数据挖掘：概述

什么是数据挖掘？

- 数据挖掘是近年来[1]从统计学和计算机科学（机器学习和数据库技术）交叉而来的新词汇，应用于科学、工程和商业领域中的大型数据库
- 数据挖掘处正在变动和发展过程中，有很多数据挖掘的定义，也有很多关于数据挖掘是什么和不是什么的讨论。本课用的术语并不是标准的，例如：偏差、分类、预测、特征＝自变量、目标＝因变量、事例＝范本＝行

---

[1]第一次关于数据挖掘和知识发现的国际会议于1995年召开

广义和狭义的数据挖掘定义

- 广义的数据挖掘定义包括传统的统计学方法；狭义的定义则强调自动和启发式方法
- 数据挖掘、数据捕捞、无特定目标的搜索
- 数据库中知识发现（KDD）

我喜欢的（定义）

- “大规模和快速的统计学”

——Darryl Pregibon

- 本人对上述定义的扩展：“大规模、快速的、简明的（统计学）”

Gartner小组（的定义）

- “数据挖掘是用模式识别、统计学、数学等方法过滤存储在数据库中大量的数据来发现新的、有意义的关系、模式和趋势的过程。”

（数据挖掘产生的）驱动力

- 市场因素：从关注产品/服务到关注客户
- 信息技术：从关注最新的收支差额到关注交易模式—数据仓库（DW）—联机分析处理（OLAP）
- 存储费用大幅度下降：（因此产生了）巨大的数据库。例如，沃尔玛2千万交易/天，10万亿字节的数据库；BlockBuster（全球最大的音像制品连锁租赁公司）:（有）3千6百万家庭（的数据）；
- 交易数据可自动获取。例如：条形码、POS机、鼠标点击、位置数据（GPS、移动电话）
- 因特网：个性化的交互、纵向的数据

核心学科

- 统计学（随着21世纪数据规模和处理速度的要求而改变）。例如：
  - 描述上：可视化
  - 模型：回归、聚类分析
- 机器学习。例如：神经网络
- 数据库检索。例如：关联规则
- 平行的发展：决策树、k-最近邻、OLAP-EDA（联机分析－电子数据交换）

数据挖掘过程

1、理解应用和目标；

2、得到研究用的数据集（通常来自数据仓库）；

3、数据清洗和预处理；

4、数据降维和投影;

5、选择数据挖掘任务；

6、选择数据挖掘算法；

7、用算法完成任务；

8、解释结果，如果需要重复步骤1－7；

数据挖掘

9、配置：集成进运作的系统。

SEMMA方法论（SAS）

- S：从数据集中抽取样本，分成训练集、验证集和测试集
- E：通过统计及图示等方法探究数据集（隐含的规律）
- M：修正：变量转换、填补数据缺省值
- M：模型：建立合适的模型，如回归、分类树、神经网络
- A：评估：用验证、测试数据集来检验模型

应用示例

- 客户关系管理
- 财务分析
- 电子商务和互联网

客户关系管理

- 目标市场
- 流失预测/流失分析
- 欺诈检测
- 信用评分

目标市场

- 商业问题：使用潜在客户列表进行直邮活动
- 解决方案：人口、地理数据结合过去购买行为数据，用数据挖掘识别确定最有希望的回应者
- 收益：更高的回应率、节约活动费用

例子：Fleet金融集团

- 重新设计客户服务结构，包括在数据仓库和营销自动化方面投资了3千8百万美元
- 从1千5百万客户中抽取的2万个样本，并用Logistic回归去预测对房屋资产贷款（home-equity）产品回应的概率
- 用CART方法去预测有利可图的客户，和及时响应也无利可图的客户；

流失分析：Telcos公司

- 商业问题：防止客户流失，避免增加倾向于流失的客户
- 解决方法：用神经网络、时间序列分析方法确定典型的易于流失和背叛的顾客的电话使用模式
- 收益：保持并更有效的促进客户

例子：法国电信

- 建设流失/客户档案系统作为主要客户的数据仓库解决方案的一部分
- 基于客户特征的预防性的CPS（客户流失预防系统），是从已知的易于流失和非易于流失的客户的例子来确定易于流失的客户的重要特征
- 早期的CPS系统用与已知的、易流失的客户的例子相匹配的模式

欺诈检测

- 商业问题：欺诈增加成本，或减少收益
- 解决方法：用神经网络、Logistic回归去确定欺诈性例子的特征，以便将来防止（类似事情发生）或更有力的检举
- 收益：通过减少不理想的客户来增加收益；

例子：马萨诸塞州汽车保险局

- 通过专家细察过去的、关于保险理赔的报告去确定欺诈的例子；
- 关于原告、事故类型、伤害类型/处理措施的一些特征（超过60）都编入数据库；
- 用降维方法去获得带有权重的变量。多元回归分步子集选择方法去识别和欺骗强关联的特征；

