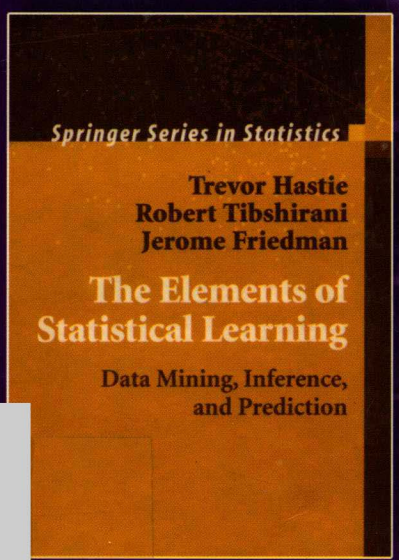
国外计算机科学教材系列

统计学习基础

——数据挖掘、推理与预测

The Elements of Statistical Learning

Data Mining, Inference, and Prediction



Trevor Hastie
[美] Robert Tibshirani 著
Jerome Friedman
范 明 柴玉梅 昝红英 等译

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

国外计算机科学教材系列

统计学习基础

——数据挖掘、推理与预测

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Trevor Hastie
[美] Robert Tibshirani 著
Jerome Friedman

范 明 柴玉梅 昝红英 等译

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

内容简介

计算和信息技术的飞速发展带来了医学、生物学、财经和营销等诸多领域的海量数据。理解这些数据是一种挑战，这导致了统计学领域新工具的发展，并延伸到诸如数据挖掘、机器学习和生物信息学等新领域。许多工具都具有共同的基础，但常常用不同的术语来表达。本书介绍了这些领域的一些重要概念。尽管应用的是统计学方法，但强调的是概念，而不是数学。许多例子附以彩图。本书内容广泛，从有指导的学习（预测）到无指导的学习，应有尽有。包括神经网络、支持向量机、分类树和提升等主题，是同类书籍中介绍得最全面的。

本书可作为高等院校相关专业本科生和研究生的教材，对于统计学相关人员、科学界和业界关注数据挖掘的人，本书值得一读。

版权贸易合同登记号　图字：01-2002-4937

图书在版编目（CIP）数据

统计学习基础——数据挖掘、推理与预测 /（美）黑斯蒂（Hastie, T.）等著；范明等译.
-北京：电子工业出版社，2004.1
（国外计算机科学教材系列）
书名原文：The Elements of Statistical Learning: Data Mining, Inference, and Prediction
ISBN 7-5053-9331-6

Ⅰ.统… Ⅱ.①黑… ②范… Ⅲ.统计学-教材 Ⅳ.C8

中国版本图书馆CIP数据核字（2003）第124311号

责任编辑：杜闽燕
印　　刷：北京兴华印刷厂
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
经　　销：各地新华书店
开　　本：787 × 1092　1/16　　印张：24.75　字数：634千字　彩插：22
印　　次：2004年1月第1次印刷
定　　价：45.00元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系电话：（010）68279077。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。

出 版 说 明

21世纪初的5至10年是我国国民经济和社会发展的重要时期，也是信息产业快速发展的关键时期。在我国加入WTO后的今天，培养一支适应国际化竞争的一流IT人才队伍是我国高等教育的重要任务之一。信息科学和技术方面人才的优劣与多寡，是我国面对国际竞争时成败的关键因素。

当前，正值我国高等教育特别是信息科学领域的教育调整、变革的重大时期，为使我国教育体制与国际化接轨，有条件的高等院校正在为某些信息学科和技术课程使用国外优秀教材和优秀原版教材，以使我国在计算机教学上尽快赶上国际先进水平。

电子工业出版社秉承多年来引进国外优秀图书的经验，翻译出版了“国外计算机科学教材系列”丛书，这套教材覆盖学科范围广、领域宽、层次多，既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。这些教材涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。同时，我们也适当引进了一些优秀英文原版教材，本着翻译版本和英文原版并重的原则，对重点图书既提供英文原版又提供相应的翻译版本。

在图书选题上，我们大都选择国外著名出版公司出版的高校教材，如Pearson Education培生教育出版集团、麦格劳－希尔教育出版集团、麻省理工学院出版社、剑桥大学出版社等。撰写教材的许多作者都是蜚声世界的教授、学者，如道格拉斯·科默（Douglas E. Comer）、威廉·斯托林斯（William Stallings）、哈维·戴特尔（Harvey M. Deitel）、尤利斯·布莱克（Uyless Black）等。

