计分析方法，那么本书就涵盖了为数据挖掘提供分析技术的两部分内容，即以机器学习为基础的和以统计学为基础的数据分析方法。其次，如果你把数据分析看作更为宏观的包含了数据挖掘在内的广义数据分析技术，那么为了引入以机器学习为出发点的智能分析技术，前期的统计分析知识则是帮助读者夯实数据分析基础的必要准备。

关于本书

R 语言是当今最为流行的统计分析语言和数据分析环境之一。它是属于 GNU 系统的一个自由、免费、源代码开放的软件，并拥有媲美于商业软件的强大统计分析和绘图功能。此外，R 语言还拥有数以万计贡献者在为其开发各种功能包，配合这些包的使用，R 的功能得到了极大拓展，几乎可以完成任何你想要的数据分析与挖掘任务。本书选择 R 语言作为描述语言和开发环境，不仅通过诸多详尽的实例来演示 R 的使用，更为那些新近接触 R 语言的读者提供了很好的入门指导。我们相信，无论你属于何种程度的 R 语言使用者，都可以很好地利用本书来增进数据分析和挖掘的技术和能力。

经典统计理论和机器学习方法为数据挖掘提供了必要的分析技术。本书系统地介绍统计分析和机器学习领域中最为重要和流行的多种技术及其基本原理，在详解有关算法的基础上，结合大量 R 语言实例演示了这些理论在实践中的使用方法。具体内容被分成三个部分，即 R 语言编程基础、基于统计的数据分析方法以及机器学习理论。统计分析与机器学习部分又具体介绍了参数估计、假设检验、极大似然估计、非参数检验方法（包括列联分析、符号检验、符号秩检验等）、方差分析、线性回归（包括岭回归和 Lasso 方法）、逻辑回归、支持向量机、聚类分析（包括 K 均值算法和 EM 算法）和人工神经网络等内容。同时，统计理论的介绍也为深化读

者对于后续机器学习部分的理解提供了很大助益。知识结构和阅读进度的安排上既兼顾了循序渐进的学习规律，亦统筹考虑了夯实基础的必要性。尽管作为一个非常宏大的话题，在有限的篇幅内我们不能将机器学习的所有方法尽述，但循着本书所提供的自学路线图，却可以建立一个十分扎实的基础以及对数据分析技术相当清晰的认识和理解。

统计学大师乔治·博克斯曾经是统计学家埃贡·皮尔逊的学生，而埃贡·皮尔逊则是统计学之父卡尔·皮尔逊的儿子。此外，乔治·博克斯还是统计学界的另一位巨擘罗纳德·费希尔的女婿。从这个角度来说，乔治·博克斯无疑集成了两位统计学宗师的学术思想，他有一句广为人们提及的名言说道："所有的模型都是错的，但其中一些是有用的。"所以，无论是基于统计的方法，还是基于机器学习的方法，最终的模型都是对现实世界的抽象，而非毫无偏差的精准描述。相关理论只有与具体分析实例相结合才有意义。而在这个所谓的结合过程中，你既不能期待一种模型（或者算法）能够解决所有的（尽管是相同类型的）问题，也不能在面对一组数据时就能（非常准确地）预先知道哪种模型（或者算法）才是最适用的。或许你该记住另外一句话："No clear reason to prefer one over another. Choice is task dependent（没有明确的原因表明一种方法胜于另外一种方法，选择通常是依赖于具体任务的）"。这也就突出了数据挖掘领域中实践的重要性，或者说由实践而来的经验之重要性。

为了力求让读者"知其然，更知其所以然"，对于晦涩的数据挖掘算法，本书都配合有完整详尽的推导过程。而包括统计数据分析在内的部分，我们更是借助 R 语言的强大能力，抽丝剥茧，逐条演示了各种检验方法、估计方法和分析方法的执行步骤，让读者深刻领悟到每一条简单函数背后所蕴藏的复杂机制。

"纸上得来终觉浅，绝知此事要躬行"，深化统计分析的基本思想，并锤炼运用 R 语言进行数据挖掘的能力，很大程度上有赖于编程实践活动。本书涉及的所有源代码，读者都可以从在线支持资源"http://blog.csdn.net/baimafujinji"中下载得到，勘误表也将实时发布到此博客上。同时欢迎读者就本书中的问题和不足与笔者展开讨论，有关问题请在上述博客中留言。

本书由左飞统稿并执笔。此外刘航、吴凯、姜萌、何鹏、胡俊、李召恒、初甲林、薛佟佟等人也参与了本书编写工作，笔者在此表示由衷的感谢。

自知论道须思量，几度无眠一文章。由于时间和能力有限，书中纰漏在所难免，真诚地希望各位读者和专家不吝批评、斧正。

目录

第 1 章

初识 R 语言

欢迎学习 R 语言！作为当今最流行的统计分析语言之一，R 语言在科学研究、生物医药、市场营销、经济分析等涉及数据统计的领域都有非常广泛而重要的应用。本章是全书的导引部分，它将帮助大家建立对 R 语言的初步认识，并通过一些简单的例子使读者熟悉 R 语言开发环境。