风险分析

- 商业问题：降低由于客户的过失而造成的贷款风险
- 解决方法：用判别分析方法于信用评分模型去构造可以区分有风险的客户的评分函数
- 收益：减少呆帐费用

财务

- 商业问题：公司债券的定价依赖于几个因素：公司的风险情况、债务时间的长短、红利、以前的历史等
- 解决方法：通过数据挖掘方法找出更准确的价格预测模型

电子商务与因特网

- 协同过滤
- 从点击率到客户

推荐系统

- 商业机会：在网络上的用户评级（Amazon.com,CDNOW.com,MovieFinder.com）。怎样用其它的客户的信息来评价一个特殊的用户
- 解决方案：用一种协同过滤的技术
- 收益：增加横向销售、进阶销售等的收入

用户点击

- 商业问题：50％的Dell计算机订单是在网上下达的，然而，保持率是0.5%，也就是0.5%Web页面的浏览者成为了客户
- 解决方法：通过一系列的点击，聚类用户并设计网站，使最终购买的客户数量最大化
- 收益：增加收入

正在形成的主要的数据挖掘应用

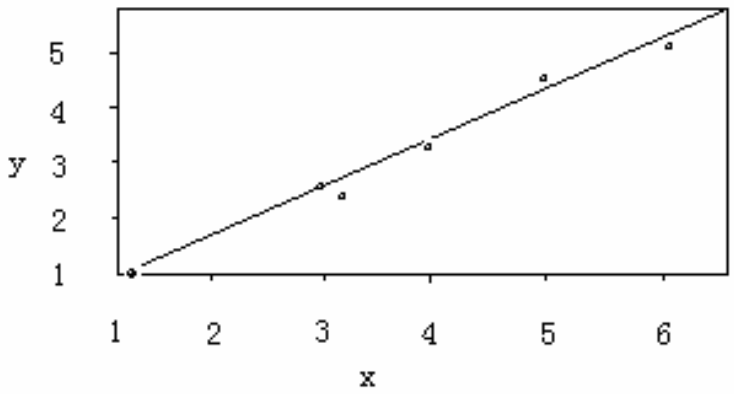
- 垃圾邮件
- 生物信息学/基因组学
- 医疗的历史数据——保险索赔
- 电子商务的个性化服务
- 射频标签：Gillette
- 安全：
  - 集装箱运输
  - 网络入侵检测

核心概念

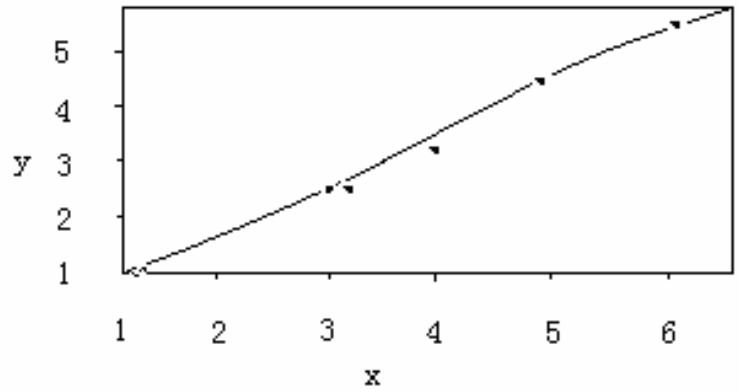
- 数据类型：
  - 数值型：连续的，包括比率和区间型
    离散的
    需要分箱的
  - 类别的
    有序的
    名义的
  - 二值的
- 过拟合与泛化
- 正则化：对模型复杂性的惩罚
- 距离度量
- 维数灾难
- 随机和分层抽样、再抽样
- 损失函数

