

大数据应用与技术丛书

Data Mining and Predictive Analytics, Second Edition

# 数据挖掘 与预测分析 (第2版)

[美] Daniel T. Larose 著  
Chantal D. Larose 著  
王念滨 宋敏 裴大茗 译



清华大学出版社

大数据应用与技术丛书

# 数据挖掘与预测分析

(第2版)

[美] Daniel T. Larose  
Chantal D. Larose

王念滨 宋敏 裴大雷

著

译



清华大学出版社

北 京

Daniel T. Larose, Chantal D. Larose  
Data Mining and Predictive Analytics, Second Edition  
EISBN: 978-1-118-11619-7  
Copyright © 2015 by John Wiley & Sons, Inc.  
All Rights Reserved. This translation published under license.

本书中文简体字版由 Wiley Publishing, Inc. 授权清华大学出版社出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

北京市版权局著作权合同登记号 图字: 01-2015-5032

Copies of this book sold without a Wiley sticker on the cover are unauthorized and illegal.

本书封面贴有 Wiley 公司防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

#### 图书在版编目(CIP)数据

数据挖掘与预测分析: 第2版 / (美) 丹尼尔·T.拉罗斯 (Daniel T.Larose) 等著; 王念滨, 宋敏, 裴大茗 译. —北京: 清华大学出版社, 2017

(大数据应用与技术丛书)

书名原文: Data Mining and Predictive Analytics, Second Edition

ISBN 978-7-302-45987-3

I. ①数… II. ①丹… ②王… ③宋… ④裴… III. ①数据采集 IV. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 312858 号

责任编辑: 王 军 于 平

封面设计: 孔祥峰

版式设计: 牛静敏

责任校对: 牛艳敏

责任印制: 李红英

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](mailto:c-service@tup.tsinghua.edu.cn)

质 量 反 馈: 010-62772015, [zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

印 装 者: 清华大学印刷厂

经 销: 全国新华书店

开 本: 185mm×260mm 印 张: 47 字 数: 1144 千字

版 次: 2017 年 2 月第 1 版 印 次: 2017 年 2 月第 1 次印刷

印 数: 1~4000

定 价: 99.80 元

---

产品编号: 064482-01

# 译者序

《数据挖掘与预测分析(第2版)》一书从解决现实世界的问题出发,介绍了当前被广泛应用于现实世界数据集合中的数据挖掘和预测分析技术。本书对数据挖掘与预测分析的讲解是以数据准备、统计分析、分类、关联规则、强化模型性能、案例研究为线索,根据技术的适用情况,结合相应的案例开展研究工作,帮助读者了解并掌握各种算法的操作和细微差异,让读者真正理解算法思想和适用环境。

本书提出的方法和技术全面、深入,几乎涵盖了当前应用中常见的各类挖掘与分析方法。对方法的介绍从概念、算法、评价等部分着手,深入浅出地加以介绍。在介绍方法的章节中增加了R语言开发园地,帮助读者利用R语言开展实际设计和开发工作,获得章节中涉及内容的结果,便于读者掌握所学内容。

本书的第I、IV、VII部分由王念滨翻译,第II、III部分由裴大茗翻译,第V、VI部分由宋敏翻译,王红滨负责R语言开发园地的翻译工作,周连科负责各章练习的翻译工作,博士研究生王瑛琦、何鸣、宋奎勇负责全书图表及附录的翻译工作。另外,硕士研究生孙静、李丝然等参加了本书的校对工作,在此一并致谢。

译者在翻译此书的过程中发现,数据挖掘与预测分析领域的许多术语国内的专家们尚未达成共识,因此在翻译过程中,主要参考了互联网释义。由于本书体量庞大,不少算法的细微之处译者尚未开展深入研究,翻译中的错误和不当之处在所难免,恳请读者批评指正。

译者

# 致 谢

## 致谢——Daniel

---

首先我要感谢我的导师，杰出的统计学教授、康涅狄格大学文理学院副院长 Dipak K. Dey 博士，以及韦斯特菲尔德州立大学数学系统计学教授 John Judge 博士。我将终生感谢你们对我的教诲。我还要感谢我在中央康涅狄格州立大学数据挖掘项目组的同事：Chun Jin 博士、Daniel S. Miller 博士、Roger Bilisoly 博士、Darius Dziuda 博士以及 Krishna Saha 博士。感谢我的女儿 Chantal，感谢我的孪生子 Tristan Spring 及 Ravel Renaissance，感谢你们让我体验到生活的真谛。

