大数据应用与技术丛书

Python 商业数据挖掘

(第 6 版)

[美] 盖丽特 • 徐茉莉(Galit Shmueli)
彼得 • C. 布鲁斯(Peter C. Bruce)
彼得 • 戈德克(Peter Gedeck)
尼廷 • R. 帕特尔(Nitin R. Patel) 著
吴文国　金柏琪 译

清华大学出版社
北　京

北京市版权局著作权合同登记号　图字：01-2020-2709

Galit Shmueli, Peter C. Bruce, Peter Gedeck, Nitin R. Patel
Data Mining for Business Analytics: Concepts, Techniques and Applications in Python
EISBN：978-1-119-54984-0

图书在版编目(CIP)数据

Python商业数据挖掘：第6版 / (美)盖丽特 • 徐茉莉(Galit Shmueli) 等著；吴文国，金柏琪译. —北京：清华大学出版社，2021.11
(大数据应用与技术丛书)
书名原文：Data Mining for Business Analytics: Concepts, Techniques and Applications in Python
ISBN 978-7-302-59024-8

Ⅰ.①P…　Ⅱ.①盖…②吴…③金…　Ⅲ.商业信息—数据采集　Ⅳ.①F713.51

中国版本图书馆 CIP 数据核字(2021)第 188712 号

责任编辑：王　军
装帧设计：孔祥峰
责任校对：成凤进
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市科茂嘉荣印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张**：27.75　　**字　　数**：801 千字
版　　次：2021 年 11 月第 1 版　　**印　　次**：2021 年 11 月第 1 次印刷
定　　价：118.00 元

产品编号：087711-01

序言一

统计学领域已经存在了 200 多年。到了 20 世纪 50 年代，统计学已发展成为一门颇受人们重视的基础学科。它的重要性在 20 世纪 90 年代随着新型巨量数据源的出现而迅速增加。在 21 世纪的前十年，人们的注意力都被吸引到生物应用上，特别是由于人类基因组测序出现的基因数据。但最近十年来，人们发现在商业领域中可用的数据急剧增加，因而人们对统计学在商业领域中应用的兴趣也急剧提升。

十年前，当我的统计学选修课吸引攻读 MBA 的全部同学时，我的同事感到十分震惊，因为那时我们学院正在为选修课能否吸引更多学生而苦恼。现在，我们开设了商业应用硕士学位课，它已成为我们学院最大的专业硕士项目。随之，我们学院的教师人数和提供的课程数也急剧增加。Google 的首席经济师 Hal Varian 在 2009 年提出了一个非常正确的观点：未来最吸引人的职业将是统计学家。

对统计人员的需求是受一个简单而不可否认的事实驱动的。在许多领域和背景下，商业分析解决方案已经给商业业绩带来了非常可观且可以测量的改善效果，因此，需要大量具有必备统计技能的人员。然而，训练学生掌握这些技能却遇到了挑战，因为学生除了需要必备的统计方法知识外，他们还必须了解与商业相关的问题，必须具有沟通技巧且能够熟练使用多个计算包。大多数统计学教材只注重对经典方法的抽象训练，没有强调实际应用，更别提在商业领域的应用了。

本书是我到目前为止见到过的介绍商业分析最全面的图书。其中包括了统计学领域几乎全部的内容——从线性回归和 Logistic 回归等经典方法，到最新的神经网络、装袋法和提升树，甚至介绍商业领域专用的方法，如社交网络分析和文本挖掘。即使算不上“圣经”，也至少是该领域的一本权威手册。本书的各个专题安排恰到好处。更重要的是，这些专题都与商业领域里的应用有关。本书最后一章专门介绍了商业分析方法在 10 个不同领域中的应用案例。

在本书的第 6 版中，作者增加了对 Python 的支持。Python 是一门日益受数据科学家欢迎的编程语言。本书详细介绍了 Python 语言在商业背景中的应用和程序实例，确保读者能够将所学的知识应用到解决实际问题中。我深信本书是任何 Python 商业分析课程必不可少的工具书。

我们学院最近在 MBA 必修基础课中新增了一门商业分析课程，我打算在这门课的教学大纲中大量引用本书内容。我确信，本书是此类课程必不可少的工具书。

Gareth James

2019 年于南加州大学马歇尔商学院

序言二

数据是新的金矿——挖掘这个金矿会给今天高度网络化和数字化的社会创造商业价值，但这需要一组传统商业课程、统计学或工程技术中未曾介绍的技术。在当前大数据背景下，某些商业企业和机构感到压力巨大，这让我想起了一句话——“最坏的情况还未到来”。往昔大数据有三个主要来源：20 多年在企业管理系统上的投入(ERP、CRM、SCM 等)，在线社交网络上的 30 多亿用户，50 多亿高端移动设备用户。与未来受物联网驱动的智能物理生态系统数据源相比，这些数据源简直是小巫见大巫。

用传感器把物理对象(如房屋、汽车、道路甚至垃圾箱、路灯等)连接到数字化控制中心的想法是与更大的“大数据”和更深度的数据分析能力同步发展的。我们离这样的智能冰箱不远了：它能自动检测到冰箱里的某样食物(如鸡蛋)吃完了，自动用手机填写网上食物购买清单，并安排跑腿兔子(Rask Rabbit)把食物送到家；它还能自动与 Uber 出租车司机谈妥一笔生意，把一份晚餐送到你的手里。以下前景也离我们不远了：道路和驾驶车辆中内嵌的传感器能够判断交通拥挤情况、跟踪道路磨损情况以及记录汽车使用情况，并且把这些信息与基于使用量的动态定价系统、保险费率和税务系统关联起来。我们大胆构想的这一新世界需要更好的数据分析方法和更强的数据处理能力的支持。

商业分析是一门新兴学科，它可以帮助我们更好地在这一新浪潮中乘风而起。商业分析这门新学科要求读者对商业基础理论具备坚实的基础，这样读者才能够提出正确的问题，使用、存储和最优地处理各种结构化和非结构化的数据源，以及利用机器学习和统计学中的方法深入理解决策过程。具备这样条件的人是当前社会的稀缺人才，但是如何培养这种人才是本书的重点。本书的特点是通过真实的、数据丰富的且可动手实践的案例来解释当前商业分析领域的核心概念，但是并没有牺牲学术上的严谨性。这是现代商业分析的基础，当前商业分析的基本思想是预测 x 对 y 的贡献。我确信你们中的某些人将会是本书第 10 版的首批读者。

在 2018 年推出 R 语言版后，新增的 Python 版是本书的重要版本之一。Python 语言越来越受到数据分析专业人士的欢迎。这两门开源编程语言已成为数据科学中最重要的统计建模工具和机器学习编程环境。

我期待本书出现在更多的论坛里，出现在经理人的培训课程里、MBA 课堂上、硕士商业分析教学计划中、大数据科学微博里。我相信一定会的！

Ravi Bapna
2019 年于明尼苏达大学卡尔森管理学院

前　言

本书最早出版于 2007 年年初，已被众多学生、从业人员和任课老师采用，包括我本人，在过去 15 年里，在线授课和面对面授课都以本书为重要参考书。本书的第 1 版是基于 Excel 加载项(加载程序是 Analytic Solver Data Mining，早先的名称是 XLMiner)的，此后不断推出 JMP 版本、R 版本和现在的 Python 版本，并推出了本书的合作站点——www.dataminingbook.com。

新推出的 Python 版本使用了免费开源的 Python 程序设计语言。本书提供了 Python 程序的输出结果以及生成这些结果的代码，也包含相关程序包和函数的使用说明，其中的核心是 scikit-learn 包。不同于计算机科学教材或统计学教材，本书的重点在于数据挖掘的基本概念以及如何用 Python 实现相关算法。我们假设读者基本熟悉 Python 语言。

对于新推出的 Python 版本，增加了另一位共同作者——Peter Gedeck，他在商业领域里具有丰富的数据科学经验。除了提供 Python 代码和输出结果外，本书也增加了最新内容和反馈意见。这些意见来自教授 MBA 课程、MS 课程、本科生课程、文凭课程和经理人培训课程的老师及学生。最重要的是，本书首次引入了有关数据伦理的内容(详见 2.9 节)。

