新媒体大数据分析

New Media Big Data Analysis

第三章 数据建模

黄振亚，朱孟潇，张凯

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/NM2025.html

助教：齐畅，朱家骏
bigdata_2025@163.com

数据建模基础

- 基本概念——数据挖掘是什么？
  - **数据挖掘**：从大量的数据中挖掘哪些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识，并据此更好的服务人们的生活。

图1-2 我们的数据丰富，但信息贫乏

数据挖掘：在你的数据中搜索知识（有趣的模式）

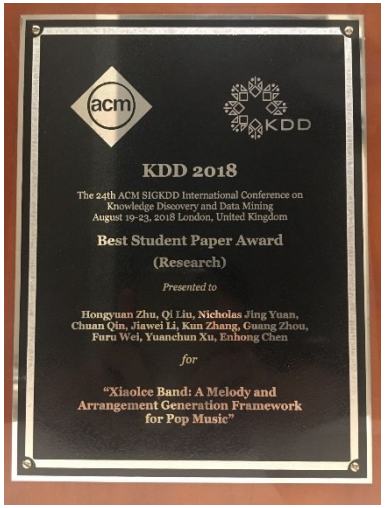
数据建模基础

- 基本概念——数据挖掘是什么？
  - 数据挖掘的近义词
    - 从数据中挖掘知识
      - Knowledge Discovery in Data
    - 知识提炼
    - 数据/模式分析
    - 数据考古
    - 数据捕捞、信息收获、资料勘探等。

- SIGKDD: Knowledge Discovery and Data Mining

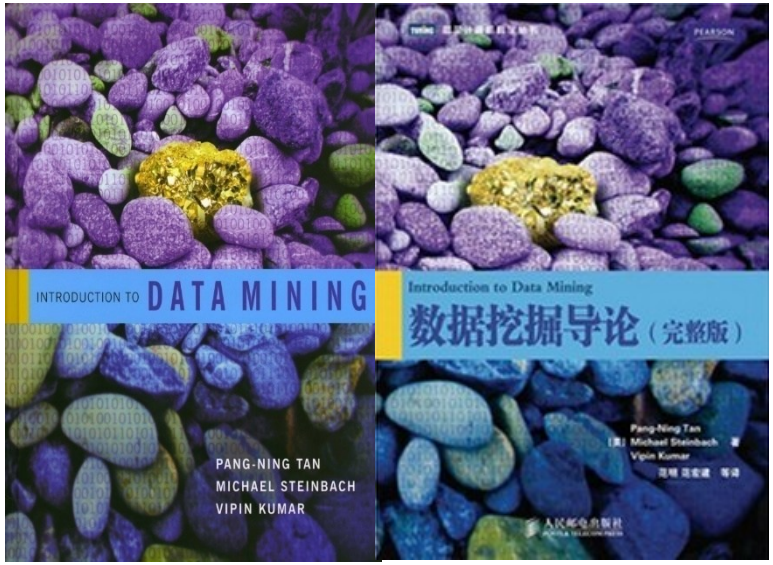
25TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING

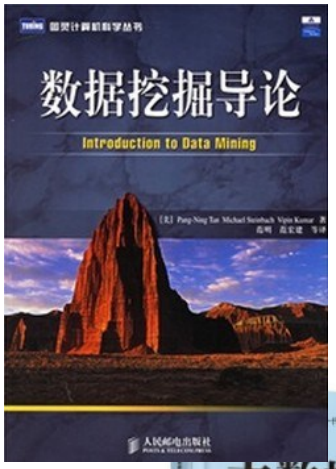
Association for Computing Machinery

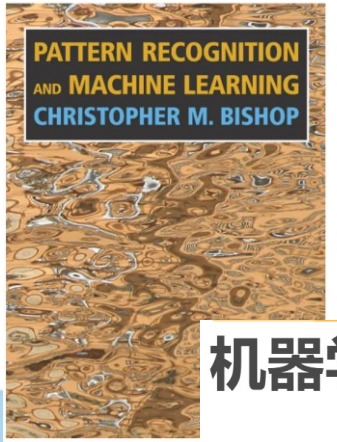


数据建模基础

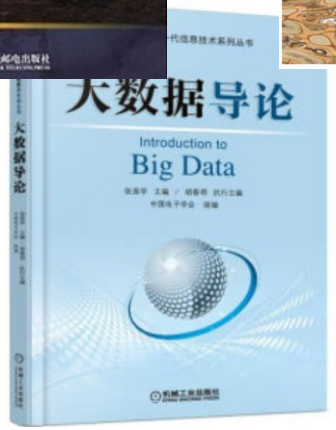
- 参考书
  - 数据挖掘导论 （Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley ）