Daniel T. Larose 博士  
中央康涅狄格大学，数据挖掘及统计学的教授

## 致谢——Chantal

---

首先我要感谢我的博士生导师、康涅狄格州立大学统计系的杰出教授 Dipak Dey 博士和副教授 Ofer Harel 博士。他们的洞察力及理解力都深深体现在我们令人激动的研究项目及我的博士论文 *Model-Based Clustering of Incomplete Data*(基于模型的不完整数据聚类)中。感谢我的父亲 Daniel，将我带入值得一生探究的数据分析领域，感谢我的母亲 Debra，感谢她对统计学的关注。最后，感谢我的兄弟姐妹们：Ravel 和 Tristan，感谢他们的洞察力、音乐及友谊。

Chantal D. Larose 硕士  
康涅狄格大学统计系

# 前言

## 什么是数据挖掘？什么是预测分析

---

数据挖掘是从大型数据集中发现有用的模式和趋势的过程。

预测分析是从大型数据集中抽取信息以便对未来的情况做出预测和估计的过程。

由 Daniel Larose 和 Chantal Larose 合著的《数据挖掘与预测分析(第 2 版)》一书能够确保读者成为这一前沿且大有前途的领域的专家。

## 为什么需要本书

---

根据 MarketsandMarkets 研究公司的调查,从 2013 年~2018 年,全球大数据市场有望以每年 26% 的速度增长,将从 2013 年的 148.7 亿美元增加到 2018 年的 463.4 亿美元<sup>1</sup>。世界范围内的公司和团体正在学习如何应用数据挖掘和预测分析以增加利润。尚未应用数据挖掘和预测分析的公司将会在 21 世纪经济的全球竞争中落伍。

在大多数领域中,人类都被数据所淹没。遗憾的是,这些花费庞大成本收集得到的数据多数都被遗弃在数据仓库中。问题是,缺乏足够的、受过良好训练的、具备将这些数据转换为人类需要的知识并就此将分类树转换为智慧的分析人员。这也是编写本书的目的所在。

McKinsey Global Institute 报告指出<sup>2</sup>:

公司在利用大数据的技能需求方面将会存在人才短缺现象。从大数据中获取价值的制约主要体现在缺乏必要的人才,特别是缺乏那些掌握统计和机器学习专门知识的人才,缺

---

1 *Big Data Market to Reach \$46.34 Billion by 2018*, by Darryl K. Taft, *eWeek*, [www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html](http://www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html), posted September 1, 2013, last accessed March 23, 2014.

2 *Big data: The next frontier for innovation, competition, and productivity*, by James Manyika et al., McKinsey Global Institute, [www.mckinsey.com](http://www.mckinsey.com), May, 2011. Last accessed March 16, 2014.

乏能够使用从大数据中获取的见识来运营公司的管理人员和分析人员。我们认为对大数据世界开展分析工作的职位比目前能够提供的缺少大约 140 000 ~ 190 000 个。此外,我们认为在美国额外还将需要 150 万位能够提出正确问题并能够有效利用大数据分析结果的管理和分析人员。

本书试图帮助解决数据分析人员短缺的问题。

数据挖掘得到越来越广泛的应用,因为它有助于增强公司从其已有的数据集合中发现有利的模式和趋势的能力。公司和团体花费了大量的金钱,收集到海量的数据,但是未能很好地利用隐藏在其数据仓库中的有价值的和可操作的信息。然而,随着数据挖掘实践变得越来越广泛,无法应用这些技术的公司将存在落后于市场的危险,将逐渐失去市场份额,因为他们的竞争对手都在使用数据挖掘,从而赢得竞争优势。

## 谁将从本书获益

---

《数据挖掘和预测分析(第2版)》一书通过逐步动手解决现实世界的现实问题,介绍了当前广泛运用于现实世界数据集合中的数据挖掘技术,这一方式将吸引管理人员、首席信息官、首席执行官、首席财务官、数据分析人员、数据库分析人员以及其他需要了解新方法以提高投资回报率的群体的注意。

利用《数据挖掘与预测分析(第2版)》,你将学习什么类型的分析能够从数据中发现最有益的知识,同时避免进入可能会导致公司投入大量资金而不能带来相应利益的误区。你将通过真正实践数据挖掘和预测分析来学习数据挖掘和预测分析。

## 危险! 数据挖掘容易被搞砸

---

能够开展数据挖掘工作的新的现有软件平台不断涌现,这将带来新的危险。这些应用处理数据非常方便,强大的数据挖掘算法以黑盒方式嵌入到软件中,导致滥用情况出现的比例更高,从而带来巨大的危险。

简言之,数据挖掘工作不容易做好。将强大的模型应用于海量数据时,一知半解特别危险。例如,对未经过预处理的数据开展分析工作可能会得出错误的结论,或者对数据集采用不适当的分析方法,又或者模型构建基于完全不正确或似是而非的假设之上。如果进行了部署,分析中存在的这些错误可能会让你付出昂贵的代价。《数据挖掘与预测分析(第2版)》一书有助于使你成为一名能够避免进入这些昂贵陷阱的精密的分析人员。

