数据挖掘概述

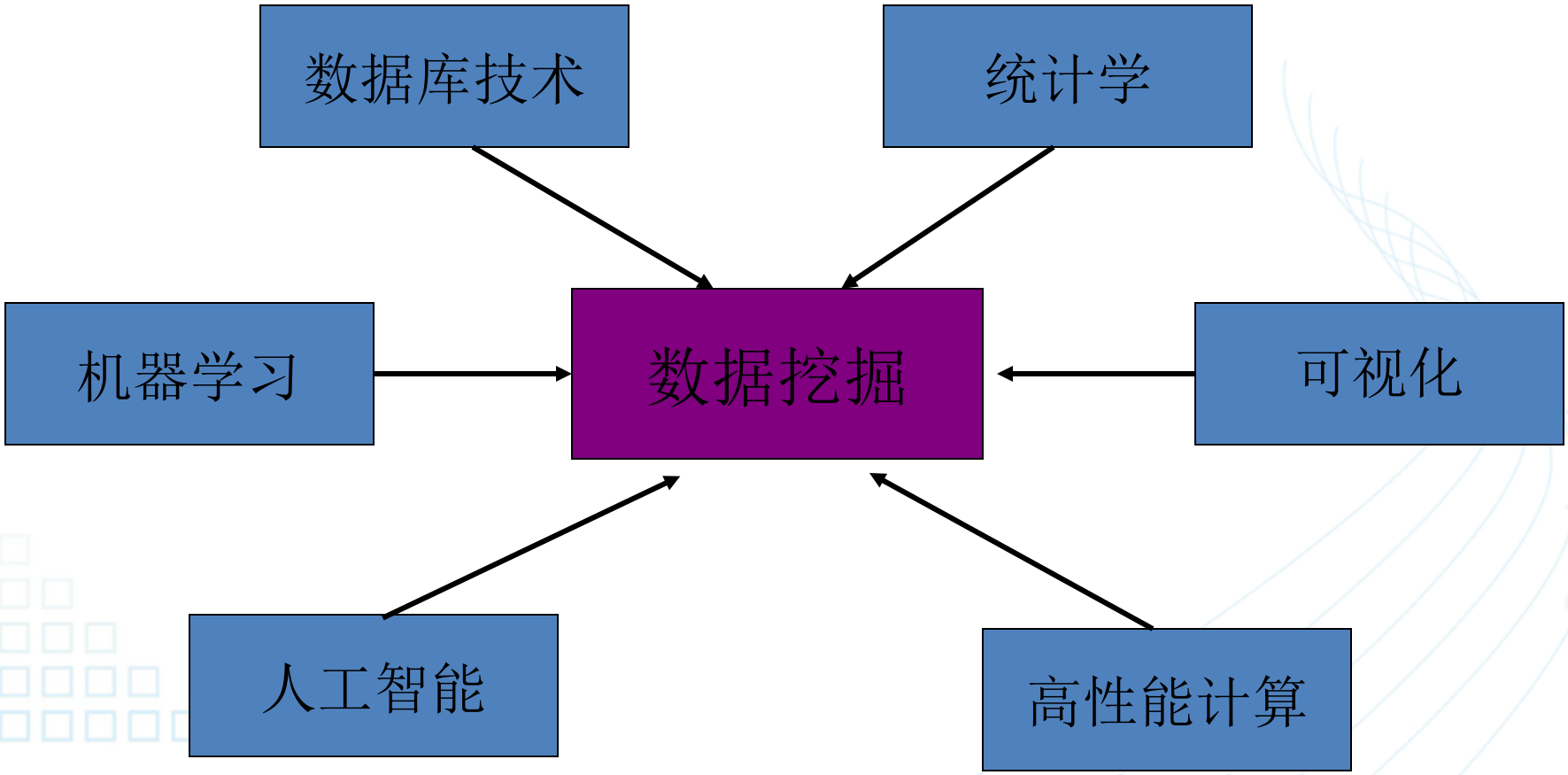

马 帅

北京航空航天大學
BEIHANG UNIVERSITY

数据挖掘的定义

- **数据挖掘是从大量数据中提取或“挖掘”知识。**
- 技术上的定义：数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- 商业角度定义：数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。
- 所谓**基于数据库的知识发现**（KDD）是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。

数据挖掘是多学科的产物



统计分析和数据挖掘的区别-孙悟空跟二郎神打仗

- **统计分析**：两人斗争4567次，其中孙悟空赢3456次。另外，孙悟空斗牛魔王，胜率是89%，二郎神斗牛魔王胜率是71%
- **关联分析**：计算死自动找到出身、教育、经验、单身四个因素。得出结论是孙悟空赢。

贫苦出身的孩子一般比皇亲国戚功夫练得刻苦；
打架经验丰富 的人因为擅长利用环境而机会更多；
在都遇得到明师的情况下，贫苦出身的孩子功夫可能会高些；
单身的人功夫总比同样环境非单身的高。

数据挖掘与统计学

统计学和数据挖掘有着共同的目标：**发现数据中的结构**。

事实上，由于它们的目标相似，一些人（尤其是统计学家）认为数据挖掘是统计学的分支。这是一个不切合实际的看法。因为数据挖掘还应用了其它领域的思想、工具和方法，尤其是计算机学科，例如数据库技术和机器学习，而且它所关注的某些领域和统计学家所关注的有很大不同。