为确保教材的选题质量和翻译质量，我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本系列教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师、博士，也有积累了几十年教学经验的老教授和博士生导师。

在该系列教材的选题、翻译和编辑加工过程中，为提高教材质量，我们做了大量细致的工作，包括对所选教材进行全面论证；选择编辑时力求达到专业对口；对排版、印制质量进行严格把关。对于英文教材中出现的错误，我们通过与作者联络和网上下载勘误表等方式，逐一进行了修订。

此外，我们还将与国外著名出版公司合作，提供一些教材的教学支持资料，希望能为授课老师提供帮助。今后，我们将继续加强与各高校教师的密切联系，为广大师生引进更多的国外优秀教材和参考书，为我国计算机科学教学体系与国际教学体系的接轨做出努力。

电子工业出版社

教材出版委员会

主　任	杨芙清	北京大学教授 中国科学院院士 北京大学信息与工程学部主任 北京大学软件工程研究所所长
委　员	王　珊	中国人民大学信息学院院长、教授
	胡道元	清华大学计算机科学与技术系教授 国际信息处理联合会通信系统中国代表
	钟玉琢	清华大学计算机科学与技术系教授 中国计算机学会多媒体专业委员会主任
	谢希仁	中国人民解放军理工大学教授 全军网络技术研究中心主任、博士生导师
	尤晋元	上海交通大学计算机科学与工程系教授 上海分布计算技术中心主任
	施伯乐	上海国际数据库研究中心主任、复旦大学教授 中国计算机学会常务理事、上海市计算机学会理事长
	邹　鹏	国防科学技术大学计算机学院教授、博士生导师 教育部计算机基础课程教学指导委员会副主任委员
	张昆藏	青岛大学信息工程学院教授

Preface of the Chinese Edition

We are very happy that our book has been translated into Chinese by Professors Fan, Chai and Zan. This means that our work has the possibility of reaching far more people than before—a prospect exciting for any scientist.

We have many Chinese speaking graduate students at Stanford Statistics, and they have assured us that these translation authors have done a very good job.

We take this opportunity to wish all our Chinese colleagues well, and hope they find this text useful.

We also can only hope that our book gets as warm a welcome in the East as it has done in the West.

With best wishes from Trevor Hastie, Rob Tibshirani and Jerome Friedman

Stanford, October 2003

中译本序

我们的书被范明教授、柴玉梅副教授和昝红英讲师翻译成中文，我们感到非常高兴。这意味我们的工作将有机会被更多人所了解——对于任何科学家，这都是令人期待和兴奋的。

在斯坦福大学统计学系，我们有许多讲中文的研究生，他们使我们确信几位译者的翻译非常出色。

借此机会，我们向所有的中国同仁问好，并希望他们喜欢本书。

热切期待我们的书在东方也能像在西方一样受到热烈欢迎。

致以良好祝愿！

Trevor Hastie, Rob Tibshirani, Jerome Friedman

2003 年 10 月于斯坦福

译者序

数据挖掘是一个多学科交叉领域，涉及数据库技术、机器学习、统计学、神经网络、模式识别、知识库、信息提取、高性能计算等诸多领域，并在工业、商务、财经、通信、医疗卫生、生物工程、科学等众多行业得到广泛的应用。在我们已经拥有收集、存储、查询大量数据的手段之后，理解这些数据、从大量数据中发现有用的知识自然成为各行各业的需要，同时也向各领域的研究者提出了新的挑战。需要是发明之母，是科学研究和科学发现的源泉。直面挑战是研究者的本能，也是激发科学研究的巨大动力。

基于知识背景，我们是从数据库或机器学习进入数据挖掘领域的。译者之一曾翻译过 Jiawei Han 和 Micheline Kamber 所著的《数据挖掘：概念与技术》一书。我们为数据挖掘领域的进展感到鼓舞，也在试图为推进数据挖掘的进展贡献微薄之力。随着对数据挖掘理解和研究的深入，我们越来越感到知识海洋的浩瀚。一个问题不断在我们的脑海中浮现：数据挖掘的数学或统计学基础是什么？

曾有几位学者向我们推荐过本书，我们也想认真读一读，但一直未能如愿。当电子工业出版社的编辑希望我们翻译一本关于数据仓库的著作时，我们表示对“The Elements of Statistical Learning: Data Mining, Inference, and Prediction”一书更感兴趣。真是无巧不成书！电子工业出版社买下了这本专著的中文版版权，并在物色翻译人员。当问起是否愿意翻译时，我们欣然同意。

