大数据
分析统计应用丛书

全国应用统计专业学位研究生教育指导委员会推荐用书

大数据 BIG DATA MINING AND STATISTICAL MACHINE LEARNING 挖掘与统计机器学习

主编 吕晓玲 宋捷

中国人民大学出版社

全国应用统计专业学位研究生教育指导委员会推荐用书

大数据 BIG DATA MINING AND STATISTICAL MACHINE LEARNING
挖掘与统计机器学习

主编 吕晓玲 宋捷

中国人民大学出版社
·北 京·

图书在版编目（CIP）数据

大数据挖掘与统计机器学习/主编吕晓玲，宋捷. —北京：中国人民大学出版社，2016.7
（大数据分析统计应用丛书）
ISBN 978-7-300-23101-3

Ⅰ.①大… Ⅱ.①吕…②宋… Ⅲ.①数据处理②机器学习 Ⅳ.①TP274②TP181

中国版本图书馆 CIP 数据核字（2016）第 153795 号

大数据分析统计应用丛书
大数据挖掘与统计机器学习
主编　吕晓玲　宋捷
Dashuju Wajue yu Tongjijiqixuexi

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京七色印务有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2016 年 7 月第 1 版
印　　张	15 插页 1	**印　　次**	2016 年 7 月第 1 次印刷
字　　数	356 000	**定　　价**	35.00 元

大数据分析统计应用丛书编委会

总　序

一

统计学是收集、分析、展示和解释数据的方法性质的一门科学。信息技术的蓬勃发展，使统计在经济、社会、管理、医学、生物、农业、工程等领域有了越来越多、越来越深入的应用。2011 年 2 月，国务院学位委员会第 28 次会议通过了新的《学位授予和人才培养学科目录（2011）》，将统计学上升为一级学科，这为统计学科建设与发展提供了难得的机遇。

一般认为，麦肯锡公司的研究部门——麦肯锡全球研究院（MGI），在 2011 年首先提出了大数据时代（age of big data）的概念，并引起了全球广泛的反响。大数据是指随着现代社会的进步和信息通信技术的发展，在政治、经济、社会、文化等各个领域形成的规模巨大、增长与传递迅速、形式复杂多样、非结构化程度高的数据或者数据集。它的来源包括传感器、移动设备、在线交易、社交网络等，其形式可以是各种空间数据、报表统计数据、文字、声音、图像、超文本等各种环境和文化数据信息等。大数据时代是一个海量数据开始广泛出现、海量数据的运用逐渐普遍的新的历史时期，也是我们需要认真研究与应对的一个新的社会环境与社会形式。

大数据时代对统计专业的学生提出了更高的要求。他们不仅需要具有扎实的统计理论基础，并且要熟练掌握各种处理大数据和统计模型分析的计算机技能，还要懂得如何提出研究问题、如何判断数据质量、如何评价模型和方法，以及如何准确清晰地呈现分析结果。这对统计教育和人才培养提出了新的目标和方向。

二

顺应时势，在教育部全国应用统计专业学位研究生教育指导委员会推动下，由中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学五所高校发起，集中统计学科、计算机学科、经济与管理学科的相关学院优势，依托应用统计专业硕士项目，组建了北京大数据分析硕士培养协同创新平台。2014 年 9 月首届实验班正式招生并开始授课。

实验班每年招收约 50～60 名学生，分别来自中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学等院校。他们均是以优异成绩进入上述高校应用统计硕士项目的本科毕业生，对大数据分析有浓厚的兴趣，立志为大数据分析领域的发展做出贡献。

大数据分析硕士的培养是为了满足政府部门和企业等用人单位利用大数据决策的需求，其核心竞争力是快速部署从大数据到知识发现和价值的能力，培养方案与国际接轨，核心内容是面向大数据的统计分析和挖掘技术。经过前期的充分论证，大数据分析硕士培养方案确定了核心必修课与分方向的选修课。必修课的重点内容为统计学和计算机科学的交叉部分，侧重于培养从大数据到价值的实践能力，包括大数据分析必备的计算机基础技能、面向大数据分析的计算机编程能力、大数据统计建模和挖掘能力。每门必修课均配备了 5 人以上的教学团队，由包括国家“千人计划”入选者、长江学者、国家杰出青年基金获得者在内的在相关领域有较高造诣的中青年学者组成。