过拟合的回归例子

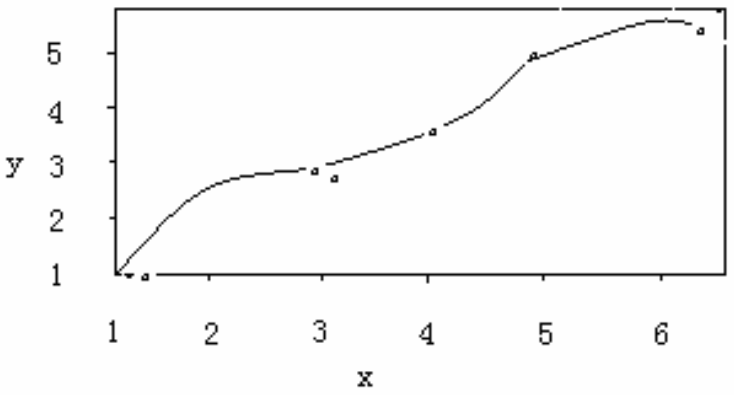
线性拟和



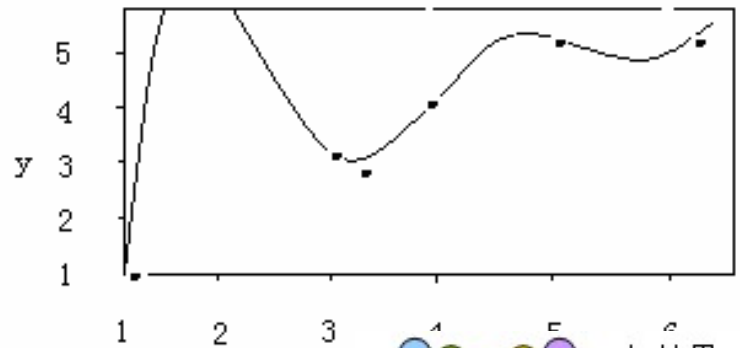
三次拟和



四次拟和



五次拟和

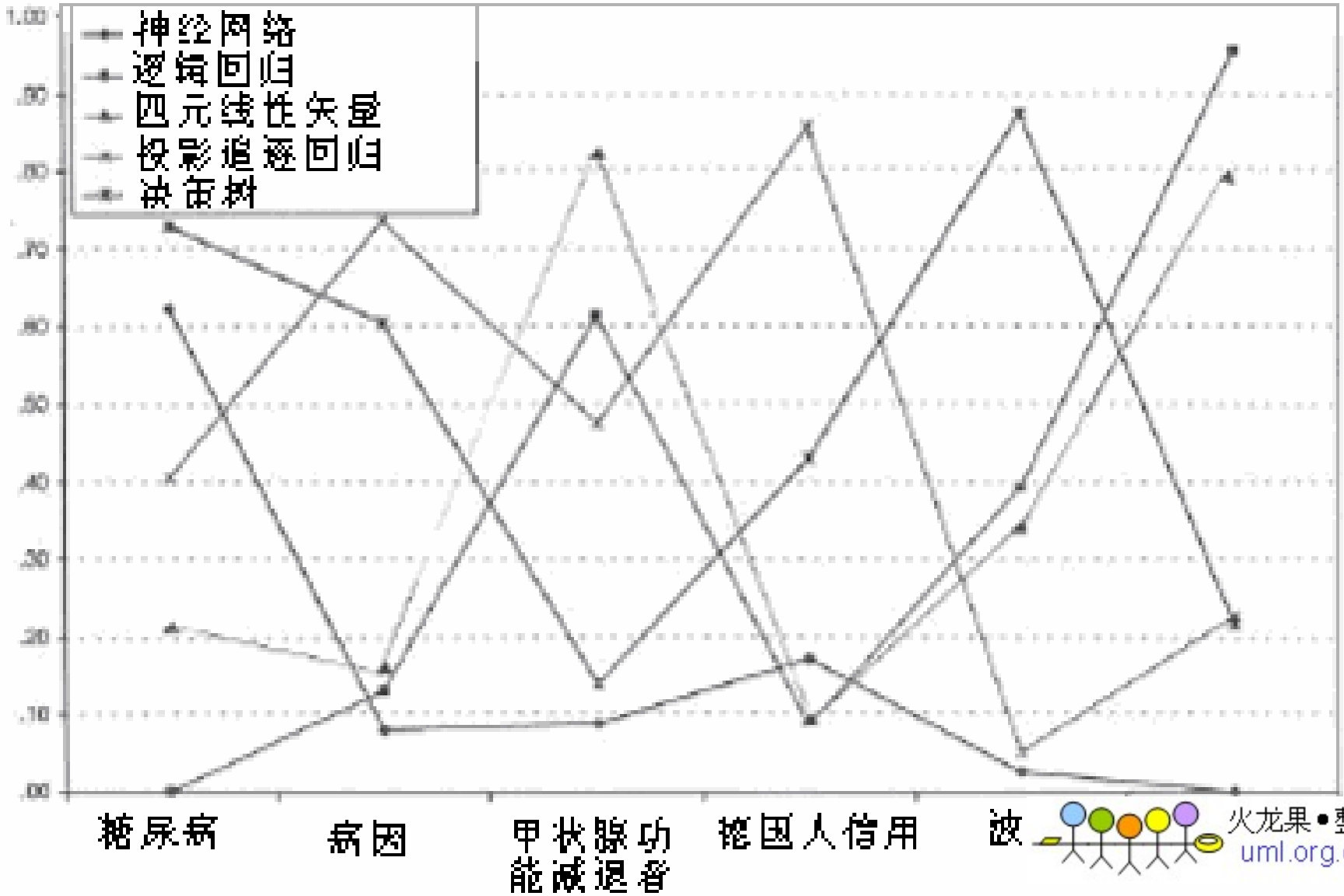


数据挖掘的典型特征

- 标准的格式是电子数据表：
  - 行：观察单元；列：变量
- 许多行和列
- 许多行有适度的列，如电话记录
- 许多列有适度的行，如基因组学
- 机会主义（通常是交易处理的副产品）
  - 不是来自设计的实验
  - 经常有异常点和缺失数据；

相关性表现例：6个数据表上的5次运算

(Lee & Elder, 1997)



课程中讨论的题目

- 有指导的技术：
  - 分类：k-最近邻、朴素贝叶斯、分类树；判别分析、Logistic回归、神经网络
  - 预测（估计）：回归、回归树、k-最近邻
- 无指导技术：
  - 聚类分析、主成分分析
  - 关联规则、协同过滤