## “白盒”方法

---

### 了解基本算法和模型结构

数据挖掘和预测分析出现问题的症结在于盲目采用“黑盒”方法，避免代价昂贵错误的最佳方法是转而采用“白盒”方法，白盒方法强调要求对软件中基本算法和统计模型结构的了解。

《数据挖掘与预测分析(第2版)》通过如下方式应用白盒方法：

- 明确地揭示为什么需要运用某一特定方法或算法。
- 让读者了解某个算法或方法是如何工作的，采用实例(小型数据集)解释，以便读者逐步了解其中的逻辑关系，从而以白盒方法了解方法或算法的内部工作模式。
- 提供将方法应用于大型、现实世界数据集的实例。
- 通过练习测试读者对概念和算法的理解程度。
- 为读者提供将数据挖掘应用于大型数据集的经验。

### 算法概览

---

《数据挖掘与预测分析(第2版)》将利用小型数据集，指引读者学习各种算法的操作和细微差异，让读者真正理解算法的内部工作情况。例如，在第21章中，我们将逐步利用小型数据库，应用 BIRCH 聚类算法(BIRCH 是层次聚类的一种方法)学习平衡迭代消减和聚类，精确地展示 BIRCH 如何针对数据集选择优化的聚类解决方法。正如我们所知，此类演示是本书针对 BIRCH 算法的独特方法。同样，在第27章中，我们将通过使用选择、交叉和变异操作算子，针对小型数据集逐步发现优化解决方案，以便读者能够更好地理解所涉及的过程。

### 将算法和模型应用到大型数据库

《数据挖掘与预测分析(第2版)》提供了大量将数据分析方法应用于大型数据库的示例。例如，第9章通过利用实际数据库，解析营养等级与谷物含量之间的关系。在第4章中，我们将主成分分析应用于实际的加利福尼亚州的人口普查数据中。所有数据集均可从本书网站 [www.dataminingconsultant.com](http://www.dataminingconsultant.com) 中获得。

### 章节练习：检查并确认读者是否了解了本章内容

《数据挖掘与预测分析(第2版)》一书的各章中包含大约 750 个练习，有助于读者了解自己对各章提供材料的理解程度，并从中体验与数字和数据打交道的乐趣。这些练习包

含概念辨析类型的练习,可帮助读者进一步梳理清楚数据挖掘中某些更具有挑战性的概念;利用数据开展工作的练习,帮助读者将特定数据挖掘算法应用到小型数据集中,从而能够逐步实现较好的解决方案。例如,在第 14 章中,我们要求读者通过该章提供的数据集获得最大后验分类。

## 动手实践：通过实际编写数据挖掘算法学习数据挖掘

本书大多数章节为读者提供了动手实践分析问题,为读者提供了运用新学的数据挖掘专业知识,解决大型数据集实际问题的方法。许多人都喜欢边学边做,而《数据挖掘与预测分析(第 2 版)》为读者提供了一个边学边做的框架。例如,在第 13 章中,读者将采用实际的信用卡审批分类数据集,构建自己的最佳 logistic 回归模型,尽可能利用从该章中学习到的方法,提供对模型强大的、可解释的支持,包括对获取的变量及标识变量的解释。

## 令人兴奋的新主题

---

《数据挖掘与预测分析(第 2 版)》一书还提供大量令人兴奋的新主题,主要包括:

- 通过利用数据驱动的误分类开销实现成本-效益分析
- 独立或多元分类模型的成本-效益分析
- 分类模型的图形化评估方法
- BIRCH 聚类
- 分段模型
- 集成方法: bagging 和 boosting 方法
- 模型投票与趋向平均
- 缺失数据的填补方法

## R 语言开发园地

---

R 语言是一种探索及分析数据集的功能强大的开源语言。使用 R 语言的分析人员可以利用大量免费的程序包、例程和图形用户界面来解决大多数数据分析问题。本书大多数章节中都为读者提供 R 语言开发园地,用 R 语言获得章节中涉及内容的结果,以及部分输出的截图。

## 附录：数据汇总与可视化

---

一些读者可能不大容易理解某些统计和图形化概念,这些概念通常会在统计课程中学习。《数据挖掘与预测分析(第 2 版)》一书提供了介绍常见概念和术语的附录,为读者更好

地理解本书的相关材料奠定基础。

## 案例研究：分析方法汇总

---

《数据挖掘与预测分析(第 2 版)》最后提供了详细的案例研究。通过对案例的研究，读者能够了解怎样将自己从书中学习到的方法融会贯通，以建立可操作的、有益的解决方案。详细的案例研究包括在以下 4 章中：

- 第 29 章 案例研究，第 1 部分：业务理解、数据预处理和探索性数据分析。
- 第 30 章 案例研究，第 2 部分：聚类与主成分分析。
- 第 31 章 案例研究，第 3 部分：建模与评估性能和可解释性。
- 第 32 章 案例研究，第 4 部分：高性能建模与评估