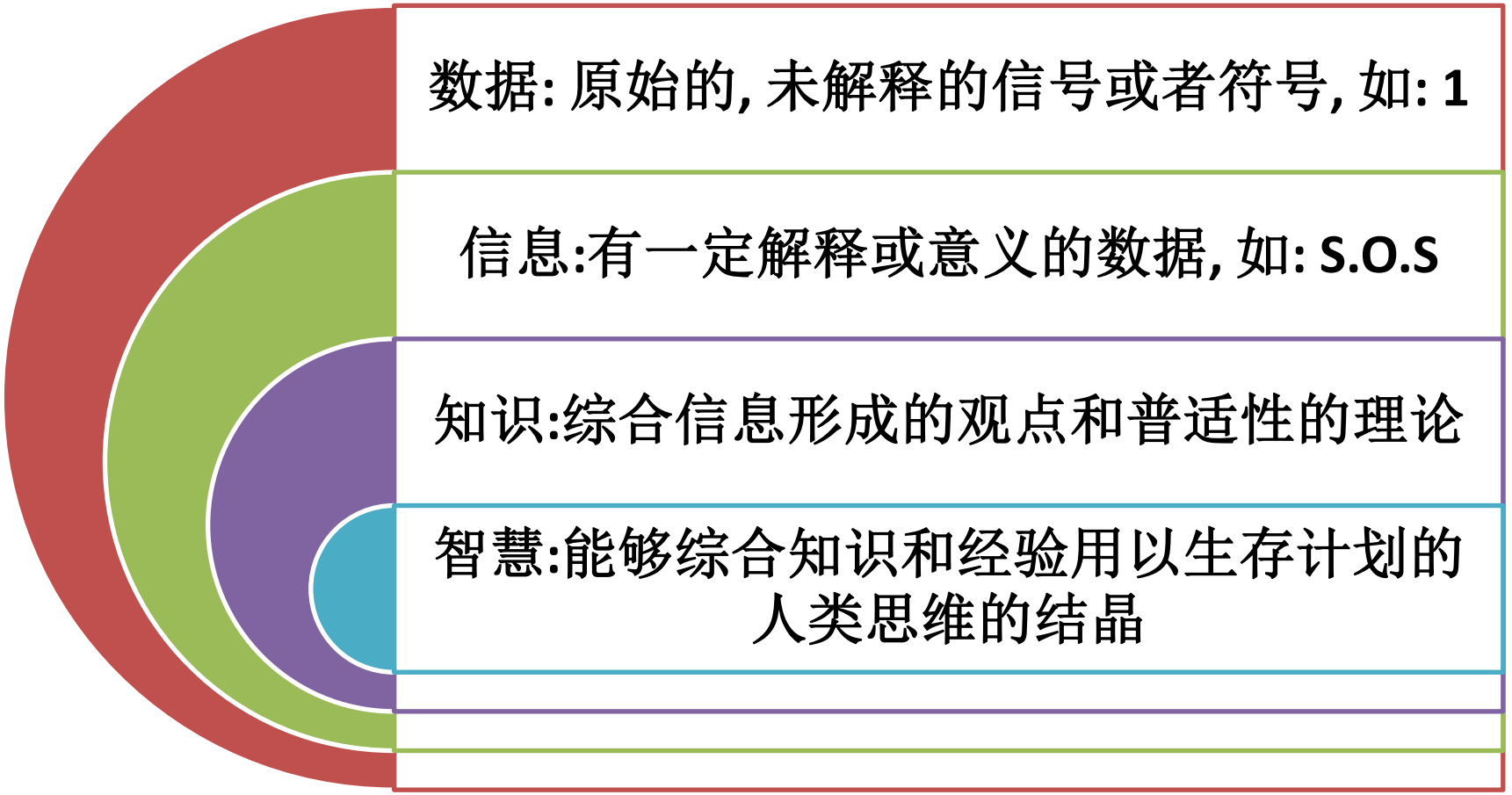
数据挖掘被关注的原因

数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要**将这些数据转换成有用的信息和知识。**获取的信息和知识可以广泛用于各种应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。

数据挖掘是信息技术自然进化的结果

- 数据库、数据仓库和Intenet等信息技术的发展。
- 计算机性能的提高和先进的体系结构的发展。
- 统计学和人工智能等方法在数据分析中的研究和应用。

四个概念的不同


数据: 原始的, 未解释的信号或者符号, 如: 1
信息:有一定解释或意义的数据, 如: S.O.S
知识:综合信息形成的观点和普适性的理论
智慧:能够综合知识和经验用以生存计划的人类思维的结晶

数据挖掘视为数据库中知识发现过程基本步骤的主要环节

数据挖掘的应用

- 电信 ：流失
- 银行：聚类（细分）， 交叉销售
- 百货公司/超市：购物篮分析 （关联规则）
- 保险：细分，交叉销售，流失（原因分析）
- 信用卡： 欺诈探测，细分
- 电子商务： 网站日志分析
- 税务部门：偷漏税行为探测
- 警察机关：犯罪行为分析
- 医学： 医疗保健

数据挖掘应用实例：市场分析和管理

- 顾客形象
  - 数据挖掘可以告诉你什麽样的顾客会买什麽样的产品（聚类或分类）
- 识别顾客需求
  - 保证为不同的顾客提供了最好的产品
  - 使用预测手段去发现什麽因素会吸引新的顾客。
- 提供汇总信息
  - 各种各样的多方位汇总信息
  - 统计的汇总信息（数据中心的趋势和变化）

数据挖掘应用实例：欺骗性检测和管理

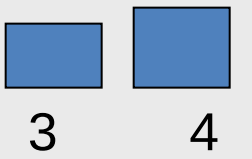
- 应用
  - 广泛应用于医疗系统, 零售系统,信用卡服务, 电信(电话卡欺骗行为), 等等.
- 实现途径
  - 利用历史性数据建立欺骗性行为模型并使用数据挖掘帮助识别同类例子
- 具体事例
  - 汽车保险：检测出那些故意制造车祸而索取保险金的人
  - 来路不明钱财的追踪: 发现可疑钱财交易(美国财政部的财政犯罪执行网)
  - 医疗保险: 检测出潜在的病人，呼叫医生和证明人

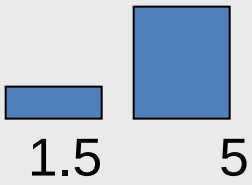
数据挖掘应用实例：欺骗性检测和管理

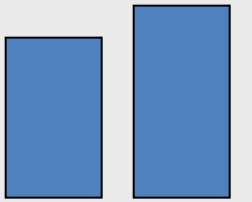
- 发现不正确的医学治疗
  - 澳大利亚医疗保险协会证明在许多情况下全面审查测试是很需要的
- 检测电话错误
  - 电话呼叫模式：呼叫目的地，持续时间，每天或每周的次数。分析与预期标准相背离的模式
- 零售
  - 分析家估计38%的零售收缩缘于雇员的不诚实。