1.1 R 语言简介

说到 R 的起源，便不得不提及上世纪 80 年代诞生于贝尔实验室的 S 语言。彼时正专注于现代统计模型和数据分析方法研究的三位科学家——约翰·钱伯斯（John Chambers）、瑞克·柏克（Rick Becker）以及后来加入的艾兰·威克斯（Allan Wilks）成功地开发了一种用来进行数据处理、统计分析和作图的解释型语言，也就是 S 语言。而曾与他们三位共事过的华盛顿大学统计学教授道格拉斯·马丁（Douglas Martin）开发了一个 S 语言的实现版本，也就是 S-PLUS 的最初版本。

马丁很快发现了 S 的潜在商业价值，但是贝尔实验室当时却没有将 S 语言商业化的设想。于是马丁便创立了 Statistical Sciences 公司，以 S-PLUS 的形式将 S 语言推向市场。所以，S-PLUS 其实就是基于 S 语言的一款商业软件，后来在 1993 年，马丁将 Statistical Sciences 卖给了 MathSoft 公司，而 S-PLUS 在 MathSoft 公司也得到了长足的发展并在商业上取得了成功。2001 年，MathSoft 公司更名为 Insightful，并将公司总部迁往西雅图。2008 年，TIBCO 公司成功将 Insightful 公司收购。

S 语言的另外一种实现版本就是本书要介绍的 R。R 最初是由新西兰奥克兰大学的罗斯·艾卡（Ross Ihaka）和罗伯特·杰特曼（Robert Gentleman）两位教授实现的，现在由“R 开发核心团队”负责开发以及维护。现在 R 是属于 GNU 系统的一个自由、免费、开源的软件。R 可以被认为是当前最为流行的一种用于数据分析和统计制图的语言及操作环境。所以当提到 R 时，

既是指一种计算机语言也是指一种软件环境。本书后面主要使用 R 这个称谓，有时也会使用 R 软件、R 语言或 R 系统来称呼它。读者应该明白尽管这些称谓各异，但是它们所指代的事物其实是统一的。

当前数据分析已经成为非常热门的话题，各行各业每天都在进行着数据分析活动。而可以用于数据分析的软件也是林林总总，例如我们所熟知的 MATLAB、Excel、SPSS、SAS、Stata、EViews 和 S-PLUS 等。那么为什么选择 R 呢？总的来说，R 具有如下一些主要特点：

- R 是一个完全免费的自由软件。尽管 S-PLUS 也是一款非常优秀的统计分析软件，但使用它需要支付一笔费用，而 R 则是一个免费的统计分析软件。
- R 支持多种操作系统。它有 UNIX、Mac OS 和 Windows 等多个版本，都是可以免费下载和使用的。它们的安装文件以及安装说明可以通过 CRAN 获得。
- R 是开放源代码的。它的源代码可自由下载使用，因此它有来自全球的热心用户为其编写软件包。借由这些软件包，R 的功能被极大地扩展，针对某些具体领域的统计分析功能被不断完善和加强。例如像经济计量、财经分析功能等就是通过扩展包实现的。
- R 具有突出的统计分析能力。R 内嵌了许多实用的统计分析功能，统计分析的结果也能被直接显示出来；一些中间结果既可保存到专门的文件中， 也可以直接用于进一步的分析。R 的功能也可以通过安装包来增强。
- R 拥有强大的绘图功能。数据分析结果可以通过专业的统计图形来呈现。内嵌的作图函数能将产生的图片展示在一个独立的窗口中，并能将之保存为各种形式的文件。如图 1-1 所示就为利用 R 绘制的统计图形。

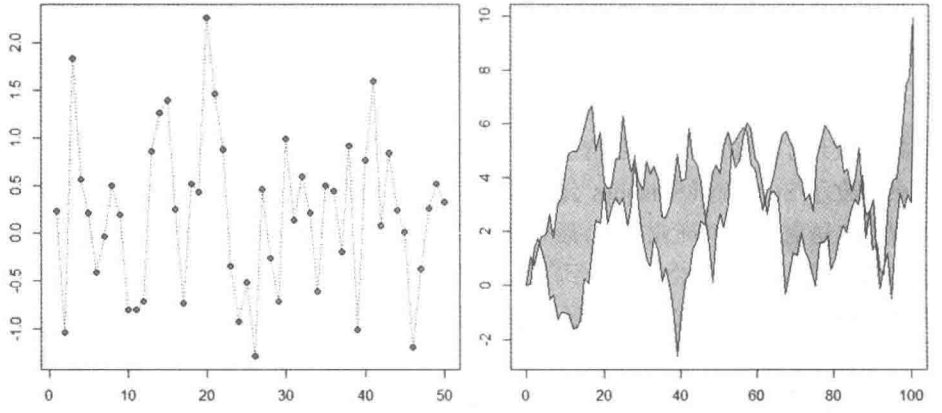


图 1-1　用 R 绘制的统计图形

- R 是面向对象的编程语言。R 比其他统计学或数学专用的编程语言有更强的面向对象功能，它提供了包括继承、多态和封装等在内的面向对象特性。