案例研究中包含大量图形、探索数据分析、预测模型、客户分析，并提供针对不同用户需求的解决方案。采用定制的数据驱动成本效益表的模型评估方法，反映分类误差的真正开销，而不是采用常见的诸如总体误差率等评估方法。因此，分析人员能够使用每位客户接触的开销对模型进行比较工作，给予接触客户的数量，预测模型能够实现多少利润。

## 本书组织结构

---

《数据挖掘与预测分析(第 2 版)》一书的组织结构有助于读者直接发现相关的逻辑。共设 32 章，包含 8 个主要部分：

- 第 I 部分是数据准备，包含有关数据预处理、探索性数据分析、降维方法等章节。
- 第 II 部分是统计分析，提供开展数据分析工作常见的经典统计方法，包括单变量统计分析 & 多元变量统计分析、简单及多元线性回归方法、为构建模型准备数据、模型构建等章节。
- 第 III 部分是分类，包含 9 章，是本书涉及内容最多的部分：其中包含  $k$ -最近邻算法、决策树、神经网络、logistic 回归、朴素贝叶斯与贝叶斯网络、模型评估技术、基于数据驱动成本的成本-效益分析、二元及  $k$  元分类模型、分类模型的图形化评估等。
- 第 IV 部分是聚类，包含层次聚类和  $k$ -均值聚类、Kohonen 网络、BIRCH 聚类、度量簇的优劣等。
- 第 V 部分是关联规则，本部分仅包含一章内容，涵盖 A Priori 关联规则以及广义规则归纳。
- 第 VI 部分是模型性能强化，提供细分模型、集成方法：bagging 和 boosting、模型投票与趋向平均等章节。
- 第 VII 部分介绍针对预测建模的其他方法，包括缺失数据填补以及遗传算法等。

- 第Ⅷ部分是案例研究：针对直邮市场的预测响应，包括4章，给出如何从直邮市场营销活动中获取最大利润的完整案例分析方法。

## 软件

---

本书使用的软件包括：

- IBP SPSS Modeler 数据挖掘软件套件
- R 开放源代码统计分析软件
- SAS Enterprise Miner
- SAS 统计分析软件
- Minitab 统计分析软件
- Weka 开放源代码数据挖掘软件

IBM SPSS Modeler 是数据挖掘领域应用最广泛的数据挖掘软件套件，该软件由 SPSS 开发([www-01.ibm.com/software/analytics/spss/products/modeler/](http://www-01.ibm.com/software/analytics/spss/products/modeler/))，本书采用了其基本软件。SAS Enterprise Miner 比 IBM Modeler 功能更强大，但学习该软件比较困难。SPSS 可以获得免费试用版(通过 Google 搜索“spss”即可下载)。Minitab 是简单易用的统计软件包，可以在该公司提供的网站 [www.minitab.com](http://www.minitab.com) 下载试用版。

## Weka：开源软件

---

Weka 机器学习平台是一种基于 GNU 通用公共许可证发布的开源软件，它包括实现多数数据挖掘任务所需要的工具集合。《数据挖掘与预测分析(第2版)》利用 Weka 3.6 开发动手实践、一步一步实例教程等，该软件可从本书的相关网站 [www.dataminingconsultant.com](http://www.dataminingconsultant.com) 获得。读者可以使用 Weka 执行如下类型的分析：logistic 回归(见第13章)、朴素贝叶斯分类(见第14章)、贝叶斯网络分类(见第14章)、遗传算法(见第27章)。有关 Weka 的更多信息可参考 [www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)。作者非常感谢 James Steck 提供了大量的 Weka 实例和练习。James Steck([jame\\_steck@comcast.net](mailto:jame_steck@comcast.net))是2005年康涅狄格州州立中央大学最早获得数据挖掘学科硕士学位的学生之一，也是最早获得研究生学术研究奖的学生。

## 本书网站 [www.dataminingconsultant.com](http://www.dataminingconsultant.com)

---

读者可以获得由 Daniel Larose 和 Chantal Larose 撰写的、Wiley InterScience 出版的数据挖掘书籍相关材料。通过该网站，或扫描本书封底的二维码，可以下载本书用到的大多数数据集，方便读者动手实践开发各种本书提到的分析方法和模型。网站还包括勘误表和比较完整的数据挖掘相关资源，涉及数据集链接、数据挖掘研究组链接以及相关的研究论

文等。

然而,本网站真正强大的原因还在于可供讲授本书的教师使用,提供的资源包括:

- 所有练习的答案,包括动手实践分析。
- 各章的 PPT,可方便教学工作。
- 示例数据挖掘课程项目,由作者亲自编写,可以在你的课程讲授中采用。
- 实际的数据集,可用于课程学习参考。
- 每章所涉及的网络资源。