收到本书英文版后，我们迫不及待地打开，真想一口气读完。然而，粗略浏览之后，我们倒吸了一口凉气。放在面前的是斯坦福大学三位统计学家 Trevor Hastie, Robert Tibshirani 和 Jerome Friedman 的力作。想读这本书是一回事，而翻译完全是另一回事。一时间，我们后悔不迭。然而，瓷器活是揽下了，有无金刚钻都只好一试。

用“临时抱佛脚”来形容我们翻译前的准备工作再恰当不过了。我们迅速浏览了一些概率论与数理统计和统计学方面的书籍，并多方寻求帮助；与此同时，我们开始通读原著。最初，进行本书的翻译有些像“逼上梁山”。然而，随着翻译工作的进展，我们才真正感到这是一种享受，因为我们已经被这本书深深地吸引。我们从书中学到了许多，并且可以肯定地说，这本书对我们未来的研究必将产生重要影响。

正如作者所言，本书试图将学习领域中许多重要的新思想汇集在一起，并且在统计学的框架下解释它们。然而，作者强调的是方法和它们的概念基础，而不是数学性质。因此，读者只要学过一门统计学的基础课程，涵盖包括线性回归在内的基本内容，阅读本书就不会有太大的困难。

本书特别适合从事数据挖掘和机器学习研究的读者阅读。尽管作者是统计学家，但他们在过去八年中一直参加神经网络、数据挖掘和机器学习会议，他们的统计学观点可以帮助读者从不同角度更好地理解学习。

全书共 14 章。第 1 章到第 6 章和第 7 章的第 7.1 节至第 7.3 节由范明翻译，昝红英校对；第 7 章其余部分和第 8 章到第 13 章由柴玉梅和王黎明翻译及校对；第 14 章由昝红英翻译，范

明校对。陈国勋认真阅读了全部译文初稿，规范了专业术语的译法并订正了一些错误。范明通读全部译稿，并最后定稿。译者还参照本书 Web 页提供的勘误表，对书中的印刷错误和疏漏进行了更正。

译者感谢电子工业出版社的工作者。在许多出版社都忙于出版“畅销书”的时候，他们坚持引进名著，是他们的远见使得本书中文版能够及时与读者见面。

译者感谢本书的三位作者 Trevor Hastie，Robert Tibshirani 和 Jerome Friedman 教授为中译本撰写序言。当请他们为中译本写序时，他们欣然同意，并坚持要在中译本出版之前先阅读部分译稿。我们按照作者的要求寄去部分章节的译稿，50 天后 Hastie 发来了他们为中译本写的序。当看到三位作者在中译本序中对译文的评价“We have many Chinese speaking graduate students at Stanford Statistics，and they have assured us that these translation authors have done a very good job.”后，我们感觉受到了最高褒奖，同时也被三位作者一丝不苟的科学精神所折服。

这里向为本书翻译做出贡献的所有人表示感谢。这是一本统计学习专著，不仅涉及统计学，而且涉及机器学习、数据挖掘。书中的例子更是取材广泛，涉及医学、生命科学、电子、语音识别等众多领域。郑州大学林诒勋教授、施仁杰教授、陈绍春教授和北京大学计算语言学研究所的于江生博士对一些数学术语的译法提出了宝贵建议；郑州大学医学院张雪培、戴丽萍博士为医学术语的翻译提供了帮助。译者的一些学生也分别阅读了部分译稿，提出了一些有益的建议。还要感谢我们的家人，感谢他(她)们的理解与支持。

作为一本交叉学科的专著，在翻译的过程中，时常需要面对新的知识。尽管我们反复讨论、多次修改，力求译文准确，但仍难免出现差错。此外，由于译者水平有限，译文中的不当之处也在所难免。译文中的错误当然应当由译者负责。但我们真诚地希望同行和读者不吝赐教。如果能把你的意见和建议发往 mfan@zzu.edu.cn，我们将不胜感激。

译　者

2003 年 6 月于郑州大学

前　言

我们被信息淹没，但却缺乏知识。

——Rutherford D. Roger

统计学领域不断受到来自科学界和产业界问题的挑战。早期，这些问题通常来自农业和工业实验，且规模相对较小。随着计算机和信息时代的到来，统计问题的规模和复杂性都有了急剧的增加。数据存储、组织和检索领域的挑战导致一个新领域“数据挖掘”的产生；生物和医学方面的统计和计算问题开创了“生物信息学”。许多领域都产生了海量数据，而统计学家的工作就是理解这些数据：提取重要的模式和趋势，理解这些数据“说瞬么”。我们称此为：从数据中学习。