大数据分析硕士培养协同创新平台是一个面向政府部门和企业等大数据分析人才需求单位开放的平台，目标是建成一个政产学研有机融和的协同创新平台。2014 年 5 月 19 日平台成立大会就汇集了《人民日报》、新华社、中央电视台、中国移动、中国联通、中国电信、全国手机媒体专业委员会、SAS（北京）有限公司、华闻传媒产业创新研究院、北京华通人商用信息有限公司、龙信数据（北京）有限公司等，成为该平台的第一批实践培养和研发基地。在 2014 年 9 月开学典礼上又有中国科学院计算机网络信息中心、中国中医科学院、商务部国际贸易学会、国家食品安全风险评估中心、北京商智通信息技术有限公司、史丹索特（北京）信息技术有限公司、北京太阳金税软件技术有限公司、北京京东叁佰陆拾度电子商务有限公司、北京知行慧科教育科技有限公司、中关村大数据产业联盟、艾瑞咨询集团 11 家单位加入平台建设的联盟协作单位。实际部门的踊跃参与说明大数据分析人才培养的巨大发展空间。为了加强大学与实际部门专家的双导师制度，开学典礼上为第一届实验班专门聘任了 26 名实际部门专家担任硕士研究生指导教师。

2015 年 1 月 15 日，大数据分析硕士培养协同创新平台联合京东、奇虎 360、艾瑞咨询集团、华通人等多家公司举办了针对学生实习的宣讲会。会后组织学生到各相关部门进行有关数据挖掘、大数据分析的实习工作，学生们得到了锻炼。

为活跃学术氛围，拓展学生视野，大数据分析硕士培养协同创新平台组织了大数据分析学术系列讲座，邀请学界、业界相关人士交流分享学术、行业前沿的经验，共同推进大数据人才培养以及学术成果的转化。

三

迄今为止，五校联合大数据分析硕士实验班已经成功开展两届。在此基础上，课程组全体教师及时收集学生反馈意见，积极组织讨论，联合中国人民大学出版社，启动了“大数据分析统计应用丛书”的编写工作。

本套丛书第一期出版四本。《大数据分析计算机基础》着重介绍数据分析必备的计算机技能，包括 Linux 操作系统与 shell 编程，数据库操作与管理；面向大数据分析的计算机编程能力，我们重点推荐了 Python 语言。《大数据探索性分析》的内容包括大数据抽样、预处理、探索性分析、可视化以及时空大数据案例。《大数据分布式计算与案例》介

绍了单机并行计算以及 Hadoop 分布式计算集群，在此基础上介绍了 HDFS 文件管理系统以及 MapReduce 框架、各种统计模型的 MapReduce 实现，此外还介绍了处理大数据最常使用的 Hive，HBase，Mahout 以及 Spark 等工具。《大数据挖掘与统计机器学习》介绍了常用的统计学习的回归和分类模型、模型评价与选择的方法、聚类和推荐系统等算法，所有方法均配有 R 语言实现案例，支持向量机和深度学习方法给出了 Python 实现案例，最后是两个数据量在 10G 以上的大数据案例分析，所有的数据和程序均可下载。相信读者在学习本套丛书的过程中，数据处理与分析能力会得到锻炼和提高。

在丛书第一期的基础上，我们也在积极策划第二期，内容包括非结构化大数据分析、大数据统计模型、统计计算与统计优化方法等，希望可以涵盖更多的数据类型与统计方法。

该丛书面向的读者主要是应用统计专业硕士，也可以作为统计专业高年级本科生、其他专业的本科生、研究生以及对大数据分析有兴趣的从业人员的参考书，希望这套丛书可以为我国大数据分析人才的培养奉献我们的绵薄之力。

丛书编委会

前　言

大数据时代的到来，使我们的生活在政治、经济、社会、文化各个领域都产生了很大改变。“数据科学”一词应运而生。如何更好地对海量数据进行分析、得出结论并做出智能决策是统计工作者面临的机遇与挑战。