## 作为教材的《数据挖掘与预测分析(第 2 版)》

---

《数据挖掘与预测分析(第 2 版)》自然适合作为 1 学期或 2 学期课程的课本,2 学期课程内容可分为数据挖掘介绍和中级数据挖掘。教师在授课时可获得如下好处:

- 数据挖掘过程介绍。
- “白盒”方法,强调理解基本算法的结构:
  - 利用玩具数据集讲授算法概览。
  - 将算法应用于大型数据集。
  - 超过 300 幅图、275 张表。
  - 包含 750 道章节练习和动手实践分析。
- 大量令人兴奋的新专题,例如基于数据驱动误分类开销的成本-效益分析。
- 详细的案例研究,有助于融会贯通前 28 章介绍的内容。
- 附录:数据汇总与可视化,包含读者可能比较生疏的统计和图形方面的概念综述。
- 对应 Web 网站,提供了上述内容详细的资源列表。

《数据挖掘与预测分析(第 2 版)》可作为本科高年级或研究生课程内容。若先有选修统计方面的课程更好,但并非必需。读者不需要具备计算机编程经验或数据库的专门知识。

# 目 录

## 第 I 部分 数据准备

|                              |    |
|------------------------------|----|
| 第 1 章 数据挖掘与预测分析概述            | 3  |
| 1.1 什么是数据挖掘和预测分析             | 3  |
| 1.2 需求: 数据挖掘技术人员             | 4  |
| 1.3 数据挖掘离不开人的参与              | 5  |
| 1.4 跨行业数据挖掘标准过程:<br>CRISP-DM | 6  |
| 1.5 数据挖掘的谬误                  | 8  |
| 1.6 数据挖掘能够完成的任务              | 9  |
| 1.6.1 描述                     | 9  |
| 1.6.2 评估                     | 10 |
| 1.6.3 预测                     | 11 |
| 1.6.4 分类                     | 11 |
| 1.6.5 聚类                     | 13 |
| 1.6.6 关联                     | 14 |
| R 语言开发园地                     | 15 |
| R 参考文献                       | 16 |
| 练习                           | 16 |
| 第 2 章 数据预处理                  | 17 |
| 2.1 需要预处理数据的原因               | 17 |
| 2.2 数据清理                     | 18 |
| 2.3 处理缺失数据                   | 19 |
| 2.4 识别错误分类                   | 22 |
| 2.5 识别离群值的图形方法               | 22 |
| 2.6 中心和散布度量                  | 24 |
| 2.7 数据变换                     | 26 |
| 2.8 min-max 规范化              | 26 |
| 2.9 Z-score 标准化              | 27 |

|                            |    |
|----------------------------|----|
| 2.10 小数定标规范化               | 28 |
| 2.11 变换为正态数据               | 28 |
| 2.12 识别离群值的数值方法            | 34 |
| 2.13 标志变量                  | 35 |
| 2.14 将分类变量转换为数值变量          | 35 |
| 2.15 数值变量分箱                | 36 |
| 2.16 对分类变量重新划分类别           | 37 |
| 2.17 添加索引字段                | 37 |
| 2.18 删除无用变量                | 38 |
| 2.19 可能不应该删除的变量            | 38 |
| 2.20 删除重复记录                | 39 |
| 2.21 ID 字段简述               | 39 |
| R 语言开发园地                   | 39 |
| R 参考文献                     | 45 |
| 练习                         | 45 |
| 第 3 章 探索性数据分析              | 49 |
| 3.1 假设检验与探索性数据分析           | 49 |
| 3.2 了解数据集                  | 49 |
| 3.3 探索分类变量                 | 52 |
| 3.4 探索数值变量                 | 58 |
| 3.5 探索多元关系                 | 62 |
| 3.6 选择感兴趣的数据子集作进一步<br>研究   | 64 |
| 3.7 使用 EDA 发现异常字段          | 64 |
| 3.8 基于预测值分级                | 65 |
| 3.9 派生新变量: 标志变量            | 67 |
| 3.10 派生新变量: 数值变量           | 69 |
| 3.11 使用 EDA 探测相关联的预测<br>变量 | 70 |

|                            |           |
|----------------------------|-----------|
| 3.12 EDA 概述 .....          | 73        |
| R 语言开发园地 .....             | 73        |
| R 参考文献 .....               | 80        |
| 练习 .....                   | 80        |
| <b>第 4 章 降维方法 .....</b>    | <b>83</b> |
| 4.1 数据挖掘中降维的必要性 .....      | 83        |
| 4.2 主成分分析 .....            | 84        |
| 4.3 将主成分分析应用于房屋数据集 .....   | 87        |
| 4.4 应提取多少个主成分 .....        | 91        |
| 4.4.1 特征值标准 .....          | 91        |
| 4.4.2 解释变异的比例标准 .....      | 92        |
| 4.4.3 最小共性标准 .....         | 92        |
| 4.4.4 坡度图标准 .....          | 92        |
| 4.5 主成分描述 .....            | 94        |
| 4.6 共性 .....               | 96        |
| 4.7 主成分验证 .....            | 97        |
| 4.8 因子分析法 .....            | 98        |
| 4.9 因子分析法在成年人数据集中的应用 ..... | 99        |
| 4.10 因子旋转 .....            | 101       |
| 4.11 用户自定义合成 .....         | 104       |
| 4.12 用户自定义合成的示例 .....      | 105       |
| R 语言开发园地 .....             | 106       |
| R 参考文献 .....               | 110       |
| 练习 .....                   | 111       |