从数据中学习的难题引发了统计科学的革命。由于计算扮演了重要角色，毫不奇怪，许多成果都是由计算机科学和工程学等其他领域的研究者做出的。

我们考虑的学习问题可以粗略地分为有指导的和无指导的。对于有指导学习，目标是根据一些输入度量预测一个结果度量值。对于无指导学习，没有结果度量，其目标是描述输入度量集合中的关联和模式。

在本书中，我们试图将学习领域中许多重要的新思想汇集在一起，并且在统计学的框架下解释它们。尽管有些数学细节是必要的，但我们强调的是方法和它们的概念基础，而不是理论性质。我们希望本书不仅能吸引统计学家，而且能吸引更广泛领域的研究者和实践者。

正如从统计学之外的研究者那里学到了许多知识一样，我们的统计学观点也可以帮助其他人更好地理解学习的不同方面。

任何事物都没有真正正确的解释，解释是为人们理解而服务的一种媒介。解释的价值是使得他人可以更富有成果地思考。

——Andreas Buja

这里要向为本书的构思和完成做出贡献的所有人员表示感谢。David Andrews，Leo Breiman，Andreas Buja，John Chambers，Bradley Efron，Geoffrey Hinton，Werner Stuetzle 和 John Tukey 对我们的工作具有重要影响。Balasubramanian Narasimhan 为我们提出了许多建议，在一些计算问题上给予了帮助，并维护了一个良好的计算环境。Shin-Ho Bang 帮我们绘制了大量的图形。Lee Wilkinson 为彩图绘制提出了宝贵意见。

Trevor Hastie
Robert Tibshirani
Jerome Friedman
斯坦福，加利福尼亚
2001 年 5 月

恬静的统计学家改变了我们的世界；不是通过发现新的事实或者开发新技术，而是通过改变我们的推理、实验和观点的形成方式……

——Ian Hacking

目　录

第1章　绪论 …… 1

第2章　有指导学习概述 …… 6

2.1　引言 …… 6

2.2　变量类型和术语 …… 6

2.3　两种简单预测方法:最小二乘方和最近邻法 …… 7

2.4　统计判决理论 …… 12

2.5　高维空间的局部方法 …… 15

2.6　统计模型、有指导学习和函数逼近 …… 19

2.7　结构化回归模型 …… 22

2.8　受限的估计方法类 …… 23

2.9　模型选择和偏倚 – 方差权衡 …… 25

文献注释 …… 26

习题 …… 27

第3章　回归的线性方法 …… 28

3.1　引言 …… 28

3.2　线性回归模型和最小二乘方 …… 28

3.3　从简单的一元回归到多元回归 …… 34

3.4　子集选择和系数收缩 …… 38

3.5　计算考虑 …… 52

文献注释 …… 52

习题 …… 53

第4章　分类的线性方法 …… 55

4.1　引言 …… 55

4.2　指示矩阵的线性回归 …… 56

4.3　线性判别分析 …… 59

4.4　逻辑斯缔回归 …… 67

4.5　分离超平面 …… 73

文献注释 …… 77

习题 …… 78

第5章　基展开与正则化 …… 80

5.1　引言 …… 80

5.2　分段多项式和样条 …… 81

5.3　过滤和特征提取 …… 88

5.4　光滑样条 …… 88

5.5　光滑参数的自动选择 …… 91

5.6 无参逻辑斯缔回归 …… 95
5.7 多维样条函数 …… 96
5.8 正则化和再生核希尔伯特空间 …… 100
5.9 小波光滑 …… 104
文献注释 …… 109
习题 …… 110
第6章 核方法 …… 115
6.1 一维核光滑方法 …… 115
6.2 选择核的宽度 …… 120
6.3 IR^p 上的局部回归 …… 121
6.4 IR^p 上结构化局部回归模型 …… 123
6.5 局部似然和其他模型 …… 125
6.6 核密度估计和分类 …… 126
6.7 径向基函数和核 …… 129
6.8 密度估计和分类的混合模型 …… 131
6.9 计算考虑 …… 132
文献注释 …… 133
习题 …… 133
第7章 模型评估与选择 …… 135
7.1 引言 …… 135
7.2 偏倚、方差和模型复杂性 …… 135
7.3 偏倚 – 方差分解 …… 137
7.4 训练误差率的乐观性 …… 140
7.5 样本内预测误差的估计 …… 142
7.6 有效的参数个数 …… 143
7.7 贝叶斯方法和 BIC …… 144
7.8 最小描述长度 …… 145
7.9 Vapnik-Chernovenkis 维 …… 147
7.10 交叉验证 …… 149
7.11 自助法 …… 152
文献注释 …… 155
习题 …… 155
第8章 模型推理和平均 …… 158
8.1 引言 …… 158
8.2 自助法和极大似然法 …… 158
8.3 贝叶斯方法 …… 162
8.4 自助法和贝叶斯推理之间的联系 …… 165
8.5 EM 算法 …… 166