本书介绍数据挖掘与统计机器学习领域最常用的模型和算法，包括最基础的线性回归和线性分类方法，以及模型选择和模型评价的概念和方法，进而介绍非线性的回归和分类方法（包括决策树与组合方法、支持向量机、神经网络以及在此基础上发展的深度学习方法）。最后介绍无监督的学习中的聚类方法和业界广泛使用的推荐系统方法。除了方法的理论讲解之外，我们给出了每种方法的 R 语言实现，以及应用 Python 语言实现深度学习和支持向量机两种方法。本书的一个亮点是最后一章给出的两个大数据案例，数据量均在 10G 左右。我们同时给出了单机版（Python、数据库、R）和分布式（Hadoop、Hive、Spark）两种实现方案。原始数据和程序代码均可在出版社提供的网址下载。

本书面向的主要读者是应用统计专业硕士，希望能够拓展到统计专业高年级的本科生以及其他各个领域有数据分析需求的学生和从业人员。对于侧重应用的初学者，可略过带星号的章节。

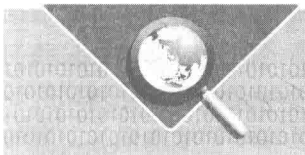
本书由吕晓玲撰写第 1 章、第 2 章、第 10 章，吕晓玲、潘蕊合写第 4 章和第 5 章，吕晓玲、宋捷合写第 3 章、第 7 章，古楠楠撰写第 6 章，褚挺进撰写第 8 章，尹建鑫撰写第 9 章，最后由吕晓玲统稿校对。

感谢北京五校联合（中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学）大数据分析硕士培养协同创新平台的所有领导和教师；感谢中国人民大学出版社的鼎力支持；感谢中国人民大学数据挖掘中心（www. rucdmc. net）的学生参与本书的写作和校对，他们是：钟琰、王小宁、刘撷芯、王高斌、安梦颖、胡见秋、范一苇、苏嘉楠、程豪、范超、要卓、李天博、林毓聪、闫晗、刘梦杭、孙亚楠、董峰池。

数据挖掘与统计机器学习是一个方兴未艾、蓬勃发展的学科领域，鉴于作者的能力和时间非常有限，本书的内容难免有不足和纰漏，还望广大读者不吝赐教，多提宝贵意见。

吕晓玲　宋捷

目　录



第1章 概 述

1.1 名词演化

数据挖掘（data mining）这一名词产生于1990年前后，迅速在学术界和商业界得到广泛应用与发展。实际上，数据挖掘与统计数据分析的目标没有什么本质的差别。按照《不列颠百科全书》，统计可以定义为收集、分析、展示、解释数据的科学。这是历史相对悠久的统计在其发展过程中逐渐形成的被世人认可的定义。它包含一系列概念、理论和方法，有一个比较稳定的知识结构和体系。数据挖掘也完全符合这个定义，但由于它的发展历史较短，初期主要由计算机科学家开创，脱离了传统统计的体系，因此有其自身的特点。数据挖掘有时也称作数据库的知识发现（Knowledge Discovery in Databases，KDD）。严格来讲这两个概念并不完全一致。同期经常被人们使用的两个名词是模式识别（pattern recognition）和人工智能（artificial intelligence）。目前使用更多的术语是机器学习（machine learning）。从统计学者的角度则称为统计机器学习（statistical machine learning）或统计学习（statistical learning）。

一般认为，麦肯锡公司的研究部门——麦肯锡全球研究院（MGI）在2011年首先提出大数据时代（age of big data）的概念，在全球引起广泛的反响。早在2001年，美国信息咨询公司Gartner的分析师Doug Laney就从数据量（volume）、多样化（variety）和快速化（velocity）三个维度分析了在数据量不断增长的过程中所面临的挑战和机遇。在大数据这一概念被广泛传播后，IBM副总裁Steven Mills于2011年在此基础上提出大数据的第四个维度——价值密度低（veracity）。人们普遍认为其蕴涵巨大的价值，但如何从中快速准确地提取真实有价值的信息是大数据处理技术的关键。

大数据，是指随着现代社会的进步和通信技术的发展，在政治、经济、社会、文化各个领域形成的规模巨大、增长与传递迅速、形式复杂多样、非结构化程度高的数据或者数据集。它的来源包括传感器、移动设备、在线交易、社交网络等，其形式可以是各种空间

数据，报表统计数据，文字、声音、图像、超文本等各种环境和文化数据信息等。大数据时代是一个海量数据开始广泛出现、海量数据的运用逐渐普遍的新的历史时期，也是我们需要认真研究与应对的一个新的社会环境与社会形式。数据科学（data science）一词应运而生。它可以被看作数学逻辑和统计批判性思维、计算机科学以及实际领域知识这三者的交集（见图 1—1）。