第 II 部分 统计分析

|                            |            |
|----------------------------|------------|
| <b>第 5 章 单变量统计分析 .....</b> | <b>117</b> |
| 5.1 数据知识发现中的数据挖掘任务 .....   | 117        |
| 5.2 用于估计和预测的统计方法 .....     | 117        |
| 5.3 统计推理 .....             | 118        |
| 5.4 我们对评估的确信程度如何 .....     | 119        |
| 5.5 均值的置信区间估计 .....        | 120        |
| 5.6 如何减少误差范围 .....         | 121        |

|                             |            |
|-----------------------------|------------|
| 5.7 比例的置信区间估计 .....         | 122        |
| 5.8 均值的假设检验 .....           | 123        |
| 5.9 拒绝零假设的证据力度的评估 .....     | 125        |
| 5.10 使用置信区间执行假设检验 .....     | 126        |
| 5.11 比例的假设检验 .....          | 127        |
| R 语言开发园地 .....              | 128        |
| R 参考文献 .....                | 129        |
| 练习 .....                    | 129        |
| <b>第 6 章 多元统计 .....</b>     | <b>133</b> |
| 6.1 描述均值差异的两样例 t-检验方法 ..... | 133        |
| 6.2 判断总体差异的两样例 Z-检验 .....   | 134        |
| 6.3 比例均匀性的测试 .....          | 135        |
| 6.4 多元数据拟合情况的卡方检验 .....     | 137        |
| 6.5 方差分析 .....              | 138        |
| R 语言开发园地 .....              | 141        |
| R 参考文献 .....                | 143        |
| 练习 .....                    | 143        |
| <b>第 7 章 数据建模准备 .....</b>   | <b>145</b> |
| 7.1 有监督学习与无监督学习 .....       | 145        |
| 7.2 统计方法与数据挖掘方法 .....       | 146        |
| 7.3 交叉验证 .....              | 146        |
| 7.4 过度拟合 .....              | 147        |
| 7.5 偏差-方差权衡 .....           | 148        |
| 7.6 平衡训练数据集 .....           | 150        |
| 7.7 建立基线性能 .....            | 151        |
| R 语言开发园地 .....              | 152        |
| R 参考文献 .....                | 153        |
| 练习 .....                    | 153        |
| <b>第 8 章 简单线性回归 .....</b>   | <b>155</b> |
| 8.1 简单线性回归示例 .....          | 155        |
| 8.2 外推的危险 .....             | 161        |