8.6 从后验中抽样的 MCMC …… 171
8.7 装袋 …… 173
8.8 模型平均和堆栈 …… 176
8.9 随机搜索:冲击 …… 178
文献注释 …… 179
习题 …… 180
第 9 章 加法模型、树和相关方法 …… 181
9.1 广义加法模型 …… 181
9.2 基于树的方法 …… 187
9.3 PRIM——凸点搜索 …… 195
9.4 MARS:多元自适应回归样条 …… 199
9.5 分层专家混合 …… 204
9.6 遗漏数据 …… 206
9.7 计算考虑 …… 207
文献注释 …… 208
习题 …… 208
第 10 章 提升和加法树 …… 210
10.1 提升方法 …… 210
10.2 提升拟合加法模型 …… 213
10.3 前向分步加法建模 …… 213
10.4 指数损失函数和 AdaBoost …… 214
10.5 为什么使用指数损失 …… 216
10.6 损失函数和健壮性 …… 216
10.7 数据挖掘的“现货”过程 …… 219
10.8 例:垃圾邮件数据 …… 220
10.9 提升树 …… 223
10.10 数值优化 …… 224
10.11 提升适当大小的树 …… 227
10.12 正则化 …… 228
10.13 可解释性 …… 232
10.14 实例 …… 235
文献注释 …… 241
习题 …… 241
第 11 章 神经网络 …… 243
11.1 引言 …… 243
11.2 投影寻踪回归 …… 243
11.3 神经网络 …… 245
11.4 拟合神经网络 …… 247

11.5 训练神经网络的一些问题 …… 249
11.6 例:模拟数据 …… 251
11.7 例:ZIP 编码数据 …… 253
11.8 讨论 …… 257
11.9 计算考虑 …… 257
文献注释 …… 257
习题 …… 258
第 12 章 支持向量机和柔性判别 …… 259
12.1 引言 …… 259
12.2 支持向量分类器 …… 259
12.3 支持向量机 …… 263
12.4 线性判别分析的推广 …… 272
12.5 柔性判别分析 …… 273
12.6 罚判别分析 …… 277
12.7 混合判别分析 …… 279
12.8 计算考虑 …… 284
文献注释 …… 284
习题 …… 285
第 13 章 原型方法和最近邻 …… 287
13.1 引言 …… 287
13.2 原型方法 …… 287
13.3 k -最近邻分类器 …… 290
13.4 自适应的最近邻方法 …… 298
13.5 计算考虑 …… 302
文献注释 …… 302
习题 …… 302
第 14 章 无指导学习 …… 305
14.1 引言 …… 305
14.2 关联规则 …… 306
14.3 聚类分析 …… 316
14.4 自组织映射 …… 335
14.5 主成分、曲线和曲面 …… 339
14.6 独立成分分析和探测性投影寻踪 …… 345
14.7 多维定标 …… 350
文献注释 …… 352
习题 …… 352
术语表 …… 356
参考文献 …… 369

第1章 绪 论

统计学习在科学、财经和工业等许多领域都起着至关重要的作用。下面是一些学习问题的例子：

- 预测一个因心脏病发作而住院的病人是否会再次心脏病发作。这种预测基于人口统计、饮食和对该病人的临床检查。
- 根据公司的业绩和经济学数据，预测今后6个月的股票股价。
- 从数字化的图像，识别手写的邮政编码中的数字。
- 根据患者血液的红外线光谱，估计糖尿病患者血液中的葡萄糖含量。
- 根据临床和人口统计学变量，确定前列腺癌风险因素。

学习科学在统计学、数据挖掘和人工智能等领域起着关键的作用，同时也与工程学和其他学科有交叉。

本书介绍从数据中学习。典型地，有结果度量，通常是量化的(如股票价格)或分类的(如心脏病发作或不发作)，我们希望根据一组特征(feature)(如饮食和临床检查)对其进行预测。假设有训练数据集(training set of data)，借此观察对象集(如人)的结果和特征度量。使用这些数据建立预测模型或学习器(learner)，使我们可以预测新的未知对象的结果。一个好的学习器可以精确地预测这种结果。