本书还包含原书第 3 版新增的如下内容：

- 社交网络分析
- 文本挖掘
- 集成方法
- 增益模型协同过滤

自第 2 版开始(基于 Analytic Solver)，以本书为教材的课程大量增加。最初，本书主要用于一学期的 MBA 选修课，现在已被用在许多商业分析学位课的教学大纲里和证书课程的教学计划里。从本科生教学计划到研究生和经理人培训计划，这些项目里的课程、时间长短不一，深浅不同。在很多情形下，本书可用在多门课程里。本书的设计思想是继续支持通用的“预测分析”或“数据挖掘”课程，但是也支持专用的商业分析教学大纲。

在专用的商业分析教学大纲中，以下课程曾使用本书。

- 预测分析——监督学习：在专用的商业分析项目里，对于预测分析主题，通常包括一系列课程。第一门课程包括本书的第Ⅰ部分至第Ⅳ部分内容。教授这门课程的老师通常根据课时适当地选择第Ⅳ部分的内容。在这类课程中，建议包括第 13 章的集成学习和第Ⅶ部分的数据分析。
- 预测分析——无监督学习：本课程介绍数据探索和可视化、降维、挖掘关系和聚类(第Ⅲ部分和第Ⅴ部分)。如果这门课程也按照“预测分析——监督学习”课程的教学计划，那么有必要分析综合应用无监督学习和监督学习的例子和方法。

- 预测分析：专门用于时间序列预测的课程需要用到第Ⅵ部分的内容。
- 高级分析：本课程综合了全部的预测分析内容(包括监督学习和无监督学习)。这门课程的重点应放在第Ⅶ部分。这部分包含了社交网络分析和文本挖掘。有的老师也会在这类课程中选择第 21 章中的案例。

在以上所有课程中，我们强烈建议增加课程设计项目，要求学生自己收集数据，或利用老师提供的数据(例如，现在有很多供数据挖掘使用的数据集)。根据我们和其他老师的经验，这些项目可让学生巩固所学的知识，并且能给学生提供一个机会，以便更好地理解数据挖掘的强大功能以及在挖掘过程中遇到的问题。

——Galit Shmueli、Peter C. Bruce、Peter Gedeck 和 Nitin R. Patel
2019 年

致　谢

我们要感谢很多人，他们帮助我们不断改进本书。本书经历了很多版本，从 2006 年的最初版，到现在最新的书名《Python 商业数据挖掘(第 6 版)》，中间还有两个 XLMiner 版本、一个 JMP 版本以及一个 R 语言版本，现在首次出现了 Python 版本。

感谢 Anthony Babinec，多年来他在 statistics.com 网站上的在线课程一直使用本书的早期版本，他向我们提供了详尽而宝贵的修改意见。Dan Toy 和 John IV 热心支持我们的项目，并对本书的初稿提供详尽和中肯的评价。Ravi Bapna 先在印度商学院，而后在明尼苏达大学的数据挖掘课里使用本书的初稿。从本书一开始，他就提出了非常有价值的评价和有益的建议。

使用本书早期版本的众多教授、助教和学生，通过直接或间接方式向我们提供了珍贵的反馈意见。他们通过富有成果的讨论、学习之旅和有意义的数据挖掘项目改进本书。具体有马里大学、MIT、印度商学院和 statistics.com 网站的老师和学生。此外还有许多大学和教学机构的老师，他们人数太多，在此无法一一列出，他们从本书的初期，就一直支持和帮助我们提高质量。Scott Nestler 从一开始就是本项目的益友。

statistics.com 网站的教学主管 Kuber Deokar，毫不吝啬地支持、帮助和关心本书的出版。我们也感谢 statistics.com 网站的教学助理 Anuja Kulkarni。Valerie Troianoc 通过 statistics.com 网站帮助了许多老师和学生，他也帮助本书走向成熟。

同事和家人一直以来也在向本书提供反馈意见和帮助。Boaz Shmueli 和 Raquelle Azran 给本书的最早两个版本提供排版意见和建议。Bruce McCullough 和 Adam Hughes 给第 1 版提供排版建议。Noa Shmueli 仔细校对了本书的第 3 版。Ran Shenberger 提出了本书版式设计的意见。Che Lin 和 Boaz Shmueli 提供了有关深度学习的反馈意见。Ken Strasma(HaystaqDNA 公司的创始人，2004 年克里竞选活动和 2008 年奥巴马竞选活动的负责人)提供了 13.2 节“增益(说服)模型”中所讲内容的背景和数据。感谢 Jen Golbeck(马里兰大学社智能实验实验室主任，*Analyzing the Social Web* 一书的作者)，他的著作启发我们在本书中增加有关社交网络分析的内容。Randall Pruim 对本书有关可视化方面的内容具有重大贡献。Inbal Yahav——R 语言版的共同作者，帮助改进了社交网络分析和文本挖掘方面的内容。

有关时间序列的内容借鉴了得克萨斯 A&M 大学 Marietta Tretter 的评论和思想。数据可视化方面的内容和本书的整体版式设计借鉴了 Stephen Few 和 Ben Shneiderman 的反馈意见和建议。

Susan Palocsay、Mia Stephens 和 Margret Bjarnadottir 提供了许多宝贵的意见和建议。特别是针对 Python 版本，Margret Bjarnadotti 提供了许多宝贵的建议。此外，还要感谢马里兰大学人机交互实验室的 Catherine Plaisant，他对第 3 章“数据可视化”中的习题和插图有很大的贡献。Gregory Piatetsky-Shapiro 是 KDNuggets.com 网站的创始人，在最初几年，他为本项目花费了许多时间，并提出

了许多有价值的建议。

感谢麻省理工学院斯隆管理学院的同事们，在本书的孕育期就得到他们的支持，他们是 Dimitris Bertsimas、James Orlin、Robert Freund、Roy Welsch、Gordon Kaufmann 和 Gabriel Bitran。作为斯隆管理学院的教学助理，Adam Mersereau 对本书的起源——以前的教学笔记和案例提出了详尽的意见。Romy Shioda 帮助我们准备了本书的几个案例和习题，Mahesh Kumar 分享了聚类分析方面的内容。

感谢马里兰大学史密斯商学院的同事们：Shrivardhan Lele、Wolfgang Jank 和 Paul Zantek，他们提供了非常实用的建议和评价。感谢 Robert Windle 和马里兰大学的 MBA 学生：Timothy Roach、Pablo Macouzet 和 Nathan Birckhead，他们提供了宝贵的数据集。感谢 MBA 学生 Rob Whitener 和 Daniel Curtis，他们贡献了热图和地图。

Anand Bodapati 提供了部分数据集和建议，微软研究院的 Jake Hofman 和 Sharad Borle 帮助我们存取数据。Suresh Ankolekar 和 Mayank Shah 帮助我们开发了几个案例，并提供了一些宝贵的教学评论。Vinni Bhandari 帮助我们撰写查尔斯图书俱乐部案例。

感谢哈佛大学的 Marvin Zelen、L. J. Wei 和 Cyrus Mehta 以及普纳大学的 Anil Gore，他们就统计学与数据挖掘的关系提出了发人深省的讨论意见。也感谢麻省理工学院系统工程部的 Richard Larson，他就数据挖掘在复杂系统模型中的作用提出了真知灼见的想法。在 20 多年前，他们帮助我们在新兴的数据挖掘领域里提出了一种平衡的哲学观点。

最后，我们要感谢 Wiley 出版社的编辑，感谢他们对本书十多年的漫长之路给予的支持和帮助。Steve Quigley 从一开始就显示出对本书的信心，并帮助我们以极快的速度完成出版过程。Curt Hinrich 的视野、提示和参与，帮助我们立刻投入本书。有了 Sarah Keegan、Mindy Okura-Marszycki、Jon Gurstelle、Kathleen Santoloci 和 Katrina Maceda 等人的不断鼓励，本书的 Python 版才最终得以出版。我们还要特别感谢 Amy Hendrickson，因为有她帮助我们排版，才有这本精美的图书。

目 录

第 I 部分

预 备 知 识

第 1 章

引　言

1.1　商业分析简介

商业分析(Business Analysis，BA)是利用量化数据做出决策的实践和艺术。这个术语对于不同的机构有不同的含义。

考虑商业分析在新闻机构数字化转型过程中所起的作用。有一份英国小报，它的读者群面向工薪阶层；该报纸推出了网络版，并在首页上做了一项测试，以了解哪些图片的点击量最大：猫、狗还是猴子。这家报纸的这个简单应用可以视为“商业分析”。而《华盛顿邮报》的读者群是一些极有影响力的人士，他们是大型国防承包商们感兴趣的对象：因为它可能是唯一可以经常看到航空母舰广告的报纸。在数字环境中，《华盛顿邮报》能够根据一天中的某个时间以及读者的位置和订阅信息来跟踪读者。通过这种方式，该报纸的在线版本可以只面向很小的一个群体——如对五角大楼预算有投票权的美国众议院或参议院军事委员会的成员。