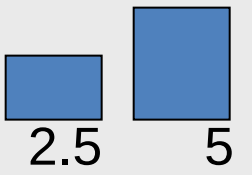
数据挖掘实例：分类问题

Examples of class A

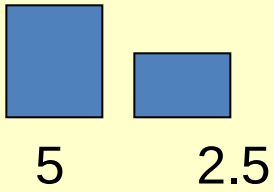


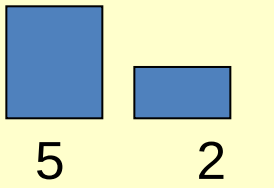


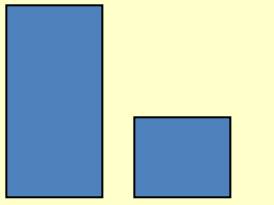


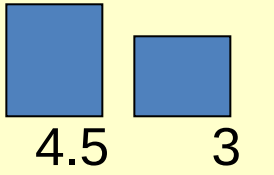


Examples of class B



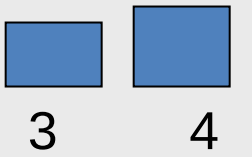


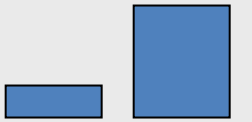


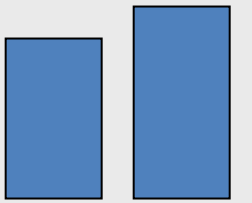


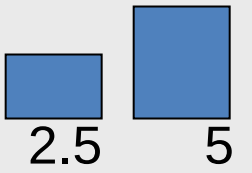
数据挖掘实例：分类问题

Examples of class A

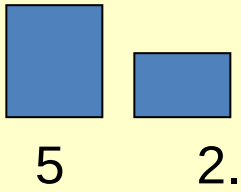


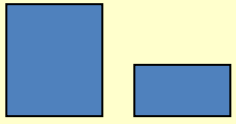


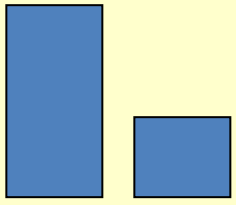


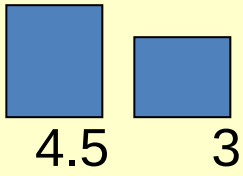


Examples of class B

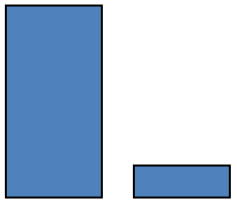


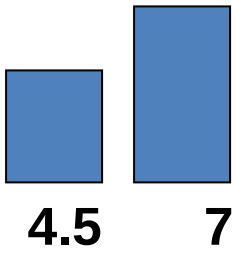






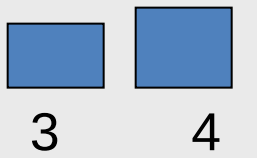
问题：如何分类？

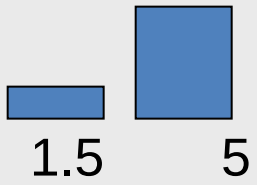


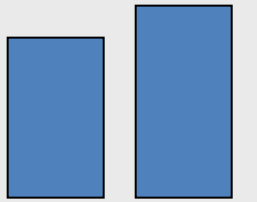


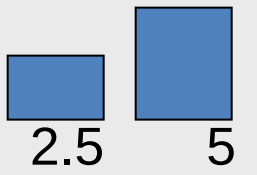
数据挖掘实例：分类问题

Examples of class A

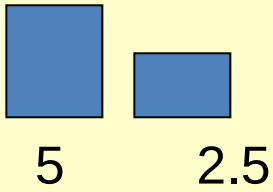


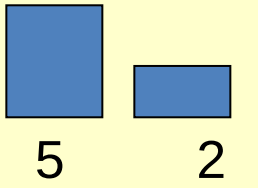


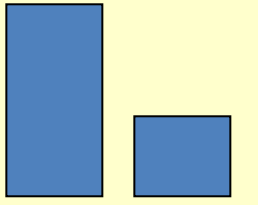


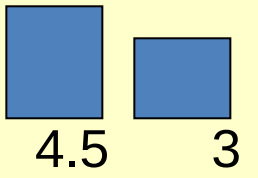


Examples of class B

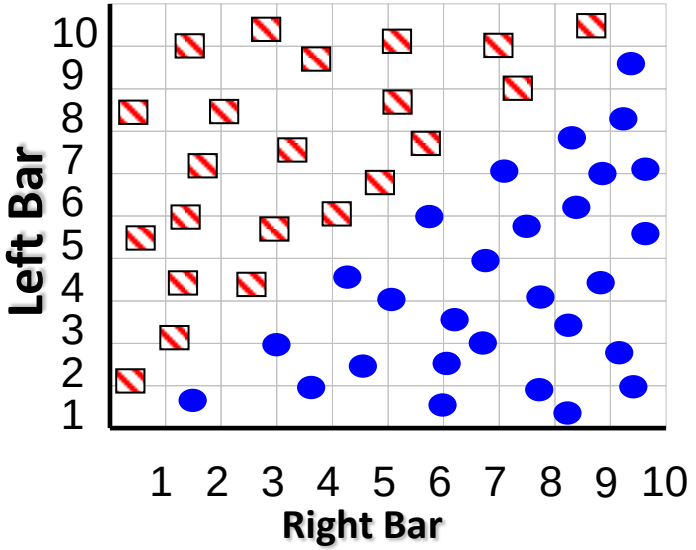






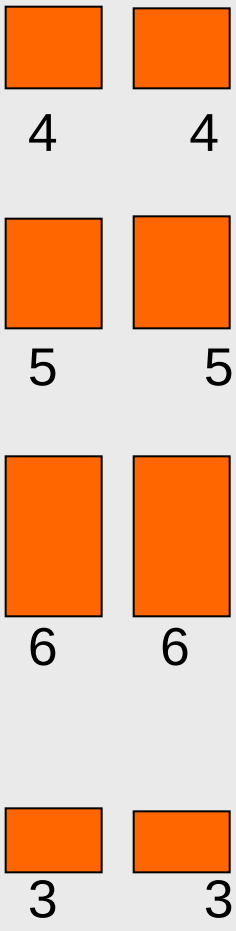


分类规则：
If the left bar is smaller than the right bar, it is an A otherwise it is a B.