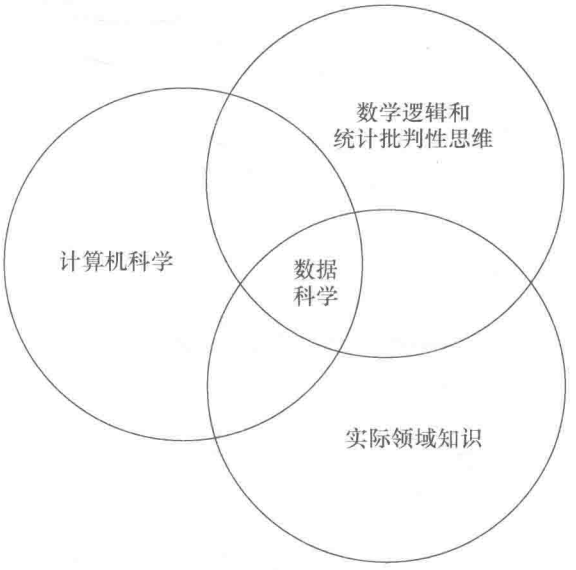


图 1—1　数据科学

1.2　基本内容

统计学是一门科学，科学的基本特征是其方法论：对世界的认识源于观测或实验所得的信息（或者数据），总结信息时会形成模型（也叫作假说或理论），模型会指导进一步的探索，直到遇到这些模型无法解释的现象，从而导致对这些模型的更新或替代。这就是科学的方法，只有用科学的方法进行探索才能称之为科学（吴喜之，2016）。统计的思维方式是归纳，也就是从数据所反映的现实中得到一般的模型，希望以此解释数据所代表的那部分世界。这和以演绎为主的数学思维方式相反，演绎是在一些人为的假定（或者一个公理系统）之下推导出各种结论。

在统计科学发展的前期，由于没有计算机，不可能应付庞大的数据量，只能在对少量数据的背景分布做出诸如独立同正态分布之类的数学假定后，建立一些数学模型，进行手工计算，并推导出由这些模型所得结果的性质，比如置信区间、相合性等。有时候这些性质是利用中心极限定理或大样本定理得到的当样本量趋于无穷时的理论性质。这些性质对总体的分布以及样本的形式都有很多的假定。这种发展方式给统计打上了很深的数学烙印。统计发展的历史体现在模型驱动的研究及教学模式上。以模型而不是数据为主导的研

究方式导致统计在某种程度上“自我封闭、自我欣赏”，结果是很可能丢掉许多属于数据科学的领域。

模型驱动的研究在前计算机时代有其合理性，但在计算机技术快速发展的大数据时代，必须转变这种模式。统计是应用的学科，将统计方法应用到各个领域，解决实际问题是统计的灵魂。在分析数据时，首先寻求现有方法，当现有方法不能满足需求时，就要根据数据的特征创造新的方法，并对其理论性质进行深入的探讨。这是统计近年来飞速发展的历程。创造模型的目的是解决实际问题。统计研究应该由问题或者数据而不是模型、数学公式所驱动。此外，为了让新的模型得到真正的应用，对模型的求解和计算提出了很高的要求，统计研究必须同时考虑算法复杂度和计算编程高效实现的问题。

目前被广泛使用的有着极高口碑的统计学教材有两本，一本是 Trevor Hastie，Robert Tibshirani，and Jerome Friedman 编写的 *The Elements of Statistical Learning—Data Mining，Inference and Prediction*（Hastie，Tibshirani and Friedman，2008），简称 ESL。第二本是 Gareth James，Daniela Witten，Trevor Hastie，and Robert Tibshirani 编写的 *An Introduction to Statistical Learning with Applications in R*（James et al.，2013），简称 ISL。第一本书面向的读者更专业一些，内容较多，理论偏难。第二本书面向的读者更广泛，内容偏基础，更强调应用，有各个方法的 R 语言实现实例。两本书均将统计学习方法分为两种，即有监督学习（supervised learning）和无监督学习（unsupervised learning）。所谓有监督学习，就是在分析问题时，数据中有一个明确的目标变量 Y（也称作因变量、响应变量、输出变量等），可以通过建立它对其他变量 X（也称为自变量、协变量、解释变量、输入变量、特征、字段等）的模型来预测。如果 Y 的取值是连续型的，则称为回归分析。如果 Y 是一个分类标签，则称作分类问题。目前广泛使用的有监督学习方法包括决策树及其组合算法、神经网络、支持向量机、最近邻居法、朴素贝叶斯方法等。无监督学习是指数据中没有明确的目标变量，通过一些方法寻找数据之间的相互关系或者模式（pattern）。无监督学习典型的例子是主成分分析、聚类和关联规则等。