商业分析，或者更一般地称为分析，包含一系列数据分析方法。许多功能强大的应用不过是数据统计、规则检查和简单的算术计算的集合。对于一些机构来说，这正是商业分析的含义。

商业分析的另一层含义，就是现在所谓的“商业智能”(Business Intelligence，BI)。BI 是指能够帮助我们更好洞察“发生了什么，正在发生什么”的数据可视化技术和报表技术。这是通过图表、表格和指示板以及检查和探索数据来实现的。以前的 BI 主要使用静态报表，但现在它已经演化为更友好、更高效的工具和方法，例如创建交互式指示板，使得用户不仅可以实时访问数据，而且能够直接与实时数据进行互动。高效的指示板往往直接绑定到公司的数据上，使得经理们能够快速查看在大型、复杂的数据库中不容易看到的信息。工业运营经理使用的就是这样一个工具，它能在一个二维显示器上显示客户订单，并用颜色和气泡大小作为附加变量，显示客户名称、产品类型、订单数量以及生产时间。

现在的商业分析通常都包括了商业智能和其他高级数据分析方法，如用来探索数据、量化和解释测量值之间的关系并预测新记录的统计模型和数据挖掘算法。回归模型等方法用来描述和量化表示数据品牌间的“平均”关系(如广告和销量的关系)、预测新记录(如某种药物对一名新患者是否有效)以及预测未来值(例如下周的网络流量)。

熟悉本书之前版本的读者可能注意到，本书关注的主题从“数据挖掘与商业智能”变成了“数据挖掘与商业分析”。这种变化反映了最近流行的术语——商业分析，并取代了之前使用的术语——商业智能，用来代表高级分析。如今，商业智能仅指数据可视化和报表生成。

谁在使用预测分析

预测分析的广泛使用，再加上越来越多的可用数据，大大提升了整个经济领域里各个机构的分析能力。

信用评分：在商业领域里，一个早已使用的预测模型技术是信用评分。信用评分并不是随意判断某人的信用度，而是基于一个预测模型，根据之前的数据预测未来的还款行为。

未来购买：最近一个例子(颇具争议)与美国连锁巨头 Target 有关。Target 使用预测模型把潜在客户划分为“怀孕用户”和“非怀孕用户”。对于“怀孕客户”，在其怀孕早期阶段，向她们发送促销信息，这使 Target 在抢夺这个很有购买力的客户群时占尽了先机。

逃税：美国国家税务局发现，当根据预测模型进行执法时，执法人员能够将注意力集中到最有可能发生税务欺诈的行为上，从而把发现偷税行为的概率提高 25 倍(Siegel, 2013)。

商业分析工具盒也包含统计实验，其中 A/B 测试是营销人员最熟悉的方法，它常用于定价政策：

- 旅游网站 Orbitz 发现，在对宾馆进行定价时，对 Mac 用户定的价格可能会比对 Windows 用户定的价格高。
- 史泰博(Staples)公司的网上商店发现，如果顾客住的地方距离史泰博的商店比较远，那么当他们购买订书机时，他们往往可以接受较高的价格。

注意，在公司背景下，商业分析只是问题的一个解决方案。例如，一名经理知道商业分析和数据挖掘是热门领域，所以决定要在自己的机构中部署它们，以找出隐藏在某个地方的潜在价值。要想成功地运用商业分析和数据挖掘，就需要对潜在价值所在的商业背景有深入了解，还需要对数据挖掘方法到底能做些什么有清楚的认识。

1.2 什么是数据挖掘

在本书中，数据挖掘是指超越计数、描述性技术、报表生成和基于业务规则的商业分析等基础性方法的一种高级商业分析方法。虽然本书也要介绍数据可视化方法，但它通常只是学习其他高级商业分析方法的第一步，本书重点放在比较高级的数据分析工具上。具体来说，包括统计方法以及经常以自动方式帮助我们决策的机器学习方法。预测通常是其中的重要组成部分之一。我们感兴趣的并不是“广告与销量有怎样的关系”，而是“此时此刻应该给某个网购者推送什么样的广告或推荐什么样的产品”，或者说我们感兴趣的是把客户划分成不同的角色(personas)，采用不同的营销手段，以及把新客户归到哪一类角色。

大数据时代加速了数据挖掘的应用。数据挖掘方法以其强大的功能和自动性，能够处理海量数据并从中提取商业价值。

1.3 数据挖掘及相关术语

分析领域正在迅猛发展，它的应用范围越来越广，应用高级分析的机构也越来越多。因此，很多术语的定义存在一定的重叠性和不一致性。

对不同的人来说，术语“数据挖掘”具有不同的含义。对一般人来说，下面这个泛化的定义有点模糊且具有贬义：在海量数据(常常是私人数据)中深挖，找出感兴趣的东西。一家大型咨询公司有一个“数据挖掘部门”，但它的主要职责是研究过去的数据并用图表表示它们，以找出总的趋势。但是令人困惑的是，在这家公司里，先进的预测模型属于“高级分析部门”的职责范围。这些机构还用其他

术语表示差不多相同的意思：“预测分析”“预测建模”“机器学习”。

数据挖掘处在统计学和机器学习(也称为“人工智能”)两个领域的交汇点。在统计学领域，用于数据探索和模型构建的各种技术已出现很久，如线性回归、逻辑回归、判别分析和主成分分析法等。但是，经典统计的核心原则——计算困难且缺少数据——并不适用于数据丰富和计算能力超强的数据挖掘应用。

Daryl Pregibon 因而将数据挖掘描述为“规模化、速度化的统计”(Pregibon，1999)。统计学与机器学习两个领域的另一个主要差别是，统计学的重点在于从一个样本推断出总体的“平均效果”。例如，“价格提高 1 美元将导致需求平均减少两箱”。与之相对，机器学习的关注点是预测个体记录，如“当价格提高 1 美元时，对于顾客甲，预测他的需求为 1 箱；而对于顾客乙，预测他的需求为 3 箱”。经典统计学强调的推断技术(判断样本中存在的某个模式或者有价值的结果是否是由偶然因素引起的)不会出现在数据挖掘里。

与统计学相比，数据挖掘以开放的方式处理大数据集，所以无法像推断方法那样对需要解决的问题施加严格的限制。因此，数据挖掘常用的方法很容易出现过拟合，即某个模型与现有的样本数据拟合得非常好，因此不仅描述了数据的结构特征，还描述了数据的随机特性。用工程学术语来讲，此模型不仅拟合了信号，还拟合了噪声。

本书用“机器学习”这个术语表示可以直接从数据中学习到局部模式的算法，特别是指按分层或迭代方式进行学习的算法。与之对应的是，我们用“统计模型”表示把全局结构应用于数据的方法。一个简单例子是线性回归模型(统计模型)与 k 近邻算法(机器学习)。在线性回归模型里，一条记录要服从应用于全部记录的总体线性方程；而在 k 近邻算法中，这条记录要根据自身的几条邻近记录进行分类。

最后，许多从业者，特别是 IT 领域和计算机科学领域的从业者，用术语“机器学习”表示本书介绍的所有方法。

1.4　大数据

数据挖掘和大数据密切相关。大数据是一个相对术语——今天的数据之所以称为大数据，既是相对于过去的数据量，也是相对于处理它们的方法和设备而言的。大数据带来的挑战可以归纳为 4 个 V：体量(volume)、速度(velocity)、多样性(variety)和真实性(veracity)。体量是指数据的量，速度是指数据生成和变化的速度，多样性是指数据的类型非常多(货币、日期、数字、文本等)，真实性是指数据是由有机分布过程(如几百万人注册服务或免费下载)生成的，不同于专为某个研究而收集的数据，这些数据总是受到某些因素的控制或质量检查。