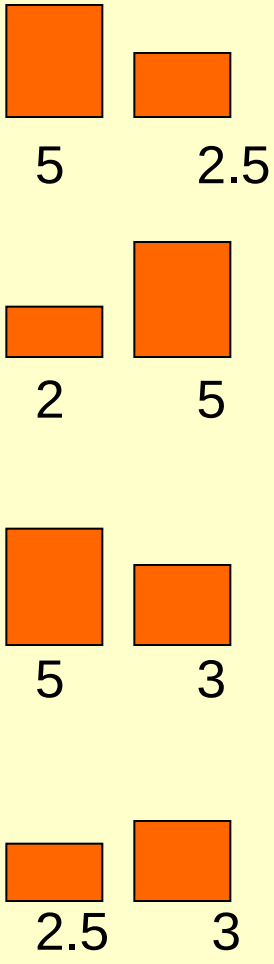


数据挖掘实例：分类问题

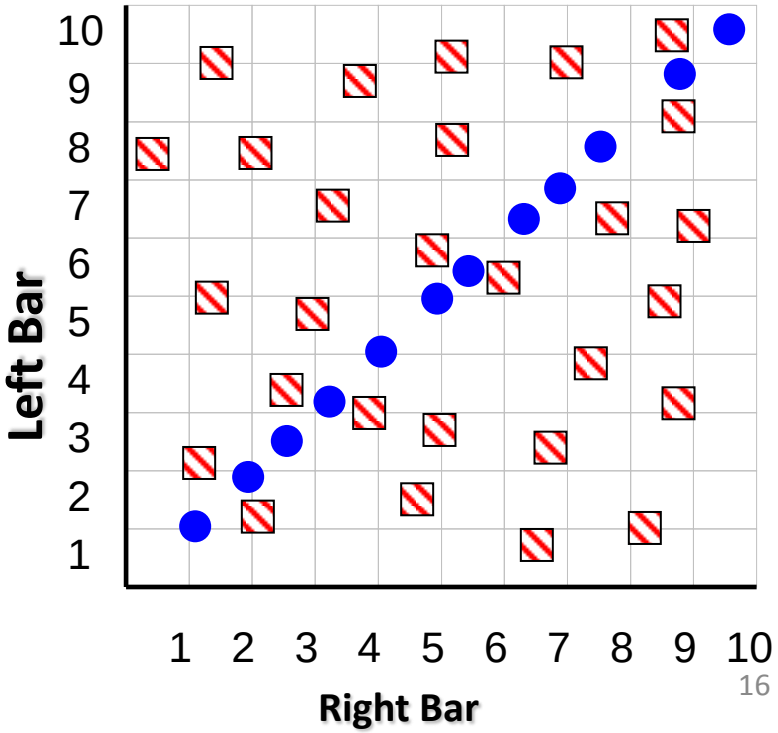
Examples of class A



Examples of class B

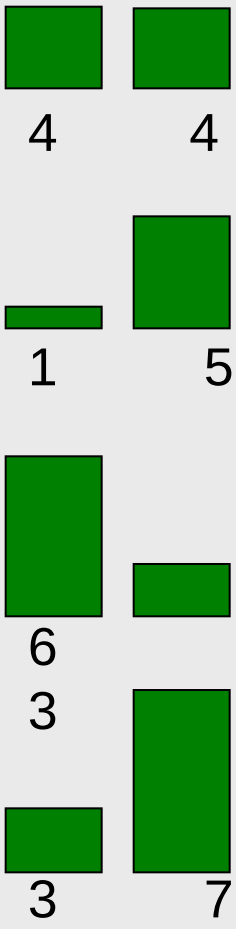


if the two bars are equal sizes, it is an *A*.
Otherwise it is a *B*.

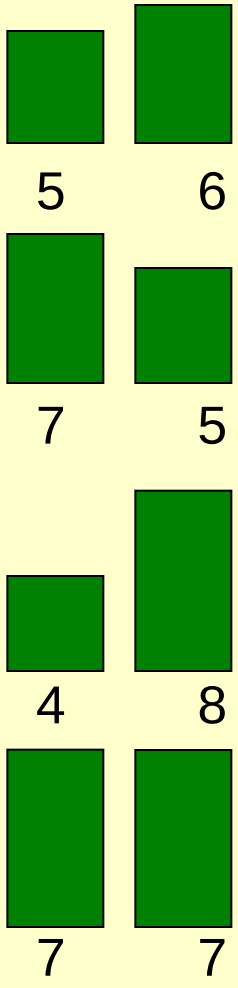


数据挖掘实例：分类问题

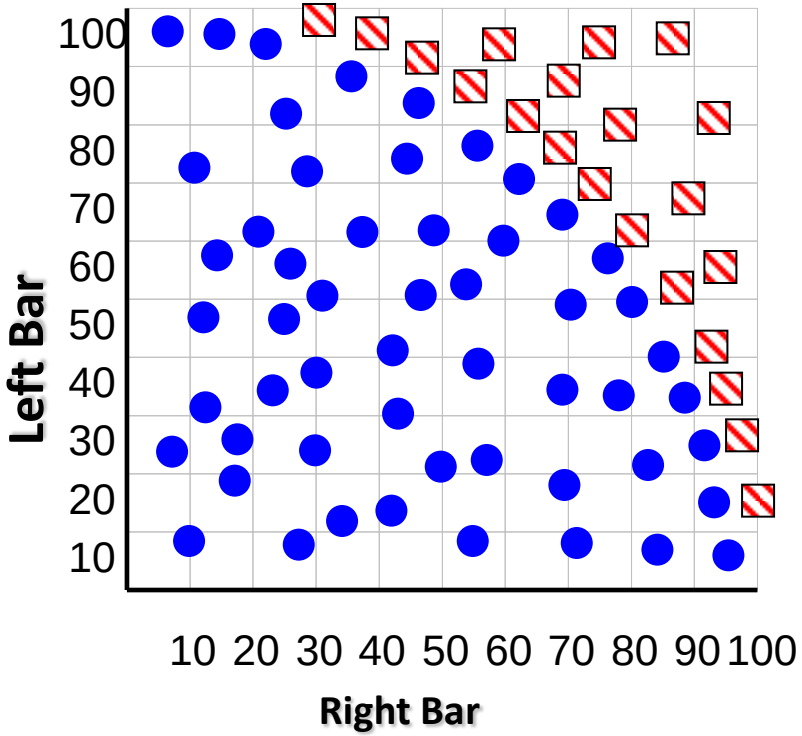
Examples of class A



Examples of class B



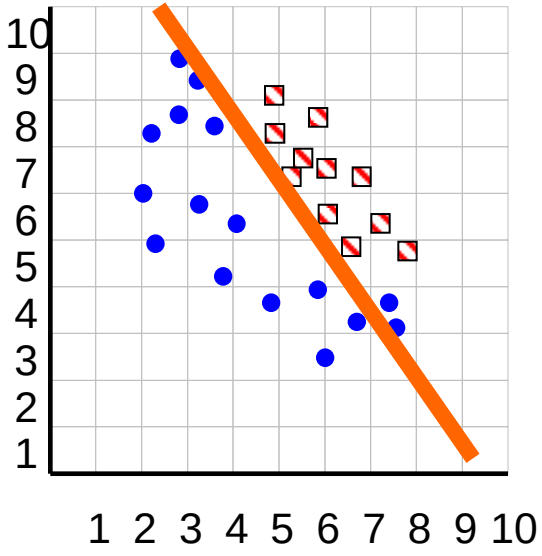
if the square of the sum of the two bars is less than or equal to 10, it is an A. Otherwise it is a B.