|                                                    |     |                                                             |     |
|----------------------------------------------------|-----|-------------------------------------------------------------|-----|
| 8.3 回归有用吗?系数的确定                                    | 162 | 9.3.8 (在给定 $x_1, x_2, \dots, x_m$ 的情况下)<br>随机选择的 $y$ 值的预测区间 | 223 |
| 8.4 估计标准误差                                         | 166 | 9.4 利用指示变量的包含范畴型预测<br>变量的回归                                 | 224 |
| 8.5 相关系数 $r$                                       | 167 | 9.5 调整 $R^2$ : 惩罚包含无用预测变量<br>的模型                            | 230 |
| 8.6 简单线性回归的方差分析表                                   | 169 | 9.6 序列平方和                                                   | 231 |
| 8.7 离群点、高杠杆率点与有影响的<br>观察点                          | 170 | 9.7 多重共线性                                                   | 233 |
| 8.8 回归方程概括                                         | 178 | 9.8 变量选择方法                                                  | 239 |
| 8.9 回归假设验证                                         | 179 | 9.8.1 有偏 $F$ -检验                                            | 239 |
| 8.10 回归推理                                          | 184 | 9.8.2 前向选择过程                                                | 240 |
| 8.11 $x$ 与 $y$ 之间关系的 $t$ -检验                       | 185 | 9.8.3 反向删除过程                                                | 241 |
| 8.12 回归直线斜率的置信区间                                   | 187 | 9.8.4 逐步选择过程                                                | 241 |
| 8.13 相关系数 $\rho$ 的置信区间                             | 188 | 9.8.5 最佳子集过程                                                | 241 |
| 8.14 给定均值的置信区间                                     | 190 | 9.8.6 “所有可能子集”过程                                            | 242 |
| 8.15 给定随机选择值的预测区间                                  | 191 | 9.9 油耗数据集                                                   | 242 |
| 8.16 获得线性特性的变换                                     | 194 | 9.10 变量选择方法的应用                                              | 243 |
| 8.17 博克斯-考克斯变换                                     | 199 | 9.10.1 应用于油耗数据集的前向<br>选择过程                                  | 244 |
| R 语言开发园地                                           | 199 | 9.10.2 应用于油耗数据集的后向<br>删除过程                                  | 245 |
| R 参考文献                                             | 205 | 9.10.3 应用于油耗数据集的逐步选<br>择过程                                  | 246 |
| 练习                                                 | 205 | 9.10.4 应用于油耗数据集的最佳子<br>集过程                                  | 246 |
| 第9章 多元回归与模型构建                                      | 213 | 9.10.5 Mallows' $C_p$ 统计量                                   | 247 |
| 9.1 多元回归示例                                         | 213 | 9.11 将主成分作为预测变量进行<br>多元回归                                   | 251 |
| 9.2 总体多元回归方程                                       | 218 | R 语言开发园地                                                    | 255 |
| 9.3 多元回归推理                                         | 219 | R 参考文献                                                      | 265 |
| 9.3.1 $y$ 与 $x_i$ 之间关系的 $t$ -检验                    | 219 | 练习                                                          | 265 |
| 9.3.2 营养等级与含糖量之间关系<br>的 $t$ -检验                    | 220 | 第III部分 分类                                                   |     |
| 9.3.3 营养等级与纤维含量之间<br>关系的 $t$ -检验                   | 220 | 第10章 $K$ -最近邻算法                                             | 273 |
| 9.3.4 总体回归模型显著性的<br>$F$ -检验                        | 221 | 10.1 分类任务                                                   | 273 |
| 9.3.5 营养等级与含糖量和纤维<br>含量之间关系的 $F$ -检验               | 222 | 10.2 $k$ -最近邻算法                                             | 274 |
| 9.3.6 特定系数 $\beta_i$ 的置信区间                         | 223 |                                                             |     |
| 9.3.7 (在给定 $x_1, x_2, \dots, x_m$ 的情况下)<br>均值的置信区间 | 223 |                                                             |     |

|        |                               |     |
|--------|-------------------------------|-----|
| 10.3   | 距离函数                          | 276 |
| 10.4   | 组合函数                          | 279 |
| 10.4.1 | 简单权重投票方式                      | 279 |
| 10.4.2 | 加权投票                          | 279 |
| 10.5   | 量化属性的相关性: 轴伸缩                 | 280 |
| 10.6   | 数据库方面的考虑                      | 281 |
| 10.7   | 将 $k$ -最近邻算法用于评估和预测           | 281 |
| 10.8   | $k$ 值的选择                      | 282 |
| 10.9   | 利用 IBM/SPSS 建模工具应用 $k$ -最近邻算法 | 283 |
|        | R 语言开发园地                      | 284 |
|        | R 参考文献                        | 286 |
|        | 练习                            | 286 |
| 第 11 章 | 决策树                           | 289 |
| 11.1   | 决策树是什么                        | 289 |
| 11.2   | 使用决策树的要求                      | 291 |
| 11.3   | 分类与回归树                        | 291 |
| 11.4   | C4.5 算法                       | 297 |
| 11.5   | 决策规则                          | 302 |
| 11.6   | 比较 C5.0 和 CART 算法应用到实际的数据     | 303 |
|        | R 语言开发园地                      | 306 |
|        | R 参考文献                        | 307 |
|        | 练习                            | 308 |
| 第 12 章 | 神经网络                          | 311 |
| 12.1   | 输入和输出编码                       | 312 |
| 12.2   | 神经网络用于评估和预测                   | 313 |
| 12.3   | 神经网络的简单示例                     | 314 |
| 12.4   | sigmoid 激活函数                  | 316 |
| 12.5   | 反向传播                          | 317 |
| 12.6   | 梯度下降法                         | 317 |
| 12.7   | 反向传播规则                        | 318 |
| 12.8   | 反向传播示例                        | 319 |
| 12.9   | 终止条件                          | 320 |