大多数大型机构都需要面对大数据带来的挑战与机遇，因为大多数常规数据处理方法生成的数据都是可以存储和分析的。把传统统计分析(如 15 个变量和 5000 条记录)中的数据与沃尔玛数据库相比较，就能理解大数据的规模。如果将传统统计研究比作表示语句结束的句点，那么沃尔玛数据库就是足球场。这可能还不包括与沃尔玛有关的其他数据，例如非结构化的社交媒体数据。

分析的难度越大，回报也越大：

- 在线约会网站 OkCupid，使用统计模型预测什么形式的消息内容最有可能收到回复。
- 挪威移动电话服务公司 Telenor，使用模型预测哪些客户最可能流失，并重点关注这些客户，因而把用户流失率降低了 37%。
- 美国保险公司 Allstate，在保险数据中加入了更多的车辆类型信息，因而把汽车理赔中伤害责任险的预测准确度提高到了原来的三倍。

以上例子都来自 Eric Siegel 的著作——《预测分析》一书。

一些非常有价值的任务，在大数据时代之前是无法实现的，例如 Google 公司的技术基础——Web 搜索技术。在早期，当搜索 Ricky Ricardo Little Red Riding Hood 关键词时，会找到与 I Love Lucy 电视剧有关的链接，或者与 Little Red Riding Hood(20 世纪 50 年代风靡美国的情景喜剧)有关的各种链接，或者与 Richardo 作为乐队领队有关的链接，以及与儿童故事 Ricky Ricardo Little Red Riding Hood 有关的链接。只有当 Google 的数据库收集到足够多的数据(包括用户的点击记录)时，排在搜索结果最前面的才会是 I Love Lucy 电视剧的链接，因为在这个混合喜剧里，Ricky 扮演了 Ricky Ricardo Little Red Riding Hood 角色。

1.5 数据科学

大数据无所不在，其规模、价值和重要性不言而喻，并由此出现一种新的职业：数据科学家。数据科学是统计学、机器学习、数学、程序设计、商业和 IT 的综合。大数据这个术语本身包含了相比上述其他学科更广的内容。很少有人在以上所有领域都有很深的造诣。在 *Analyzing the Analyzers* (Harris et al., 2013)一书里，作者把大多数数据科学家所具备的技术集描写为 T 形状——表示在某一领域里有很深的功底(相当于 T 的竖线)，而在其他领域里比较肤浅(相当于 T 的顶部水平线)。

在一场数据科学的大型会议期间(Strata + Hadoop World, October 2014)，大多数参会人员都认为，编程是一项关键技术，但是也有一部分人不这样认为。大数据是数据科学发展的驱动力，但是大多数数据科学家实际上并没有把他们的大部分时间花在太字节规模或更大规模的数据上。

如此规模的海量数据只有在模型的部署阶段才会遇到。也只有在这个阶段才会遇到很多挑战，其中大多数与 IT 和编程问题有关，涉及数据处理以及如何把各个不同组件集成到一个系统中。大多数工作必须在这个阶段之前完成。这些早期的规划和原型阶段正是本书的重点——如何使我们开发的统计和机器学习模型最终能融入部署的系统里？针对不同类型的数据和问题，要用什么样的方法？这些方法如何起作用？它们有哪些要求、优点和缺点？如何评估它们的性能？

1.6 为什么有这么多不同的方法

读者从本书或者其他与数据挖掘有关的资源可以看出，预测和分类有很多不同方法。读者可能会问，为什么有这么多方法共存？是否某些方法比其他方法更好？以上问题的答案是，每个方法都有优点，也都有缺点。一个方法是否有用，与数据量、数据中存在的模式、数据是否满足这个方法的某些重要假设、数据中存在多少噪声、分析的特定目的等因素有关。图 1.1 是一个很好的例子，它试图根据家庭收入水平和住房面积大小等综合因素把购买驾驶式割草机的家庭(黑色实心圆)和没有购买驾驶式割草机的家庭(灰色空心圆)区分开来。第一个方法(左图)试图寻找水平线和竖直线把这两类客户分开，而第二个方法(右图)则试图寻找一条对角线把他们分开。

不同的方法会得到不同的结果，它们的性能也不一样。因此，在数据挖掘领域，行业的习惯性做法是先试用几种不同的方法，而后选择一种看起来比较容易实现当前目标的方法。

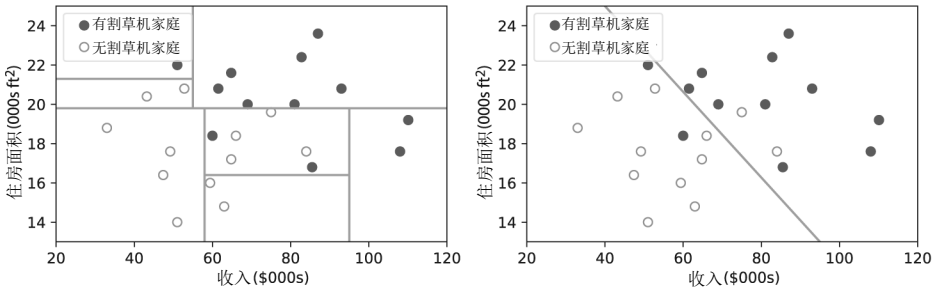

图 1.1　用两种不同方法区分有割草机家庭与无割草机家庭

1.7　术语与符号

由于数据挖掘的混合性，它的从业人员经常用不同的术语表示同一个东西。例如，在机器学习(人工智能)领域，把需要预测的变量称为结果变量或目标变量。在统计学家看来，它们是因变量或响应变量。下面归纳了这些术语。

- 算法(Algorithm)：实现某个特定数据挖掘方法的具体过程，如分类树、判别分析等。
- 属性(Attribute)：参考预测变量。
- 实例(Case)：参考观测对象。
- 分类变量(Categorical Variable)：只能取某几个固定值中的一个。例如，航班的飞行状态只能是准时、延误或取消。
- 置信(Confidence)：表示“当某人购买了 A 和 B 时，也可能会购买 C”这种关联规则的性能指标。置信是一种条件概率，表示当购买了 A 和 B 时，购买 C 的概率。置信在统计学中有着更广泛的意义(如置信区间)，表示对选择不同样本的结果进行估计时出现的误差程度。
- 因变量(Dependent Variable)：参考响应变量。
- 估计(Estimation)：参考预测变量。
- 因子变量(Factor Variable)：参考分类变量。
- 特征(Feature)：参考预测变量。
- 保留数据或保留集(Holdout Data 或 Holdout Set)：它们不是用来拟合模型的，而是用来评估模型性能的样本数据。本书使用验证集和测试集，不使用保留集这个术语。
- 输入变量(Input Variable)：参考预测变量。
- 模型(Model)：应用于数据集的算法，模型要包括参数设置(许多算法都提供了参数，允许用户进行调整)。
- 观测对象(Observation)：数据分析的客体，可从中获取测量数据(如一位顾客、一笔交易等)，它还有其他叫法，如实例、样本、样品、案例、记录、模式、行。在电子表格里，每一行通常代表一条记录，每一列代表一个变量。注意这里的“样本”不同于统计学中的含义。在统计学中，“样本”代表若干观测对象的集合。
- 结果变量(Outcome Variable)：参考响应。
- $P(A|B)$：条件概率，表示事件 A 在事件 B 已发生前提下的概率。可以读成“给定事件 B，事件 A 发生的概率”。
- 预测(Prediction)：对连续结果变量的值进行预测，也叫估计。

- 预测变量(Predictor)：一种变量，通常用 X 表示，它是预测模型的输入变量，也叫特征量、输入变量或自变量。从数据库角度，也可称为字段。
- 剖面(Profile)：从某个观测对象上得到的一组测量值(如某个人的身高、体重和年龄)。
- 记录(Record)：参考观测对象。
- 响应(Response)：一种变量，通常用 Y 表示，在监督学习中则是被预测的变量，也称因变量、输出变量、目标变量、结果变量。
- 样本(Sample)：在统计学领域，样本表示一组观测对象的集合；而在机器学习中，样本代表单个观测对象。
- 分数(score)：预测值或预测类。给新数据计分是指使用训练数据集得到的模型预测新数据的结果。
- 成功类(Success Class)：二元结果中重要的类别(例如，表示购买/非购买结果的购买者)。
- 监督学习(Supervised Learning)：向算法(如 Logistic 回归或回归树等)提供的记录已有结果量，算法通过学习来预测结果量未知的新记录。
- 目标变量(Target)：参考响应。
- 测试数据或测试集(Test Data 或 Test Set)：在模型构建过程中的最后阶段或者评估模型的最终性能并借此选择模型时使用的数据集。
- 训练数据或训练集(Training Data 或 Training Set)：在模型拟合过程中使用的数据集。
- 无监督学习(Unsupervised Learning)：无监督学习是指从数据中学习模式，而不是预测我们感兴趣的输出值。
- 验证数据或验证集(Validation Data 或 Validation Set)：这部分数据用来评估模型的拟合程度，或者用来调整模型，或者从多个试用模型中选择最佳模型。
- 变量(Variable)：记录的任何测量结果，既包括输入变量(X)，也包括输出变量(Y)。