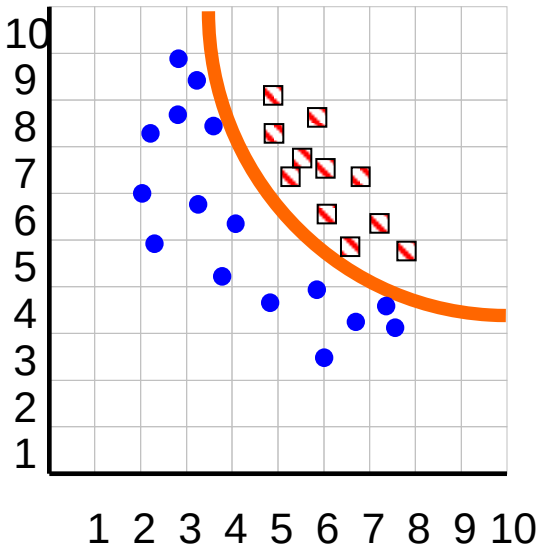
数据挖掘目标

目标：找到f(x)拟合已观测数据！

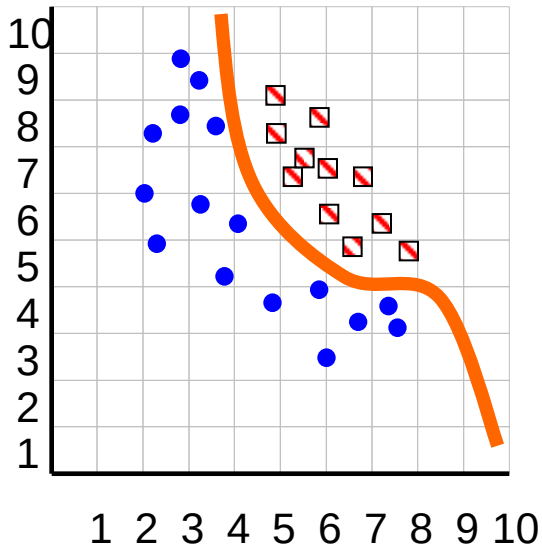
Accuracy = 94%



Accuracy = 100%



Accuracy = 100%



数据挖掘：分类问题

- **按训练方式，机器学习可分为：**

（1）有监督的学习；有训练样本，学习机通过学习获得训练样本包含的知识，并用其作为判断测试样本的类别的依据。

（2）无监督的学习：无训练样本，仅根据测试样本的在特征空间分布情况判断其类别。

（3）半监督的学习：有少量训练样本，学习机以从训练样本获得的知识为基础，结合测试样本的分布情况逐步修正已有知识，并判断测试样本的类别。

（4）强化学习：没有训练样本，但有对学习机每一步是否更接近目标的奖惩措施。

数据挖掘方法实例

- 关联规则
- 决策树
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

关联规则挖掘

- 关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。设$\boldsymbol{I}=\{i_1,i_2,\ldots,i_m\}$是项的集合。设任务相关的数据$D$是数据库事务的集合，其中每个事务$T$是项的集合，使得$T\subseteq \boldsymbol{I}$。设$A$是一个项集，事务$T$包含$A$当且仅当$A\subseteq T$。

事先已知

事先未知

- 超市中什么产品会一起购买？— 啤酒和尿布
- 在买了一台PC之后下一步会购买?
- 哪种DNA对这种药物敏感?

定义：从数据集中找出对象或项集之间同时发生的关联或顺序关系。

数据挖掘方法实例

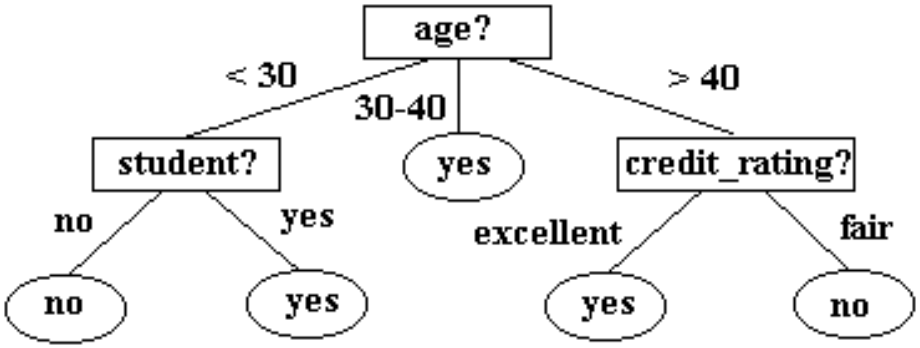
- 关联规则
- 决策树
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

数据挖掘方法-决策树

- 决策树学习是归纳推理算法。它是一种逼近离散函数的方法，且对噪声数据有很好的健壮性。在这种方法中学习到的知识被表示为决策树，决策树也能再被表示为多个if-then的规则，以提高可读性。

实例：通过年龄、收入、是否为学生、信用记录来预测用户是否会购买电脑

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

数据挖掘方法-朴素贝叶斯（Naive Bayes）分类器

- 朴素贝叶斯分类器是一种基于贝叶斯理论的分类器。它的特点是以概率形式表达所有形式的不确定，学习和推理都由概率规则实现，学习的结果可以解释为对不同可能的信任程度。
- $P(H)$是**先验概率**，或H的先验概率。$P(H|X)$是**后验概率**，或条件X下，H的后验概率。后验概率$P(H|X)$比先验概率$P(H)$基于更多的信息。$P(H)$是独立于X的。

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} = \frac{P(X \mid H)P(H)}{\sum_{H} P(H)P(X \mid H)}$$

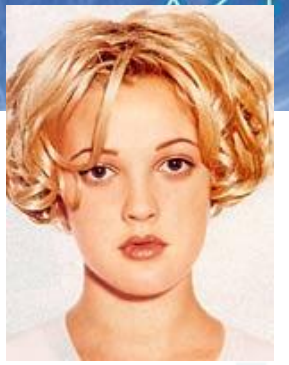
朴素贝叶斯分类能够奏效的前提是，$P(X/H)$相对比较容易计算。假定X表示红色和圆的，H表示假定X是苹果；则$P(X/H)$表示已知苹果，它既红又圆的概率。

假设有如下两个类别：

c_1 = male, and c_2 = female.