|        |                             |     |
|--------|-----------------------------|-----|
| 12.10  | 学习率                         | 321 |
| 12.11  | 动量项                         | 322 |
| 12.12  | 敏感性分析                       | 323 |
| 12.13  | 神经网络建模应用                    | 324 |
|        | R 语言开发园地                    | 326 |
|        | R 参考文献                      | 328 |
|        | 练习                          | 328 |
| 第 13 章 | logistic 回归                 | 331 |
| 13.1   | logistic 回归简单示例             | 331 |
| 13.2   | 最大似然估计                      | 333 |
| 13.3   | 解释 logistic 回归的输出           | 334 |
| 13.4   | 推理: 这些预测有显著性吗               | 335 |
| 13.5   | 概率比比率与相对风险                  | 337 |
| 13.6   | 对二分 logistic 回归预测的解释        | 339 |
| 13.7   | 对应用于多元预测变量的 logistic 回归的解释  | 342 |
| 13.8   | 对应用于连续型预测变量的 logistic 回归的解释 | 346 |
| 13.9   | 线性假设                        | 351 |
| 13.10  | 零单元问题                       | 353 |
| 13.11  | 多元 logistic 回归              | 355 |
| 13.12  | 引入高阶项处理非线性                  | 359 |
| 13.13  | logistic 回归模型的验证            | 366 |
| 13.14  | WEKA: 应用 logistic 回归的实践分析   | 370 |
|        | R 语言开发园地                    | 374 |
|        | R 参考文献                      | 380 |
|        | 练习                          | 380 |
| 第 14 章 | 朴素贝叶斯与贝叶斯网络                 | 385 |
| 14.1   | 贝叶斯方法                       | 385 |
| 14.2   | 最大后验(MAP)分类                 | 387 |
| 14.3   | 后验概率比                       | 391 |
| 14.4   | 数据平衡                        | 393 |
| 14.5   | 朴素贝叶斯分类                     | 394 |

|                                  |     |                                          |     |
|----------------------------------|-----|------------------------------------------|-----|
| 14.6 解释对数后验概率比.....              | 397 | 第 16 章 基于数据驱动成本的                         |     |
| 14.7 零单元问题.....                  | 398 | 成本-效益分析.....                             | 439 |
| 14.8 朴素贝叶斯分类中的数值型<br>预测变量.....   | 399 | 16.1 在行调整条件下的决策<br>不变性.....              | 439 |
| 14.9 WEKA: 使用朴素贝叶斯开展<br>分析.....  | 402 | 16.2 正分类标准.....                          | 440 |
| 14.10 贝叶斯信念网络.....               | 406 | 16.3 正分类标准的示范.....                       | 442 |
| 14.11 衣物购买示例.....                | 407 | 16.4 构建成本矩阵.....                         | 444 |
| 14.12 利用贝叶斯网络发现概率.....           | 409 | 16.5 在缩放条件下的决策不变性.....                   | 445 |
| R 语言开发园地.....                    | 413 | 16.6 直接成本和机会成本.....                      | 446 |
| R 参考文献.....                      | 417 | 16.7 案例研究: 基于数据驱动误<br>分类成本的成本-效益分析.....  | 446 |
| 练习.....                          | 417 | 16.8 再平衡作为误分类成本的<br>代理.....              | 450 |
| 第 15 章 模型评估技术.....               | 421 | R 语言开发园地.....                            | 452 |
| 15.1 用于描述任务的模型评估<br>技术.....      | 421 | R 参考文献.....                              | 455 |
| 15.2 用于评估和预测任务的模型<br>评估技术.....   | 422 | 练习.....                                  | 455 |
| 15.3 用于分类任务的模型评估<br>方法.....      | 423 | 第 17 章 三元和 $k$ 元分类模型的成本-<br>效益分析.....    | 459 |
| 15.4 准确率和总误差率.....               | 425 | 17.1 三元目标的分类评估变量.....                    | 459 |
| 15.5 灵敏性和特效性.....                | 426 | 17.2 三元分类评估度量在贷款审批<br>问题中的应用.....        | 462 |
| 15.6 假正类率和假负类率.....              | 427 | 17.3 三元贷款分类问题的数据驱动<br>成本-效益分析.....       | 466 |
| 15.7 真正类、真负类、假正类、<br>假负类的比例..... | 427 | 17.4 比较使用/不使用数据驱动误分<br>类成本的 CART 模型..... | 467 |
| 15.8 通过误分类成本调整来反映<br>现实关注点.....  | 429 | 17.5 一般的 $k$ 元目标的分类评估<br>度量.....         | 470 |
| 15.9 决策成本/效益分析.....              | 430 | 17.6 $k$ 元分类中评估度量和数据驱动<br>误分类成本的示例.....  | 472 |
| 15.10 提升图表和增益图表.....             | 431 | R 语言开发园地.....                            | 474 |
| 15.11 整合模型评估与模型建立.....           | 434 | R 参考文献.....                              | 475 |
| 15.12 结果融合: 应用一系列<br>模型.....     | 435 | 练习.....                                  | 475 |
| R 语言开发园地.....                    | 436 | 第 18 章 分类模型的图形化评估.....                   | 477 |
| R 参考文献.....                      | 436 | 18.1 回顾提升图表和增益图表.....                    | 477 |
| 练习.....                          | 437 | 18.2 使用误分类成本的提升图表<br>和增益图表.....          | 477 |