1.8 本书的线路图

本书涵盖了许多广泛使用的预测方法和分类方法，还包括了数据挖掘方法。图 1.2 概括了数据挖掘的主要步骤和对应的章号，主题后面的圆括号中的数字表示章号。表 1.1 从另一个角度——按数据类型和数据结构——介绍了数据挖掘的主要步骤。

本书分为 8 大部分。第Ⅰ部分(第 1 和 2 章)概述了数据挖掘及相关内容。第Ⅱ部分(第 3 和 4 章)分析了早期阶段的数据探索和降维方法。

第Ⅲ部分(第 5 章)讨论了性能评价问题。虽然只有一章，但是这一章讨论了很多内容，从预测性能指标到误分类成本。这一部分介绍的一些原理对于监督学习方法的正确评估和不同监督方法的比较至关重要。

第Ⅳ部分包括 8 章内容(第 6~13 章)，介绍了许多常用的监督学习方法(用于分类和预测)。在这一部分，我们按算法的复杂性、普及程度和理解的难易程度组织内容。第 13 章对前面介绍的全部方法进行了组合。

第Ⅴ部分重点介绍挖掘数据之间的关系，此外还讨论了关联规则和协同过滤(第 14 章)，第 15 章介绍了聚类分析。

第Ⅵ部分包括 3 章内容(第 16~18 章)，重点放在了时间序列预测上。第 16 章介绍与时间序列有关的一般性问题，第 17 和 18 章介绍两种流行的预测方法：基于回归的预测和光滑法。

第Ⅶ部分(第 19 和 20 章)从更广的范围讨论两个数据分析主题：社交网络分析和文本挖掘。它们

能把数据挖掘方法应用于专用的数据结构：社交网络和文本。

最后的第Ⅷ部分介绍了几个案例。

虽然读者可以按本书的章节顺序阅读每一章的专题内容，但是也可以单独阅读每一章。建议读者在开始阅读第Ⅳ部分和第Ⅴ部分之前，务必阅读第Ⅰ、第Ⅱ、第Ⅲ部分的内容。同样在第Ⅵ部分，建议先阅读第 16 章，再阅读随后的第 17 和 18 章。

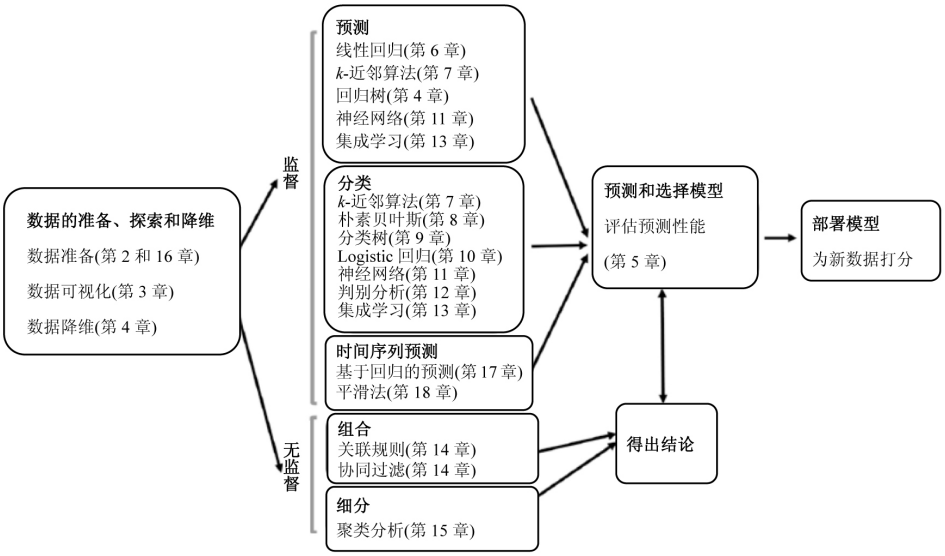

图 1.2　数据挖掘过程

表 1.1　本书根据数据性质对数据挖掘方法做了分类*

	监督学习		无监督学习
	连续响应	分类响应	无响应
连续预测变量	线性回归(6)	*k*-近邻算法(7)	主成分(4)
	k-近邻算法(7)	Logistic 回归(10)	协同过滤(14)
	神经网络(11)	神经网络(11)	聚类分析(15)
	集成算法(13)	判别分析(1)	
		集成算法(13)	
分类预测变量	线性回归(6)	朴素贝叶斯(8)	关联规则(14)
	回归树(9)	分类树(9)	协同过滤(14)
	神经网络(11)	Logistic 回归	
	集成算法(13)	神经网络(11)	
		集成算法(13)	

*圆括号里的数字表示章号。

使用 Python 和 Jupyter Notebook

Python 是一门功能强大的通用编程语言，它有很多应用：从简短的脚本程序到企业级的大型应用程序。现在有很多可用于科学计算的免费开源库和工具，而且数量不断增多。要想进一步了解 Python 及其应用，请访问 www.python.org。

为了方便读者实际体验数据挖掘过程，本书使用了 Jupyter Notebook(参考图 1.3 中的例子)。它们提供了一个交互式计算环境。在这个环境里，我们很容易把代码的执行、丰富的文本格式和图形等集为一体。

Python 语言被广泛应用于学术界和产业界，而且在数据挖掘和数据分析等领域越来越受欢迎，这是因为用户很容易获得可用于数据分析、数据可视化和机器学习等各个领域的程序包。其中涵盖了统计和数据挖掘领域里的各种技术——分类、预测、关联分析和文本挖掘、时间序列预测、数据探索和数据压缩，此外还提供了各种监督数据挖掘工具——神经网络、分类和回归树、*k*-近邻分类、朴素贝叶斯、Logistic 回归、线性回归和判别分析，这些都可以用于预测模型。Python 提供的程序包还涉及无监督算法——关联规则、协同过滤、主成分分析、*k*-均值聚类和层次聚类以及可视化工具和数据处理工具。同一个方法可能会在多个包中得到实现，正如我们在本书里不断提到的那样。本书的示例、习题和案例都是用 Python 语言实现的。

- 下载：对于 Python 和 Jupyter Notebook，建议使用 anaconda 访问 www.anaconda.com，按网站上的指示进行下载。
- 安装：请安装 anaconda 和 anaconda-navigator，后者用来安装和更新单个 Python 包。
- 使用：可从 anaconda-navigator 启动 Jupyter，也可从命令行启动 Jupyter。