预测名字为 “*drew*” 是male or female, i.e *p*(male | *drew*) or *p*(female | *drew*)哪个大？

(Note: “Drew can be a male or female name”)



Drew Barrymore

Drew Carey

给定性别为 “male”,名字为 “*drew*” 的概率?

性别为 male的概率?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

名字为 “*drew*”的概率?
(actually irrelevant, since it is that same for all classes)

Officer Drew

This is Officer Drew (who arrested me in 1997). Is Officer Drew a Male or Female?

假设数据库中已有名字和性别的对应数据。可以以此为已观测数据来应用贝叶斯分类器。

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Officer Drew

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid \textit{drew}) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid \textit{drew}) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

Officer Drew 的性别更可能为Female.

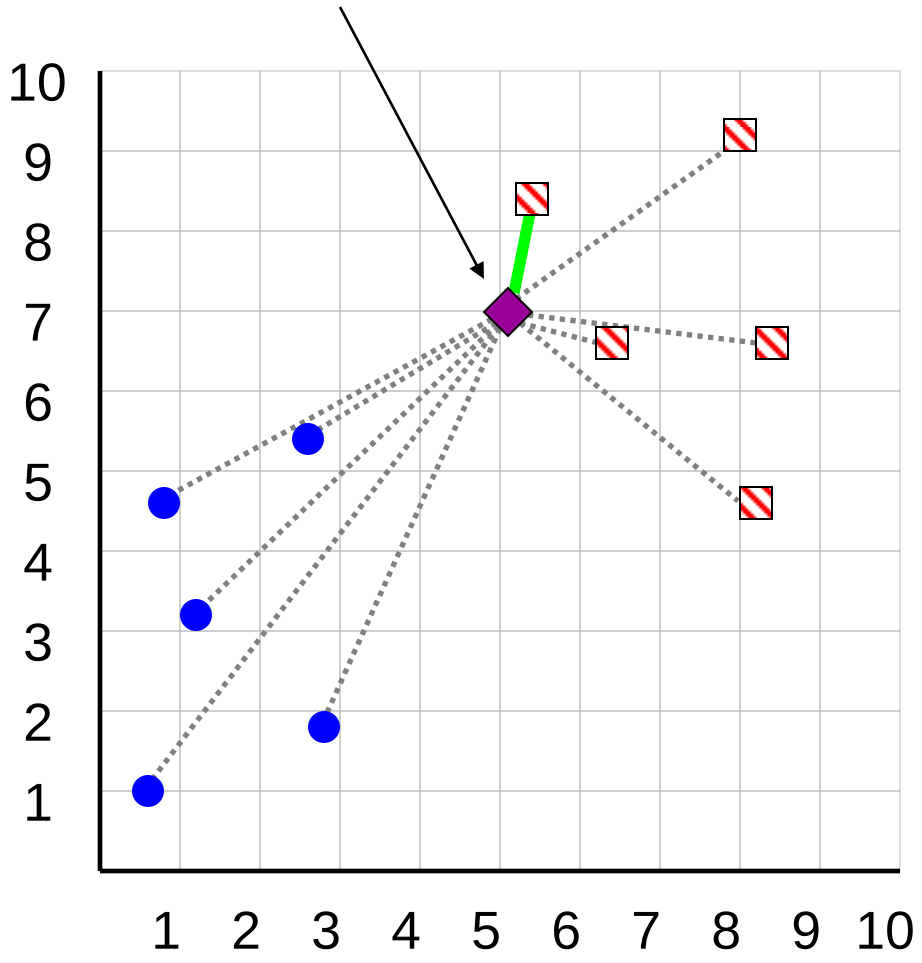
数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

数据挖掘方法实例 K-最近邻分类

- K-近邻（K-NN）分类是基于范例的分类方法，它的基本思想是：给定待分类样本后，考虑在训练样本集中与该待分类样本距离最近（最相似）的K个样本，根据这K 个样本中大多数样本所属的类别判定待分类样本的类别。
- 它的特例是1- NN，即分类时选出待分类样本的最近邻，并以此最近邻的类标记来判断样本的类。
- K-NN算法的优点在于它有较高的精确程度，研究表明，K-NN的分类效果要明显好于朴素贝叶斯分类、决策树分类。

数据挖掘方法实例 K-最近邻分类



Evelyn Fix
1904-1965

Joe Hodges
1922-2000

If the **nearest** instance to the previously unseen instance **is a A**
class is **A**
else
class is **B**