上面描述的例子称为有指导学习(supervised learning)问题。之所以称它为“有指导的”，是因为有结果变量指导学习过程。在无指导学习(unsupervised learning)问题中，只能观察特征，而没有结果度量。我们的任务只是描述数据组织或聚类的方式。本书大部分讨论的是有指导学习；关于无指导学习问题的研究不多，仅在最后一章介绍。

下面是本书讨论的实际学习问题的一些例子。

例1 垃圾邮件

本例的数据包括4601封电子邮件信息，研究预测电子邮件是否为垃圾邮件。目标是设计一个垃圾邮件自动检测器，在把邮件放进用户信箱之前过滤掉垃圾邮件。对于所有4601封电子邮件，都知道真实结果(电子邮件类型)email或spam，以及电子邮件中最常出现的57个词和标点符号的相对频率。这是一个有指导学习问题，结果类变量为email/spam。该问题也称分类(classification)问题。

表1.1列出了单词和字符，并显示了它们在email和spam之间的最大平均差异。

表1.1 电子邮件信息中指定的单词或字符的平均百分比。选取显示email和spam之间差别最大的单词和字符

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

我们的学习方法必须决定使用哪些特征以及如何使用。例如，可以使用下面的规则：

```
if ( %george < 0.6) & ( %you > 1.5)   then spam
                                      else email
```

规则的另一种形式可以是：

```
if (0.2· %you - 0.3· %george) > 0   then spam
                                    else email
```

对于该问题，并非所有错误都是等同的。我们想避免过滤掉好的电子邮件，尽管使垃圾邮件通过不是所希望的，但也不会导致太大问题。本书将讨论处理该学习问题的多种不同方法。

例 2 前列腺癌

该例的数据如图 1.1 所示，取自 Stamey 等人(1989)的研究。该研究考察了 97 位准备做前列腺根治术病人的前列腺特殊抗原(prostate specific antigen，PSA)水平与一些临床指标之间的相关性。