要想获得有关本书最新及更全面的信息，请访问 www.dataminingbook.com。

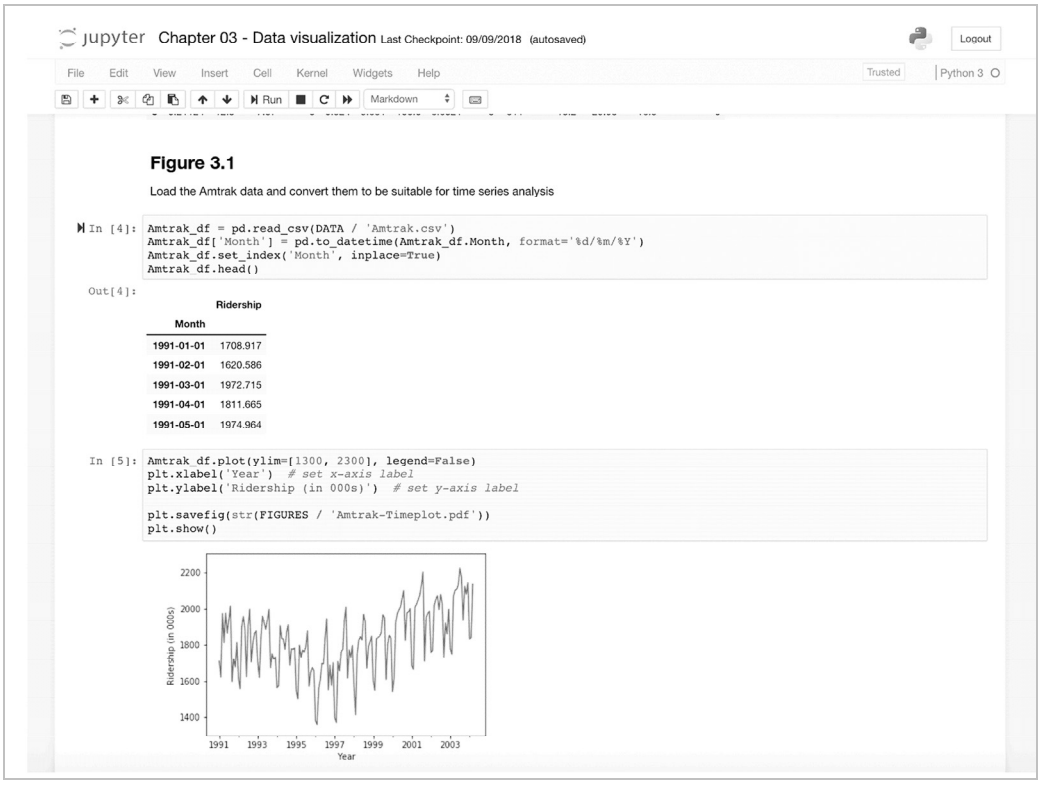

图 1.3　Jupyter Notebook 的工作界面

第2章

数据挖掘过程概述

本章概述数据挖掘的主要步骤。数据挖掘过程起始于目标明确的定义，终止于模型的部署。图 2.1 给出了数据挖掘的主要步骤。此外，本章还介绍与数据收集、数据清理和预处理等有关的一些问题。我们引入了数据分割的概念，分割生成的训练集用来训练模型，分割生成的另一个数据集(验证集)用来评价模型的性能，同时也解释为什么这样做有助于避免过拟合的出现。最后，我们用一个实例来说明模型的构建步骤。

图 2.1　数据建模过程的流程图

2.1　引言

在第 1 章，我们已经介绍了数据挖掘的几个通用定义。在本章，我们还将论述常被称为数据挖掘的各种方法。本书的核心内容是预测分析、分类和预测任务、模式发现。这些内容已成为绝大多数大公司商业分析的核心内容。

本书并未详细讨论的两个比较简单的数据库方法是联机分析处理(Online Analytical Processing，OLAP)和 SQL(Structured Query Language，结构化查询语言)。这两个方法也常被当作数据挖掘方法。OLAP 和 SQL 对数据库的搜索在本质上是属于描述性的，而且基于由用户定义的业务规则(例如，find all credit card customers in a certain zip code with annual charges > $20,000, who own their home and who pay the entire amount of their monthly bill at least 95% of the time，翻译为中文就是，在某个邮编区域里找到所有花费大于 20 000 美元、拥有自住房屋且每月全额还款超过 95%的信用卡用户)。虽然 SQL 查询常用来获取数据挖掘所需要的数据，但是它们并没有涉及统计建模和自动算法。

2.2　数据挖掘的核心思想

2.2.1　分类

分类可能是数据分析的最基本形式。收到录用通知书的人会有两种反应：接受或不接受。贷款的申请人存在三种情形：按时还款、延迟还款或宣布破产。信用卡交易可以是正常交易或欺诈交易。网络上传输的数据包可能包含良性内容或有害内容。车队里的巴士可以提供服务或者不能提供服务。患

者可能已经恢复、继续治疗或已病故。

数据挖掘的常见任务是分析分类未知的数据或者发生在未来的数据，试图预测数据中包含哪些类型或者数据属于哪一种类型。从分类已知的相似数据中找出规则，然后把它们应用于分类未知的数据。

2.2.2 预测

预测与分类相似，只是预测的对象是数值型变量(如采购金额)而非分类变量(如购买者或非购买者)。当然，用分类算法也可以预测类别，但是本书所说的预测专指连续数值型变量的预测(在某些数据挖掘文献中，估计和回归用来代表连续变量的预测，而预测既指连续数据，也指分类数据)。

2.2.3 关联规则与推荐系统

包含客户交易记录的大型数据库凭借自身的优势，对客户选购的商品进行关联分析，即分析“哪些商品会与哪些商品一起购买”。关联规则或亲和性分析，专门用来在大型数据中找到商品间普遍存在的关联模式。例如，杂货店可以利用关联信息合理放置商品。他们可以利用这些规则举行每周一次的促销活动或者将商品搭配销售。医院可以从患者住院数据库推导出关联规则，并利用这些关联规则找到“在有了什么样的症状后有可能会出现什么症状”的答案，因而可以预测患者再次住院的症状。

Amazon.com 和 Netflix.com 等网站的在线推荐系统采用了协同过滤技术。这是一种根据客户在过去购物历史中表现出来的偏好和口味、评分、浏览记录和其他表示个人偏好的度量指标实现推荐服务的方法。与关联规则相对应的是协同过滤。前者生成适用于总人口的关联规则，后者则在个体级生成“什么商品与什么商品搭配”等规则。因此，协同过滤常用在推荐系统中，目的是给偏好广泛的用户提供个性化的推荐服务。

2.2.4 预测分析

分类和预测在某些情况下还包括关联规则和协同过滤，这些技术都可以用在预测分析中。预测分析这个术语有时也包括聚类划分等模式识别方法。

2.2.5 数据规约与降维技术

当变量的个数受到限制时，或者当大量记录被划分为同质组时，数据挖掘算法的性能通常会得到改善。例如，分析师不会直接处理成千上万个不同类型的产品，而是把它们分类成数量较少的组，然后为每个组单独建立模型。营销主管可能希望把客户分成不同的角色，因为只有在把客户分类为同质组时才能定义角色。把大量的记录(或实例)整合成较小数据集的过程称为数据规约(Data Reduction)，通常把减小实例数量的方法称为聚类(clustering)。

通常把压缩变量个数的方法称为降维。降维是在部署数据挖掘方法之前十分常用的初始步骤，旨在提高预测能力并改善可管理性和可解释性。

2.2.6 数据探索和可视化

数据处理早期阶段的一个步骤是探索数据。探索常用于数据清理和变换，也用于可视发现和"假设生成”。

用于数据探索的方法包括各种数据汇总和摘要，不管是数值形式还是图形形式。数据探索还包括

单独分析每个变量以及变量间的关系。进行数据探索的目的是发现数据中存在的模式和异常现象。使用图形和数据面板的数据探索方法被称为数据可视化或可视化分析。对于数值型变量，可以利用直方图或箱线图了解数据的数值分布，检测奇异值(异常观测对象)，发现其他与分析任务有关的信息。同样，对于分类变量，数据探索也可以使用图表。另外，对于数值型变量，我们可以使用两个变量的散点图，了解它们可能存在的关系，同样可以检测数据的奇异值。在图形中添加颜色和交互导航可以大大改善可视化效果。

2.2.7　监督学习与无监督学习

各种数据挖掘方法之间的根本区别是监督学习和无监督学习。监督学习是一些用于分类和预测的算法。使用监督学习时，数据的结果变量(如已购买或没有购买)必须已知，这样的数据也称为标签数据，因为每条记录都有一个标签(结果值)。训练数据是供分类算法或预测算法学习或训练用的，可通过训练数据获取预测变量与结果变量的关系。当算法经过学习后，就可以应用于其他也包含标签数据的样本(也称为验证数据)，不过在这些数据中，结果值已知，但是故意隐藏起来，用来与预测结果进行比较，以确定算法的性能。如果需要比较多个模型，最好保存第三份样本，这份样本的结果变量也是已知的，把它应用于最终选定的模型，以确定预测效果。现在，最终选定的模型就可以用来预测新实例(结果值未知)的结果值。

简单的线性回归就是监督学习算法的典型代表(读者可能已经在统计学导论课堂上学习过这种算法)。Y 变量(已知)是结果变量，X 变量是预测变量。生成一条回归线，使得实际值与使用这条回归线得到的预测值之差的平方和最小。得到了回归线后，我们就可以用它预测新实例的结果值。