A
B

数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

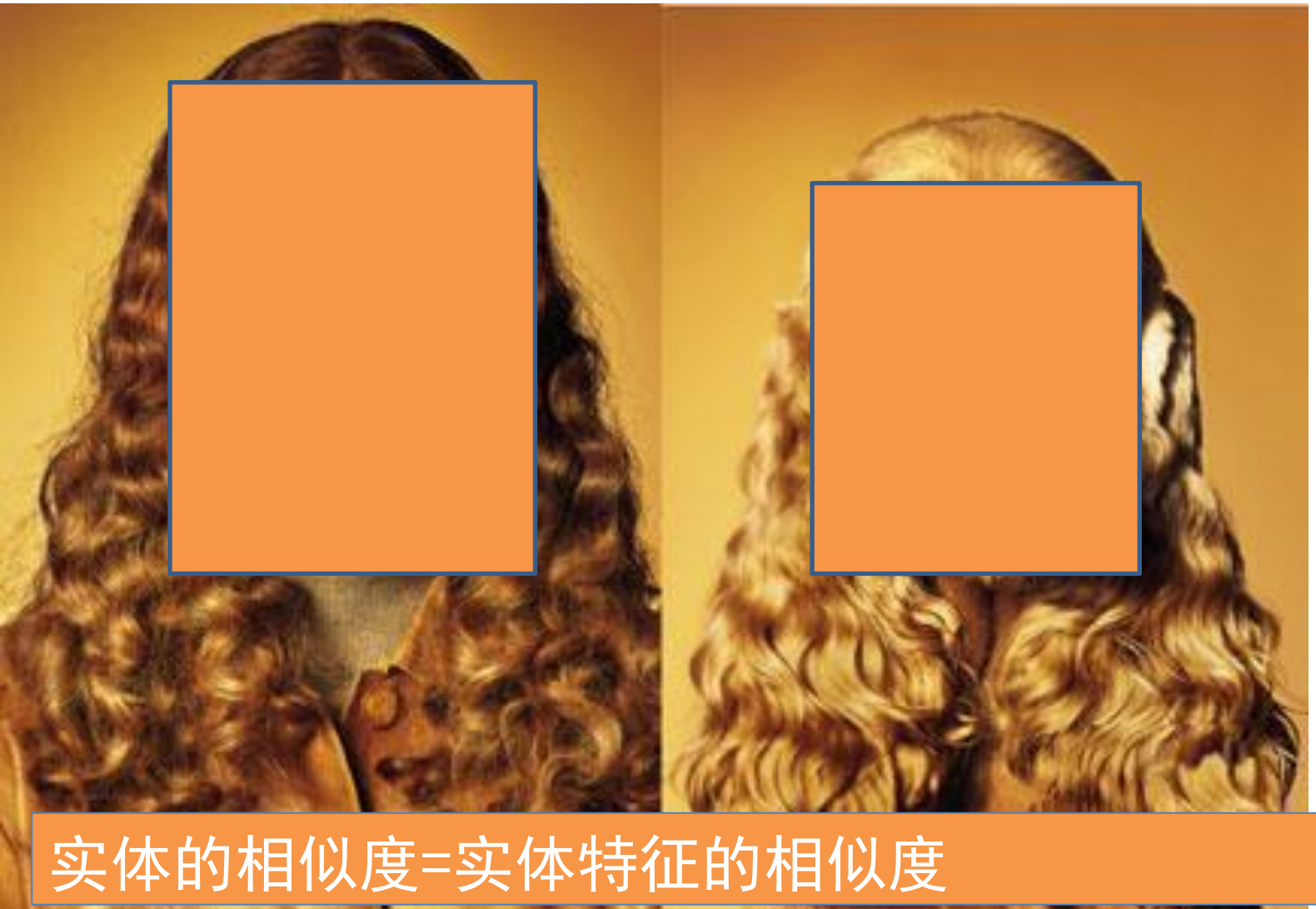
聚类方法分类

- 基于层次的方法：层次的方法对给定数据集合进行层次的分解。根据层次的分解如何形成，层次的方法可以被分为凝聚或分裂方法。（Chameleon，CURE，BIRCH）
- 基于密度的方法：只要临近区域的密度超过某个阈值，就继续聚类。避免仅生成球状聚类。（DBSCAN，OPTICS，DENCLUE）
- 基于网格的方法：基于网格的方法把对象空间量化为有限数目的单元，所有的聚类操作都在这个量化的空间上进行。这种方法的主要优点是它的处理速度很快。（STING，CLIQUE，WaveCluster）
- 基于模型的方法：为每个簇假设一个模型，发现数据对模型的最好匹配。（COBWEB，CLASSIT，AutoClass）

聚类–定义相似度

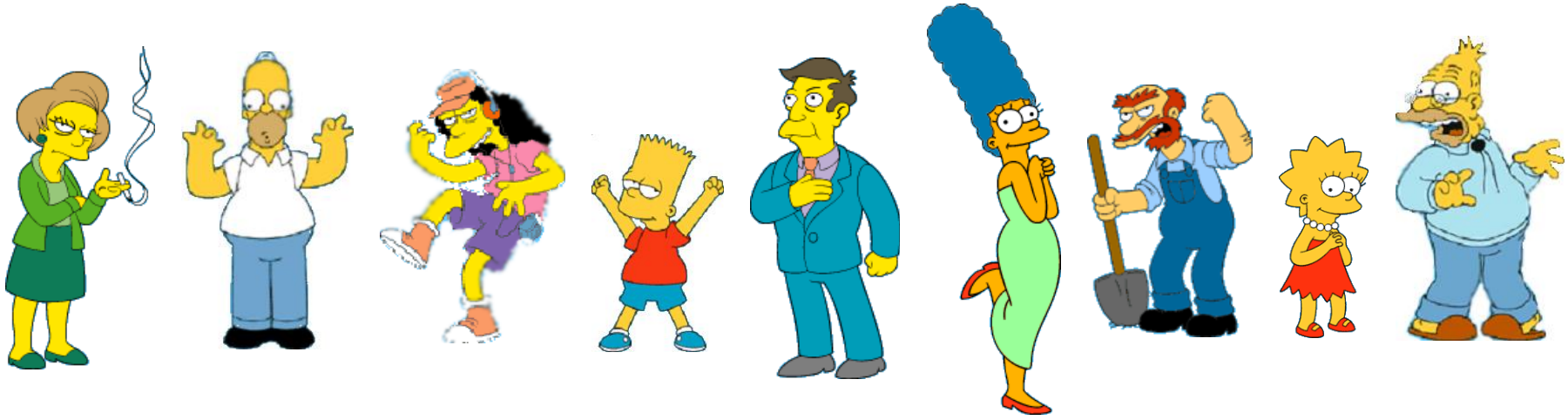
The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary

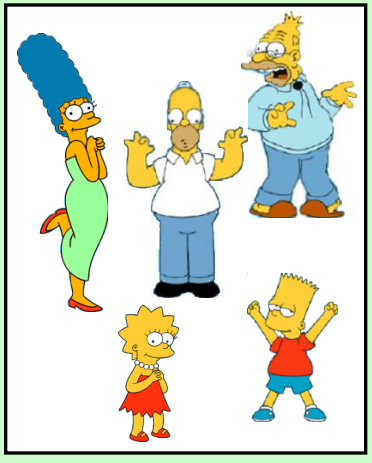


实体的相似度=实体特征的相似度

可能的聚类方式

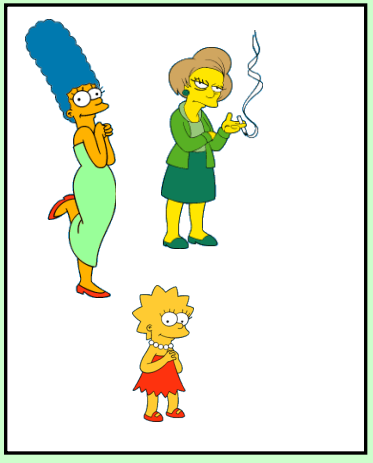


Clustering is subjective !