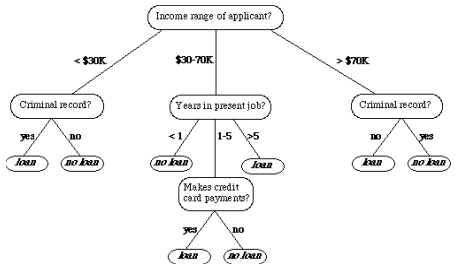
机器学习



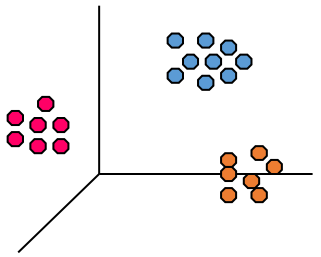
数据建模基础

- 数据挖掘有哪些典型任务?

分类与预测



聚类



数据

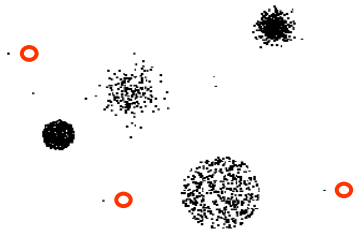
	T		H		P	
	L	H	L	H	L	H
J	-6.0	8.8	60	100	986	1044
F	-2.8	10.9	48	100	973	1025
M	-5.6	17.7	34	100	976	1037
A	-1.2	22.2	27	100	996	1036
M	-0.8	27.8	25	100	1003	1034
J	5.2	29.1	26	100	998	1030
J	9.8	30.6	23	99	997	1027
A	5.6	26.1	31	100	992	1029
S	5.2	24.8	35	100	998	1028
O	-0.4	21.3	42	100	990	1031
N	-7.6	17.3	55	100	963	1023
D	-10.4	9.2	53	100	987	1039

table 17a

2010 monthly weather variation, Cambridge (UK)

关联分析

异常检测



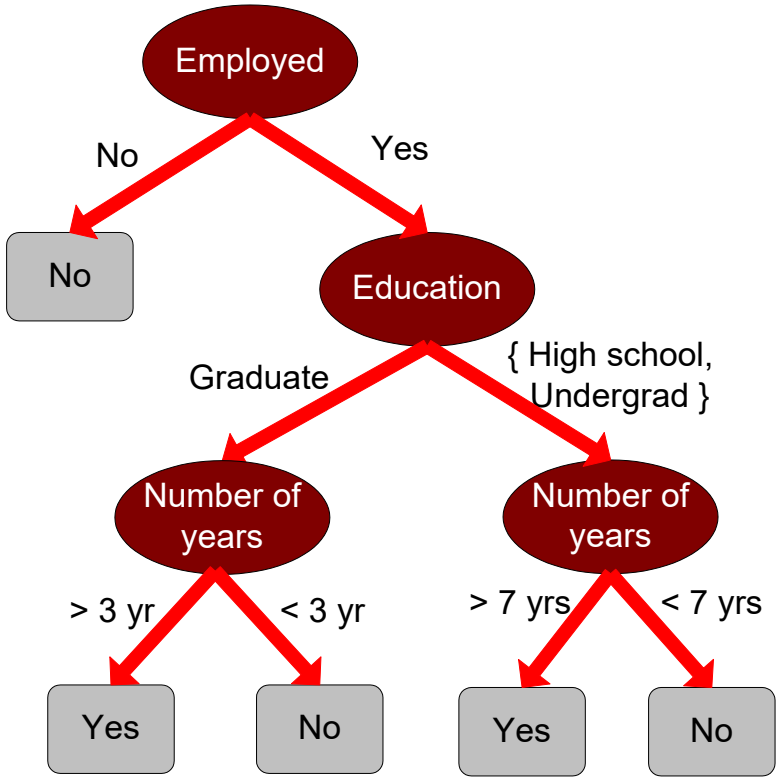
数据建模基础

- 数据挖掘任务——分类与预测 (Classification, Prediction)
  - 预测性建模（监督学习）
  - 寻找一个模型：特征–>类别的函数