无监督学习算法是指那些可以应用于以下场合的算法：不含需要预测或分类的结果变量。由于没有“学习过程”，因此不需要从已知结果值的实例中进行学习。关联规则、降维方法和聚类技术都是无监督学习算法。

有时监督学习和无监督学习也可以一起使用。例如，首先用无监督学习把贷款申请人分成不同风险等级的组，然后用监督学习对每个组单独进行预测，预测贷款违约风险。

监督学习需要好的监督

在某些情形下，因变量的值(即标签)之所以已知，是因为它们是数据本身不可缺少的一部分。网络日志会记录某个用户是否单击指定的链接。银行记录会显示某笔贷款是否按时还款。但是在另一些情形下，必须通过人工标注方法给因变量设定值，并且需要生成足够多的标签数据去训练模型。电子邮件必须用人工方法标注为垃圾邮件或合法邮件。在法律发现研究中，必须把文档标注为相关文档或非相关文档。不管哪种情况，监督质量不高的话，数据挖掘算法就可能会被数据误导。

2014 年 1 月 5 日，Gene Weingarten 在《华盛顿邮报》上报道了这样一件事情：一条非常怪的短语 defiantly recommend 如何通过自动校正系统悄然进入英语词汇里。defiantly 比起 definitely 更接近另一个很容易拼错的单词 definatly。当用户输入拼写错误的单词 definatly 时，早期的 Google 搜索引擎会将其当作 defiantly。在理想的监督学习模型里，人们要引导自动校正过程，即拒绝 defiantly 这个单词，并替换为 definitely。谷歌的算法应该通过学习知道 definitely 是 definatly 的最佳首选匹配词。

2.3　数据挖掘步骤

本书的重点是理解和使用数据挖掘算法(下面的步骤(4)~(7))。然而，我们发现，商业分析项目中出现的一些最严重错误往往是由于对问题没有理解正确造成的——必须在深入算法细节之前正确地理解

问题本身。下面列出典型的数据挖掘步骤。

(1) 深入分析数据挖掘项目的目的。利益相关者将如何使用项目的结果？项目的结果对谁有影响？数据分析只是一次性投入还是需要持续投入。

(2) 收集分析数据挖掘项目所需要的数据。我们通常需要从大型数据库中选取部分样本数据，供分析使用。样本与总体的关系决定了数据挖掘结果推广到样本外数据的能力。这一步可能需要合并来自多个不同数据库或资源的数据，这些数据可能是公司内部的数据(如顾客过去的购买历史)，也可能来自外部(信用等级)。虽然数据挖掘需要处理非常大的数据库，但是通常只需要分析几千或几万条记录即可。

(3) 探索、整理和预处理数据。这一步能确保数据处于合理状态。缺失数据如何处理？数值是否处于合理的范围内？是否有明显的奇异值？可以用可视化方法检查数据。例如，用一系列的散点图显示每个变量与其他变量的关系。此外，我们需要确保字段的定义、度量单位、时间间隔等属性的一致性。在这一步，经常需要通过已有变量创建新的变量。例如，使用开始时间和结束时间计算持续时间。

(4) 如有必要，压缩数据的维数。降维的主要目的是删除无关的变量或对变量进行变换处理。例如，把"money spent"变换成"spent>$100"和"spent≤$100"或者创建新变量(可创建一个新变量来表示客户至少购买了其中一件商品)。你必须理解每个变量的意义以及每个变量是否有必要都包含在模型里。

(5) 决定数据挖掘的任务(分类任务、预处理任务或聚类任务等)。这一步需要把一般的问题或疑问转换为更具体的数据挖掘问题。

(6) 分割数据(用于监督任务)。如果任务属于监督类型(即分类或预测型)，那么需要随机分割数据集。可以把数据集分割成三个子集：训练集、验证集和测试集。

(7) 选择数据挖掘方法(回归、神经网络、层次聚类等)。

(8) 用算法实现任务。这一步通常是一个交互过程，我们需要尝试各种不同实现方案，并且经常需要尝试同一种算法的不同实现方法(在算法里选择不同的变量或配置参数)。在适当情况下，可利用算法在验证数据上的性能反馈信息改进模型。

(9) 解释算法的结果。这一步需要做出选择，为实际部署模型选择最优的算法。如有可能，还需要把最终选取的算法应用于测试集以确定算法的性能(回想一下，为了调整算法，可能需要在验证集上测试算法，于是验证数据成为拟合过程的一部分，这很可能低估了最终选择的模型在部署过程中产生的错误)。

(10) 部署模型。这一步需要把模型集成为可以运行的系统，并且运行在真实数据上，用于决策或采取行动。例如，可能要把模型应用于潜在客户的购物清单，采取的行动可能是“向预测出的购买金额大于 10 美元的客户发送邮件”。这一步的关键任务是给新记录“计分”，或者使用选择的模型预测每一条新记录的结果变量(“分数”)。

上述这些步骤可以归纳为 SEMMA，这是由 SAS 软件公司最早提出的一个概念。

- Sample(采样)：从数据集中选取一个样本集，并把它分割成训练集、验证集和测试集。
- Explore(探索)：用统计方法和可视化方法探索数据集。
- Modify(调整)：变换变量和插补缺失值。
- Model(建模)：拟合预测模型(例如回归树、人工神经网络)。
- Assess(估计)：用验证集比较各个模型。

IBM 的 SPSS Modeler(原来的 SPSS-Clementine)建模软件也有一个类似的概念，称为 CRISP-DM(CRoss-Industry Standard Process for Data Mining，数据挖掘跨行业标准过程)。这些平台提供的预测建模过程都包含同样的步骤。

2.4　前期步骤

2.4.1　数据集的组织

通常数据集是按以下方式组织和显示的：变量代表列，记录代表行。我们以一个数据集为例来说明数据的组织结构，这个数据集中的内容是 2014 年波士顿 West Roxbury 小区的房价。它有 14 个变量和 5000 条房源记录。数据集里的每一行代表一个房子。第一个房子的估价是 344 200 美元，应交税 4430 美元，面积为 9965 平方英尺(1 平方英尺约 0.093 平方米)，建于 1880 年。在监督学习算法里，其中一个变量是因变量，通常位于数据集的第一列或最后一列。本例的因变量是总房价(TOTAL VALUE)，位于数据集的第一列。

2.4.2　预测 West Roxbury 小区的房价

互联网已经给房地产业带来巨变。现在，房产经纪人都通过网站发布房源和房价，市场上到处都有独立房屋和公寓的估价，甚至出现了通常不在市场上挂牌的住宅单元的估价。在编写本书的时候，Zillow(www.zillow.com)是一家在美国颇受欢迎的在线地产网站。2014 年，Zillow 收购了其最主要的竞争对手 Trulia 公司。截至 2015 年，Zillow 公司网站已成为美国人查看房价的最重要平台，正因为如此，它也成了房产经纪人最主要的在线广告网站。本来房产经纪人可以获得高达 6%的佣金，这每年给他们带来十分可观的盈余，但是现在他们需要给 Zillow 公司交广告费，因此他们的利润大受影响(事实上，这正是 Zillow 公司商业模式的关键——把 6%的佣金从房产经纪人那里转移到自己的腰包里)。

Zillow 公司发布的房产价格被称为 Z 估价(Zestimates)，这些数据直接来源于公开的城市房屋信息，这些信息是评估房产税的依据。因此，房产经纪人想要寻求一种可以取代 Zillow 公司的办法。

一种简单的办法就是使用由市政当局确定的房产估价，但是这些估价不包含房屋的所有附属物，并且没有考虑房屋结构的改变或其他附加物的添加而带来的变化。另外，市政当局使用的评估方法往往不够透明，并不能反映市场的真实情况。但是，城市地产数据可以作为建立模型的基础，之后再在模型中添加其他数据。

现在我们分析波士顿地产评估数据。数据主要来自波士顿，可以用来预测房产价格。WestRoxbury.csv 数据文件包含了 West Roxbury 小区每个业主自住的独立房屋的价格信息，既包括众多预测变量的数据，也包括因变量的数据——房产的评估价(Total Value)。这个数据集一共有 14 个变量和 5802 条记录。表 2.1 详细描述了这个数据集中每个变量的含义[1]。表 2.2 展示了这个数据集的一个样本集[2]。

前面曾提到，在标题行的下面，每一行代表一个房子。例如，第一个房子的评估价是$344 200 美元(TOTAL VALUE)，应交税 4430 美元，面积为 9965 平方英尺，建于 1880 年，上下两层，共有 6 个房间。