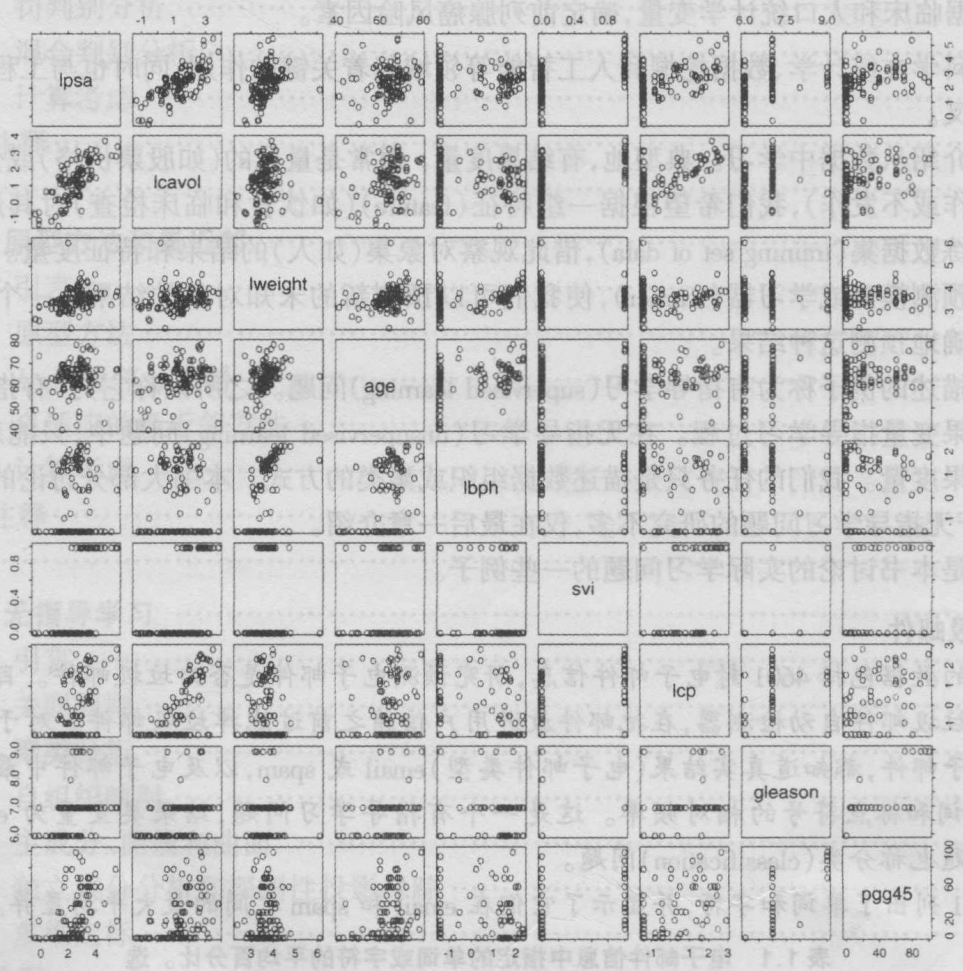


图 1.1 前列腺癌数据的散点图矩阵。第一行依次显示对每个预测子的响应。其中 svi 和 gleason 两个预测子是分类的

本例的目的是从一些指标预测 PSA 的记录值 lpsa。指标包括肿瘤体积记录值 lcavol、前列腺重量记录值 lweight、年龄 age、良性前列腺增生量 lbph、精囊浸润 svi、包膜穿透记录值 lcp、Gl-

eason 积分 gleason 和 Gleason 积分 4 或 5 所占的百分比 pgg45。图 1.1 是这些变量的散点图矩阵。一些指标与 lpsa 的相关性是显而易见的，但是靠肉眼构造一个好的预测模型很困难。这是一个有指导学习问题，称为回归问题(regression problem)，因为输出度量是定量的。

例 3 手写体数字识别

该例的数据取自美国邮政信封上手写的邮政编码。每幅图像从一个五位的邮政编码截取，隔离成单个数字。图像是 16×16 的八位灰度图，每个点的亮度从 0 到 255。一些样本图像在图 1.2 中给出。

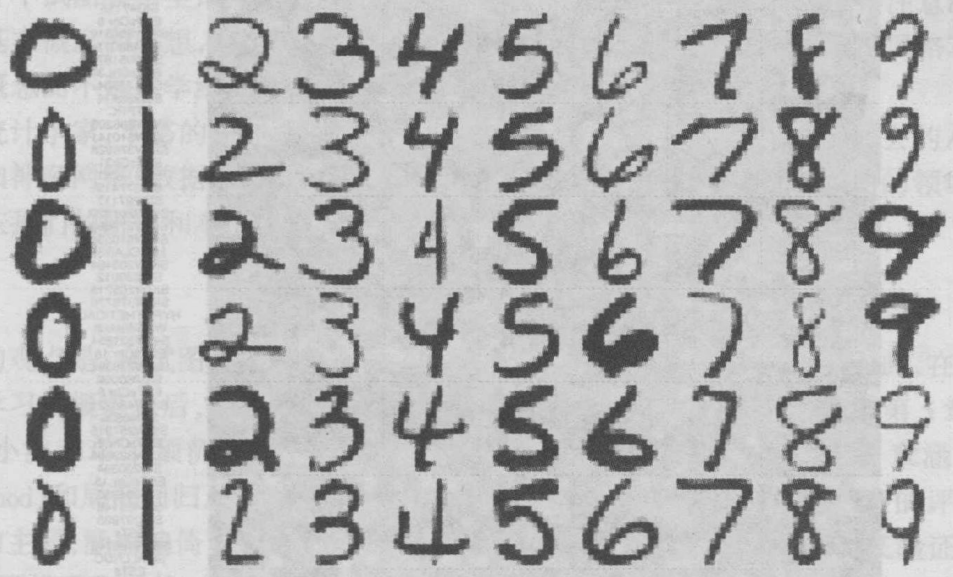


图 1.2 取自美国邮政信封的手写体数字

图像已被规范化，使它们具有大致相同的尺寸和方位。本例的任务是从 16×16 的点亮度矩阵，快速准确地识别图像{0,1,…,9}。如果足够准确，结果算法就可以用于信件自动分拣过程。这是一个分类问题，要求错误率很低，以避免邮件误投。为了获得低误差率，有些对象可以归入“不知道”类，并通过手工分拣。

例 4 DNA 表达微阵列分析

DNA 代表脱氧核糖核酸，是构成人类染色体的基本物质。通过测定出现在细胞某基因中的 mRNA(信使核糖核酸)总量，可用 DNA 微阵列来测量细胞中的基因表达。微阵列是生物学的一项突破性技术，便于同时对细胞单个样本中数以千计的基因进行定量研究。

下面介绍 DNA 微阵列如何工作。数千基因的核苷酸序列印在一个玻璃片上。一个目标样本和一个参照样本用红绿染色标记，并均与玻璃片上的 DNA 杂交。通过荧光检查器，测量每个位点上 RNA 的记录(红/绿)强度。结果是数千个值，通常在 -6~6 之间，测量目标样本中每个基因相对于参照样本的表达水平。正值表示目标样本的表达水平高于参照样本的表达水平，负值相反。