本书面向的主要读者是应用统计专业硕士，希望能够拓展到统计专业高年级的本科生以及其他各个领域有数据分析需求的学生和从业人员。从内容选择和章节安排上，我们借鉴了上述两本经典教材。在理论难度方面要高于 ISL 这本书，但没有达到 ESL 的水平，介于两者之间。类似 ISL，本书每一章在最后给出实际数据分析的上机实践以及 R 程序代码。在内容的选取方面，我们首先介绍最简单、最基础的线性回归方法（第 2 章）和线性分类方法（第 3 章），之后介绍重要的模型评价和模型选择的概念和方法（第 4 章）。非线性的回归和分类方法包括决策树与组合方法（第 5 章），神经网络以及在此基础上发展的深度学习方法（第 6 章）和支持向量机方法（第 7 章）。其中第 6 章的深度学习方法是近期迅速发展起来的，ESL 和 ISL 两本书并没有介绍。第 8 章着重介绍了无监督学习中的聚类方法。第 9 章介绍了目前业界广泛使用的推荐系统方法，这也是 ESL 和 ISL 两本书没有介绍的内容。对于业界广泛使用的深度学习和支持向量机方法，R 语言实现效率偏低，因此我们还介绍了 Python 调用更专业的程序包快速实现这两种算法的方法。最后，本书的一个亮点就是在第 10 章给出了两个大数据案例，数据量均在 10G 左右，作为初试大数据的读者，我们认为是非常合适的。我们给出了单机版（Python、数据库、R）和分布式（Hadoop、Hive、Spark）两种实现方案。读者可以在出版社的网址下载相应的数据和程

序。遗憾的是，本书并没有包含最近邻、朴素贝叶斯、图模型、非参数方法、关联规则、Pagerank 等一些常用的统计学习方法，有兴趣的读者请参阅其他文献。

1.3 数据智慧

2016 年第 1 期《中国计算机学会通讯》刊登了美国加州大学伯克利分校统计系郁彬教授（美国科学院、美国艺术与科学学院院士）的一篇中译版的文章——《数据科学中的数据智慧》，英文原文的网址链接是 http://www. odbms. org/2015/04/data-wisdom-for-data-science/。

在此，我们想引用郁彬教授的文章作为第 1 章的结束语。郁彬教授深入地讨论了应用统计方法解决实际问题应该注意的事项，明确提出“数据智慧”应该是应用统计学概念的核心。希望读者可以认真阅读这篇文章并思考：在大数据时代，统计数据分析工作者的任务和使命是什么？我们怎样才能正确应用统计方法解决实际问题？

在大数据时代，学术界和工业界的大量研究都是关于如何以一种可扩展和高效率的方式对数据进行存储、交换和计算（通过统计方法和算法）。这些研究非常重要。然而，只有对数据智慧（data wisdom）给予同等程度的重视，大数据（或者小数据）才能转化为真正有用的知识和可被采纳的信息。换言之，我们要充分认识到，只有拥有足够数量的数据，才有可能对复杂度较高的问题给出较可靠的答案。“数据智慧”对于我们从数据中提取有效信息和确保没有误用或夸大原始数据是至关重要的。

“数据智慧”一词是我对应用统计学核心部分的重新定义。这些核心部分在伟大的统计学家（或者说是数据科学家）约翰·图基（John W. Tukey）的文章（“The Future of Data Analysis,” *The Annals of Mathematical Statistics*, Volume 33, Number 1, 1962, pp. 1-67）和乔治·伯克斯（George Box）的文章（“Science and Statistics,” *Journal of the American Statistical Association*, Volume 71, Issue 356, 1976）中都有详细介绍。