1　完整的数据字典可从 https://data.boston.gov/dataset/property-assessment 网站下载，这里修改了几个变量名。

2　数据来自 https://data.boston.gov/dataset/property-assessment 的地产评估文件 FY2014，这些数据已经过简单的整理。

表 2.1　West Roxbury 小区房价数据集中的变量及其说明

变量	说明
TOTAL VALUE	房子的总估价，单位为千美元
TAX	税单上的总税额，等于房子的总估价乘上税率，单位是美元
LOT SQ FT	占地面积(单位为平方英尺)
YR BUILT	房子的建造年份
GROSS AREA	楼面面积，单位为平方英尺
LIVING AREA	居住面积，单位为平方英尺
FLOORS	楼层数
ROOMS	房间个数
BEDROOMS	卧室个数
FULL BATH	全浴室个数
HALF BATH	半浴室个数
KITCHEN	厨房个数
FIREPLACE	壁炉个数
REMODEL	房子的改造时间(Recent/Old/None)

表 2.2　West Roxbury 小区房价数据集中的前 10 条记录

TOTAL VALUE	TAX	LOT SQ FT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BED ROOMS	FUL BATH	HALF BATH	KITCHEN	FIRE PLACE	REMODEL
344.2	4330	9965	1880	2436	1352	2	6	3	1	1	1	0	None
412.6	5190	6590	1945	3108	1976	2	10	4	2	1	1	0	Recent
330.1	4152	7500	1890	2294	1371	2	8	4	1	1	1	0	None
498.6	6272	13,773	1957	5032	2608	1	9	5	1	1	1	1	None
331.5	4170	5000	1910	2370	1438	2	7	3	2	0	1	0	None
337.4	4244	5142	1950	2124	1060	1	6	3	1	0	1	1	Old
359.4	4521	5000	1954	3220	1916	2	7	3	1	1	1	0	None
320.4	4030	10,000	1950	2208	1200	1	6	3	1	0	1	0	None
333.5	4195	6835	1958	2582	1092	1	5	3	1	0	1	1	Recent
409.4	5150	5093	1900	4818	2992	2	8	4	2	0	1	0	None

2.4.3　在 Python 程序中载入并浏览数据

Pandas 支持载入多种格式的数据文件，但通常我们希望使用 csv(comma separated values，逗号分隔值)格式的数据文件。如果数据文件是 xlsx(或 xls)格式，那么可以在 Excel 里把数据保存为 csv 格式，具体方法是选择 File | Save as | Save as type: CSV (Comma delimited) (*.csv) | Save 命令。在 Excel 中处理 csv 文件时，需要注意两点：

1) 在 Excel 里打开 csv 文件时，会自动去掉起前导作用的 0，这会损坏像邮政编码这样的数据。

2) 在 Excel 里保存 csv 文件时，只会保存显示的数字。如果需要精确到小数点后的某几位数字，那么必须确保在保存之前，先显示这些位上的数字。

如果计算机中已经安装了 Python 和 Pandas，而且 WestRoxbury.csv 文件已保存为 csv 格式，那么仅仅执行表 2.3 中的代码就可以把这个数据文件载入 Python 中。

表 2.3　用 Pandas 读取数据文件

打开 Anaconda-Navigator，启动 Jupyter Notebook，这会打开一个新的浏览器窗口。导航到 csv 文件所在的文件夹，打开一个新的 Python Notebook。用户可以把名称"Untitled"改为更具描述性的其他名称，如 WestRoxbury。

把下面的代码粘贴到输入区域，单击 Run 按钮执行程序，系统会在每个输入区域中显示执行结果。如果想看到额外的输出结果，那么需要使用 print 函数。但为了提高可读性，后面的大部分程序都没有使用 print 函数。

载入数据文件并生成子集

```
# Import required packages
import pandas as pd

# Load data
housing_df = pd.read_csv('WestRoxbury.csv')
housing_df.shape # find the dimension of data frame
housing_df.head() # show the first five rows
print(housing_df) # show all the data

# Rename columns: replace spaces with '_' to allow dot notation
housing_df = housing_df.rename(columns={'TOTAL VALUE ': 'TOTAL_VALUE'}) # explicit
housing_df.columns = [s.strip().replace(' ', '_') for s in housing_df.columns] # all columns

# Practice showing the first four rows of the data
housing_df.loc[0:3] # loc[a:b] gives rows a to b, inclusive
housing_df.iloc[0:4] # iloc[a:b] gives rows a to b-1

# Different ways of showing the first 10 values in column TOTAL_VALUE
housing_df['TOTAL_VALUE'].iloc[0:10]
housing_df.iloc[0:10]['TOTAL_VALUE']
housing_df.iloc[0:10].TOTAL_VALUE # use dot notation if the column name has no spaces

# Show the fifth row of the first 10 columns
housing_df.iloc[4][0:10]
housing_df.iloc[4, 0:10]
housing_df.iloc[4:5, 0:10] # use a slice to return a data frame

# Use pd.concat to combine non-consecutive columns into a new data frame.
# The axis argument specifies the dimension along which the
# concatenation happens, 0=rows, 1=columns.
pd.concat([housing_df.iloc[4:6,0:2], housing_df.iloc[4:6,4:6]], axis=1)

# To specify a full column, use:
housing.iloc[:,0:1]
housing.TOTAL_VALUE
housing_df['TOTAL_VALUE'][0:10] # show the first 10 rows of the first column

# Descriptive statistics
```

```
print('Number of rows ', len(housing_df['TOTAL_VALUE'])) # show length of first column
print('Mean of TOTAL_VALUE ', housing_df['TOTAL_VALUE'].mean()) # show mean of column
housing_df.describe() # show summary statistics for each column
```

Pandas 使用数据框架(DataFrame)格式保存来自 csv 文件的数据。如果 csv 文件有列标题，那么它们会自动保存为数据的列名。数据框架是 Pandas 中的最基本对象，它是一种特别有用的数据处理工具。数据框架由行和列组成。行就是每个实例(即每个房子)的观测值，列就是变量(如 TOTAL VALUE、TAX)的值。通过表 2.3 里的程序，读者应该熟悉了在进行数据分析之前所需要执行的基本步骤：确定数据的大小和维数(行数和列数)，浏览全部数据，仅显示所选的行和列，对某些变量进行统计汇总。注意，前面有#符号的内容属于注释。

在 Python 中，切片操作是数据框架和列表的一种很有用的数据访问方法。通常用切片方法返回一个对象，这个对象是某个序列的一部分，或是数据框架的部分列和部分行。例如，TAX[0:4]是变量 TAX 的一个切片，它返回 TAX 的前 5 个值(但严格来说，实际上只返回前 4 个值)。注意，在 Python 语言中，索引是从 0 而不是从 1 开始的。Pandas 通过两个方法来访问数据框架的行数据[1]：loc()和 iloc()。loc()方法更通用，它允许用标签访问数据，而 iloc()方法只允许用整数下标访问数据。为了表示连续几行记录，可以使用切片操作，如 0:9。根据规定，Python 中的切片不包括切片的结束索引。iloc()方法符合这个规定，而 loc()方法不符合这个规定，因为 loc()方法包括切片的结束索引。因此，当使用 iloc()和 loc()方法时，一定要注意两者的这一差别。

2.4.4 Python 包的导入

在表 2.3 中，我们导入了可用于处理数据的 Python 包——Pandas。在本章的后面，我们还会用到其他包。可使用下面几行代码导入这些所需的包。

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression
```

pd、np 和 sm 是数据科学领域十分常用的缩写符。

2.4.5 从数据库获得采样数据

通常，我们不是对整个数据库进行数据挖掘。数据挖掘算法通常会受到各种各样的限制，比如受记录条数或变量个数的限制。这些限制可能是由计算速度、计算能力或软件本身的局限性而引起的。即使受这些限制，许多算法在较小样本集上也会执行得较快。

通常，只需要用几千条记录就可以创建精确的模型。因此，我们希望从一个大型数据库生成一个样本子集，用它建立模型。表 2.4 展示了使用 Pandas 包生成样本子集的程序。

1 在面向对象编程里，方法与函数稍有不同。为简单起见，我们认为方法是与对象(数据框架)密切关联的函数，使用方法可以访问对象的数据。