类别

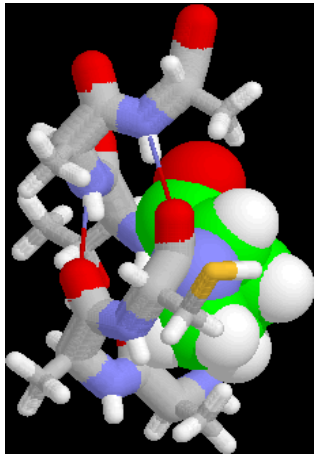
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness（信誉）



数据建模基础

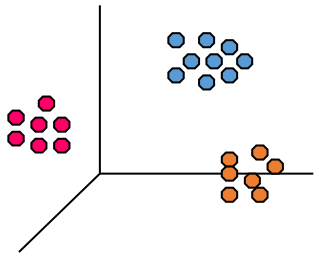
- 数据挖掘任务——分类与预测 (Classification, Prediction)
  - 邮件分类（垃圾邮件）
  - 将新闻故事分类为财经、天气、娱乐、体育等
  - 判断信用卡交易是合法的还是欺诈
  - 将蛋白质的二级结构分类为α -螺旋、β -薄片或随机螺旋
  - 预测肿瘤细胞是良性还是恶性
  - 识别网络空间的入侵者
  - 推荐系统
  - 。。。

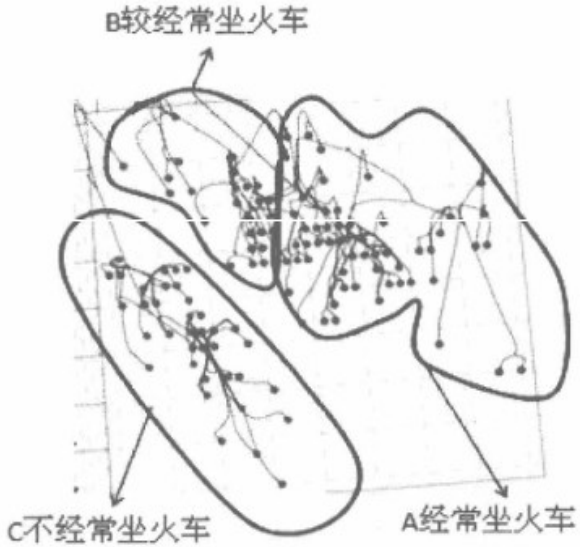


数据建模基础

- 数据挖掘任务——聚类(Clustering, unsupervised learning)
  - 例如：铁路票价制定
  - 问：如何制定合适的票价提高上座率?
  - 方案：将旅客进行聚类分析，根据旅客乘坐高铁的频率 提供不同的优惠政策。合适的定价是提高高铁上座率的保障

聚类





数据建模基础

- 数据挖掘任务——聚类(Clustering)
  - 例：搜索词条聚类(Query clustering)
  - “USTC”，“中科大”，“中国科大”，“中国科学技术大学”
  - “长城”，“颐和园”，“故宫”

数据建模基础

- 数据挖掘任务——关联分析(Association Analysis)
  - 例如： “啤酒与尿布”
  - 在一次圣诞节的顾客消费行为分析中，沃尔玛意外发现跟尿布一起购买最多的商品竟然是啤酒。经过深入分析后，卖场立即对两类商品的空间距离与价格都进行了调整，结果尿布与啤酒销量双双大增。

萨姆·沃尔顿
沃尔玛公司创始人

轰动一时的啤酒与尿布关联规则

数据建模基础

- 数据挖掘任务——异常检测 (Anomaly Detection)
  - 检测非正常行为
  - 应用:
    - 信用卡诈骗检测
    - 网络入侵检测