第一图书网, tushu007.com

<<数据挖掘方法与模型>>

图书基本信息

书名：<<数据挖掘方法与模型>>

13位ISBN编号：9787040309683

10位ISBN编号：7040309688

出版时间：2011-3

出版时间：高等教育出版社

作者：拉罗斯

页数：287

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

内容概要

当下，由于强大的数据挖掘软件平台很容易获得，草率地使用数据挖掘方法和技术将导致挖掘的结果混淆难解。
这种失误往往源自盲目使用“黑盒子”方法进行数据挖掘，而最好的避免途径就是使用“自盒子”方法，理解隐藏在软件背后的算法和统计模型结构。

本书分为7章，第1章是对降维方法的介绍，这是数据挖掘技术的一个先决条件；第2章至第6章为经典的数据挖掘算法和技术，包括一元回归模型、多元回归模型、逻辑回归模型、贝叶斯网络分析以及遗传算法，通过实际案例引导读者由已预处理的数据使用不同的挖掘技术从而得出所需结论；第7章为基于数据挖掘过程模型上的多个案例研究，通过多个领域的案例来阐述算法和技术是如何被运用的。

本书可作为数据挖掘课程教学用书，适用于高年级本科生和研究生的教学，也可供科研人员参考使用。

作者简介

作者：（美国）拉罗斯（Daniel T.Larose） 译者：刘燕权 胡赛全 冯新平 等Daniel T.Larose，博士，美国中康涅狄格州立大学统计学教授。
设计、开发并主持了世界上第一个在线数据挖掘管理科学硕士学位课程及教学，创立了中康涅狄格州立大学数据挖掘研究室Data Mining@CCSU。
研究兴趣包括数据挖掘、统计分析等。
发表多篇论文，出版学术专著5部。
刘燕权，博士，美国南康涅狄格州立大学终身正教授、校理事会理事，美福布赖特学者（2009-2010），北京大学、南京大学、清华大学、中国科学院研究生院、南京理工大学、内蒙古大学等客座教授。
研究方向为计算机科学与理论、数据挖掘、软件工程项目开发与管理、数字图书馆、信息技术理论与实践、多媒体设计及应用等。
发表论文及学术专著80余篇（部）。

第一图书网, tushu007.com

<<数据挖掘方法与模型>>

书籍目录

第1章　降维方法
1.1　数据挖掘中降低维度的必要性
1.2　主成分分析法
1.2.1　主成分分析应用于房屋数据集
1.2.2　应提取多少个主成分
1.3　因子分析法
1.3.1　因子分析法在成年人数据集中的应用
1.3.2?因子旋转
1.4　用户自定义合成
总结
参考文献
练习题
第2章　回归模型
2.1　简单线性回归实例
2.2　最小二乘法估计
2.3　决定系数
2.4　估计值的标准误差
2.5　相关系数
2.6　方差分析表
2.7　异常点、高杠杆点和强影响观测值
2.8　回归模型
2.9　回归推断
2.9.1　x和y之间线性关系的t检验
2.9.2　回归直线斜率的置信区间
2.9.3　给定x条件下，Y均值的置信区间
2.9.4　给定x条件下，Y随机选择值的预测区间
2.10　回归假设检验
2.11　实例：棒球数据集
2.12　实例：加利福尼亚州数据集
2.13　线性变换实现
总结
参考文献
练习题
第3章　多元回归和建模
3.1　多元回归实例
3.2　多元回归模型
3.3　多元回归推断
3.4　含有分类预测变量的回归
3.4.1　调整R2：对包含无用预测变量的惩罚模式
3.4.2　序贯的误差平方和
3.5　多重共线性
3.6　变量选择方法
3.6.1　偏F检验
3.6.2　向前选择程序
3.6.3　向后排除程序

3.6.4　逐步选择程序
3.6.5　最优子集程序
3.6.6　所有可能的子集选择程序
3.7变量选择方法的应用
3.7.1　向前选择程序应用于谷物数据集
3.7.2　向后排除程序应用于谷物数据集
3.7.3　逐步选择程序应用于谷物数据集
3.7.4　最优子集程序应用于谷物数据集
……
第4章　逻辑回归
第5章　朴素贝叶斯估计和贝叶斯网络
第6章　遗传算法
第7章　案便研究：直邮营销的回应建模问题
总结
参考文献

章节摘录

版权页：插图：通常用于数据挖掘的数据库可能有上百万条记录和数千个变量。
所有变量都是独立而没有任何关联的现象是不常见的。
如《数据中发掘知识：数据挖掘引言》中所提及的那样，数据分析人员需要防范多重共线性，即预测变量之间相互关联的情形。
多重共线性会导致解空间的不稳定，从而可能导致结果的不连贯。
如在多元回归中，即使单个变量的回归结果均不显著，预测变量的多重共线性集可能导致回归整体相对显著。
即使上述的不稳定性得以避免，包含具有高度相关性变量的模型往往会强调其某一特定成分，该成分实质上被重复计算。
贝尔曼指出，样本量需要符合一个多元函数，该函数跟随变量数呈现指数关系递增。
换句话说，高维空间本身具有稀疏性。
正如这个经验法则告诉我们的，在一维空间的正态分布中，有68%的值介于正负标准差之间，而在10维多元正态分布中，只有0.02%的数据属于类似的高维空间。
在考察预测变量和回应变量之间的关系时，过多地使用预测变量会不必要地复杂化分析过程。
这违反了简约原则，即应将预测变量的数目保持在可控的范围内。
另一方面，过多的变量会妨碍查找规律的建立，因为新的数据对所有变量作出的反应很可能和建模中采用的数据反应不同。
此外，仅在变量层面上分析可能会忽略变量之间的潜在联系。
例如，几个预测变量可能落人仅反映数据某一方面特征的一个组（一个因素或一个组成部分（components））内。

编辑推荐

《数据挖掘方法与模型》：国外信息技术优秀图书选择。

第一图书网, tushu007.com

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com