基因表达数据集将一系列 DNA 微阵列实验的表达值收集在一起，每一列代表一个实验。因此，有数千行代表个体基因，数十列代表样本。在图 1.3 的特例中，有 6830 个基因(行)和 64 个样本(列)，为了简洁，只显示了 100 行随机选样。该图以热度图(heat map)的形式显示数据，颜色变化从绿(负)到红(正)。样本取自不同病人的 64 个癌瘤。

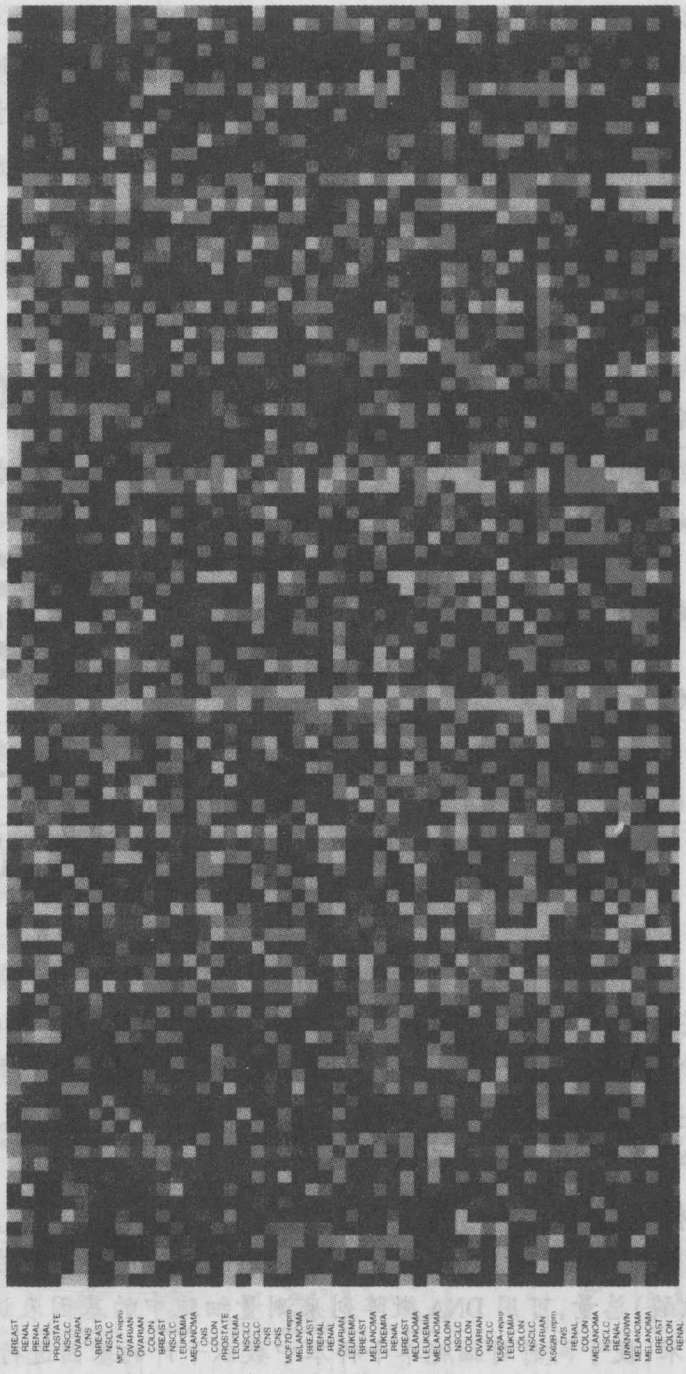


图 1.3 DNA 微阵列数据：人体瘤数据 6830 个基因(行)和 64 个样本(列)的表达水平矩阵。只显示100行的随机选样。显示的是热度图，从鲜绿(负，低显性)到鲜红(正，高显性)。遗漏的值为灰色。行和列以随机次序显示(见彩页)

我们面对的挑战是理解基因和样本是如何组织的。典型的问题包括：

(a) 根据基因的表达图解，哪些样本最相似？

(b) 根据样本的表达图解，哪些基因最相似？

(c) 对于某些癌样本，某些基因显示很高(或很低)的表达水平吗？

可以将该任务看做回归问题，它具有两个分类预测变量——基因和样本；响应变量是表达