将统计学核心部分重新命名为“数据智慧”非常必要，因为它比“应用统计学”这个术语起到更好的概括作用。对于这一点，最好能让统计学领域之外的人也了解到。因为这样一个有信息量的名称可以使人们意识到应用统计作为数据科学一部分的重要性。

依据维基百科对“智慧”词条进行解释的第一句话，我想说：“数据智慧”是将领域知识、数学和方法论与经验、理解、常识、洞察力以及良好的判断力相结合，思辨性地理解数据并依据数据做决策的一种能力。

“数据智慧”是数学、自然科学和人文主义三方面能力的融合，是科学和艺术的结合。如果没有实践经验者的指导，仅通过读书很难学习到“数据智慧”。学习它的最好方法就是和拥有它的人一起共事。当然，我们也可以通过问答的方式来帮助你形成和培养“数据智慧”能力。我这里有 10 个基本问题，我鼓励人们在开始从事数据分析项目时或者在项目进行过程中经常问问自己这些问题。这些问题是按照一定顺序

排列的，但是在不断重复的数据分析过程中，这个顺序完全可以打乱。

这些问题也许无法详尽、彻底地解释“数据智慧”，但是它们体现了“数据智慧”的一些特点。

1. 要回答的问题

数据科学问题最初往往来自统计学或者数据科学以外的学科。例如，神经科学中的一个问题：大脑是如何工作的？或银行业中的一个问题：该向哪组顾客推广新服务？要解决这些问题，统计学家必须与这些领域的专家进行合作。这些专家会提供有助于解决问题的领域知识、早期的研究成果、更广阔的视角，甚至可能对该问题进行重新定义。而与这些专家（他们往往很忙）建立联系需要很强的人际交流技巧。

与领域专家的交流对于数据科学项目的成功是必不可少的。在数据来源充足的情况下，经常发生的事情是在收集数据前还没有精确定义要回答的问题。我们发现自己处在图基所说的“探索性数据分析”（Exploratory Data Analysis，EDA）的游戏中。我们寻找需要回答的问题，然后不断地重复统计调查过程（就像乔治·伯克斯的文章中所述）。由于误差的存在，我们谨慎地避免对数据中出现的模式进行过拟合。例如，当同一份数据既用于对问题进行建模又用于对问题进行验证时，就会发生过拟合。避免过拟合的黄金准则就是将数据进行分割，在分割时考虑到数据潜在的结构（如相关性、聚类性、异质性），使分割后的每部分数据都能代表原始数据。其中一部分用来探索问题，另一部分通过预测或者建模来回答问题。

2. 数据收集

什么样的数据与第1条中要回答的问题最相关？

实验设计（统计学的一个分支）和主动学习（机器学习的一个分支）中的方法有助于回答这个问题。即使在数据收集好了以后考虑这个问题也是很有必要的。因为对理想的数据收集机制的理解可以暴露出实际数据收集过程的缺陷，能够指导下一步分析的方向。

下面的问题会对提问有所帮助：数据是如何收集的？在哪些地点？在什么时间段？是谁收集的？用什么设备收集？中途更换过操作人员和设备吗？总之，试着想象自己在数据收集现场。

3. 数据含义

数据中的某个数值代表什么含义？它测量了什么？它是否测量了需要测量的？哪些环节可能会出错？在哪些统计假设下可以认为数据收集没有问题？（对数据收集过程的详细了解在这里会很有帮助。）

4. 相关性

收集来的数据能够完全或部分回答要研究的问题吗？如果不能，还需要收集其他什么数据？第2条中提到的要点在此处同样适用。

5. 问题转化

如何将第1条中的问题转化成一个与数据相关的统计问题，使之能够很好地回答原始问题？有多种转换方式吗？比如，我们可以把问题转换成一个与统计模型有关的预测问题或者统计推断问题吗？在选择模型前，请列出与回答实质性问题相关的每一种转化方式的优点和缺点。

6. 可比性

各数据单元是不是可比的，或经过标准化处理后可视为可交换的？苹果和橘子是否被组合在一起？数据单元是不是相互独立的？两列数据是不是同一个变量的副本？

7. 可视化

观察数据（或其子集），制作一维或二维图表，并检验这些数据的统计量。询问数据范围是什么？数据正常吗？是否有缺失值？使用多种颜色和动态图来标明这些问题。是否有意料之外的情况？值得注意的是，我们大脑皮层的30%是用来处理图像的，所以可视化方法在挖掘数据模式和遇到特殊情况时非常有效。通常情况下，为了找到大数据的模式，在某些模型建立之后使用可视化方法最有用，比如计算残差并进行可视化展示。