Simpson's Family

School Employees

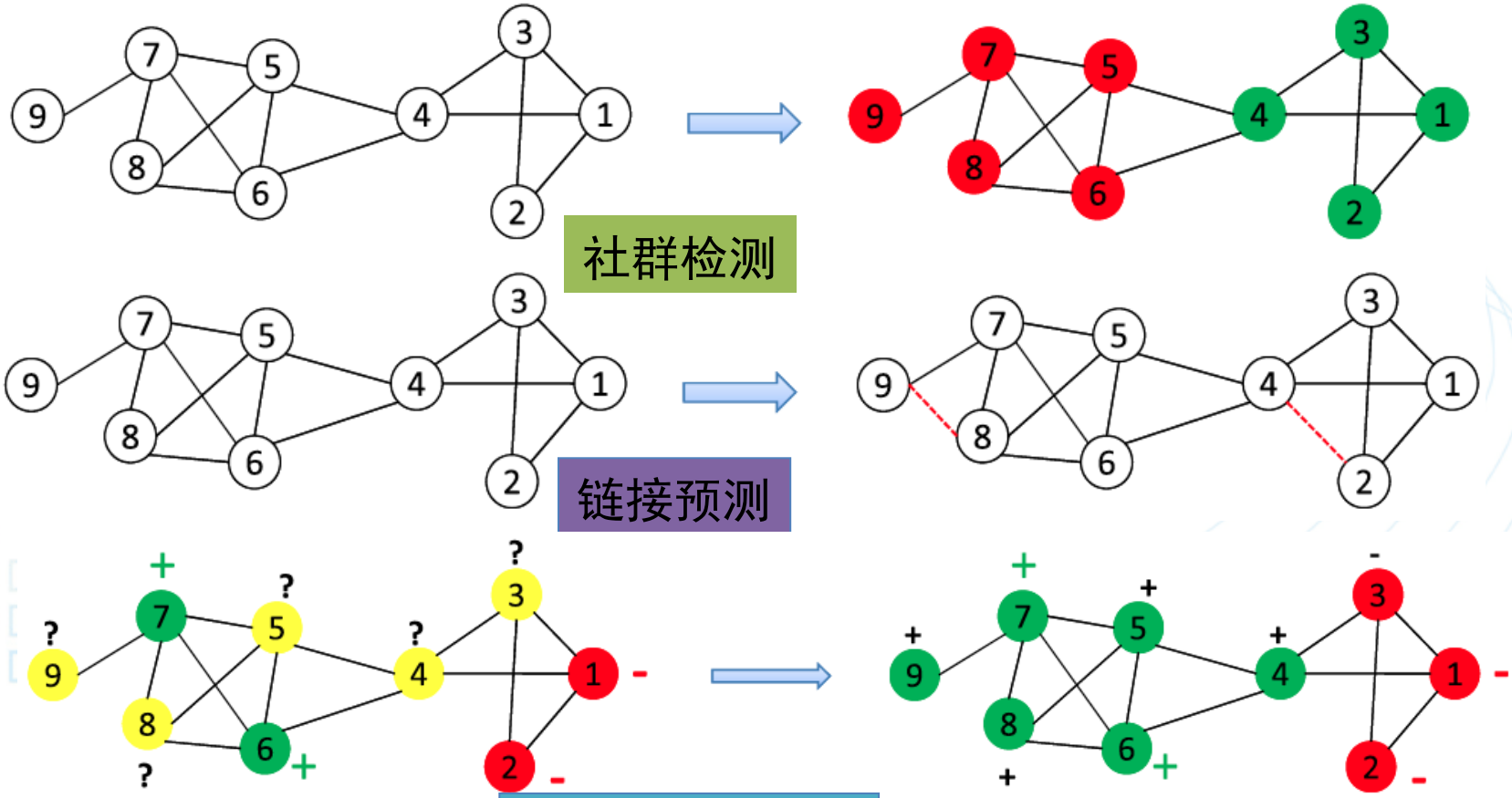


Females

Males

社交网络应用：分类、聚类和推荐

- 数据挖掘技术在社交媒体分析中的应用
  - Tag suggestion, Product/Friend/Group Recommendation



性能评估

分类精度评价指标

- 理想的分类器应该将所有属于某一类的样本标记为该类；且不将任何一个不属于该类的样本标记为该类。可以采有两个指标用来评价分类器的性能：准确率（查准率）和召回率 (查全率)。对于某一特定类别C_i，

- 准确率(P) = $\frac{\text{分类属于}C_i\text{且实际属于}C_i\text{的样本数}}{\text{分类属于}C_i\text{的样本数}}$

- 召回率(R) = $\frac{\text{分类属于}C_i\text{且实际属于}C_i\text{的样本数}}{\text{实际属于}C_i\text{的样本数}}$

分类精度评价指标（续）

- 对于同一分类器，这准确率和查全率的变化趋势通常是相反的，片面追求其中一个指标而完全不顾及另一个是没有意义的。
- 为综合考虑准确率和查全率，可以使用一种能够全面评价分类器性能的指标：F-1。
- $$F\text{-}1 = \frac{2 \times 查准率 \times 查全率}{查准率 + 查全率}$$
- F-1综合考虑了上述两指标，且偏向于准确率和查全率中较小的一个，只有当准确率和查全率都较大时，F-1指标才会比较大。

分类精度评价指标（续）

- 多数分类器可以通过调整参数获得不同的准确率和查全率，当分类器的参数调节到正好使准确率和查全率相等时，该值称为P/R无损耗(平衡)点。它也是一种综合考虑准确率和查全率的指标。
- 在综合考虑全部类别的条件下，精确度(Accuracy)也是一个常用的指标，它是指所有分类正确的样本数在所有样本中所占的比例。
- 精确度(A) = $\frac{\text{所有分类标记正确的样本数}}{\text{全部样本总数}}$


谢谢