8. 随机性

统计推断的概念（比如 p 值和置信区间）都依赖于随机性。数据中的随机性是什么含义？我们要使统计模型的随机性尽可能明确。哪些领域知识支持统计模型中的随机性描述？一个表现统计模型中随机性的最好例子是因果关系分析中内曼-鲁宾（Neyman-Rubin）的随机分组原理（在AB检验中也会使用）。

9. 稳定性

你会使用哪些现有的方法？不同的方法会得出同一个定性的结论吗？举个例子，如果数据单元是可交换的，可以通过添加噪声或二次抽样对数据进行随机扰动（一般来说，应确定二次抽样样本遵守原样本的底层结构，如相关性、聚类特性和异质性，这样二次抽样样本才能较好地代表原始数据），这样做得出的结论依然成立吗？我们只相信那些能通过稳定性检验的方法，稳定性检验简单易行，能够抗过拟合和过多假阳性的发现，具有可重复性（要了解关于稳定性重要程度的更多信息，请参见文章“Stability”（http://projecteuclid.org/euclid.bj/1377612862））。

可重复性研究最近在学术界引起很多关注（请参见《自然》（*Nature*）特刊（http://www.nature.com/news/reproducibility-1.17552））。《科学》（*Science*）的主编玛西亚·麦克纳特（Marcia McNutt）指出，“实验再现是科学家用以增加结论信度的一种重要方法”。同样，商业和政府实体也应该要求从数据分析中得出的结论在用新的同质数据检验时是可重复的。

10. 结果验证

如何知道数据分析做得好不好呢？衡量标准是什么？可以考虑用其他类型的数据或者先验知识来验证，不过可能需要收集新的数据。

在数据分析时还有许多其他问题要考虑，但我希望上面这些问题能使你对如何获取“数据智慧”产生一点感觉。作为一个统计学家，这些问题的答案需要在统计学之外获得。要找到可靠的答案，有效的信息源包括“死的”（如科学文献、报告、书籍）和“活的”（如人）。出色的人际交流技能使寻找正确信息源的过程简单许多，即使在寻求“死的”信息源的过程中也是这样。因此，为了获取充足的有用信息，人际交流技能变得更加重要，因为在我的经验中，知识渊博的人通常是你最好的指路人。

第 2 章　线性回归方法

本章介绍最常用的线性回归方法。2.1 节回顾经典多元线性回归的基础内容。为了解决多元回归的多重共线性以及高维问题的自变量选择问题，2.2 节介绍两种压缩回归方法——岭回归和 Lasso 回归。2.3 节进一步介绍求解 Lasso 的最小角回归算法以及 Lasso 解的理论性质。2.4 节给出损失函数加罚的建模框架，并介绍不同损失函数和罚函数组合的各种回归模型。2.5 节给出上机实践的例子。

2.1　多元线性回归

2.1.1　多元线性回归模型

1. 多元线性回归模型及其矩阵表示

设 y 是一个可观测的随机变量，它受到 p 个因素 x_1，x_2，…，x_p（根据具体情况，它们可以是非随机变量，但更多时候是随机变量，此时考虑的是给定 x_1，x_2，…，x_p 情况下，y 的条件分布）和随机因素 ε 的影响，y 与 x_1，x_2，…，x_p 有如下线性关系：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\varepsilon \tag{2.1}$$

式中，β_0，β_1，…，β_p 是 $p+1$ 个未知参数；ε 是不可测的随机误差，服从一定的分布。通常假设服从均值为 0，方差为 σ^2 的分布。若进一步假定服从正态分布 $N(0,\sigma^2)$，则会有更多的结论。我们称式（2.1）为多元线性回归模型，称 y 为被解释变量（因变量），$x_i(i=1,2,\cdots,p)$ 为解释变量（自变量），称 $E(y)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$ 为理论回归方程。

对于一个实际问题，要建立多元回归方程，首先要估计未知参数 β_0，β_1，…，β_p，为此我们要进行 n 次独立观测，得到 n 组样本数据 $(x_{i1},x_{i2},\cdots,x_{ip};y_i)(i=1,2,\cdots,n)$，它们满足式（2.1），即有