# 数据挖掘算法及其延伸

# Outline:

## Data Ming概念、用处及分析

## 数据归一化、处理及计算

## 关联规则、建立及实现

## 聚类算法

## 分类算法

# Data Mining

### 动机

1. 数据的指数级增长
2. 数据的收集和可用性
3. 数据库种类的多样性
4. 大量的数据需被收集和存储
5. 提取和使用数据背后隐藏的信息，存在巨大的竞争压力
6. 需要更强的数据分析技术和研究方向上的资助
7. 知识方面的匮乏
8. 大量数据集的自动分析

### 概念

1. 在大量的数据库中发现有用、能够理解的模式
2. 交叉性学科
3. 大规模、高维度的数据
4. 数据的复杂度较高
5. 应用程序的更新和复杂

### 特征、用途

1. 大量的数据集
2. 从历史数据中，自动搜索感兴趣的模式
3. 快速、大规模
4. 易于更新、易于实践
5. 有相应的决策支持

### 模式类型

**Associated rules（关联规则 ）**

**Detect sets of attributes or items that frequently co-occur in many database records and rules among them.**

**Classification(分类)**

**Build a model of classes on training dataset, and then, assign a new record to one of several predefined classes.**

**Presentation:Decision Tree**

**Clustering(聚类)**

**Partition the dataset into groups such that elements in a group hae lower inter-group similarity and higher intra-group similarity.**

**(将数据集划分成组，使得组中的元素具有较低的组间相似性和较高的组内相似性。)**

**Sequence mining(序列挖掘)**

**Give a set of sequences, find the complete set of frequent subsequences.**

**(给出一组序列，找到一组完整的频繁子序列。)**

**Anomaly detection(异常检测)**

**Give a set of n objects, and k, the number of expected anomolies, find the top k objects that are considerably dissimilar or inconsistent with the remaining data.**

**(给出一组n个对象，并且k（预期的不规则数）找到与剩余数据非常不一致或不一致的前k个对象。)**

**Community Analysis(社区分析)**

**Media mining(媒体挖掘)**

**Recommender Systems(推荐系统)**

**Recommend products that would be interesting to**

**Individuals.**

**Build a function, f : U X I 🡪 R, for user set U and item set I.**

### 数据类型

**Database-oriented data sets and applications**

**Relational Database关系数据库**

**Structured data(结构性数据)**

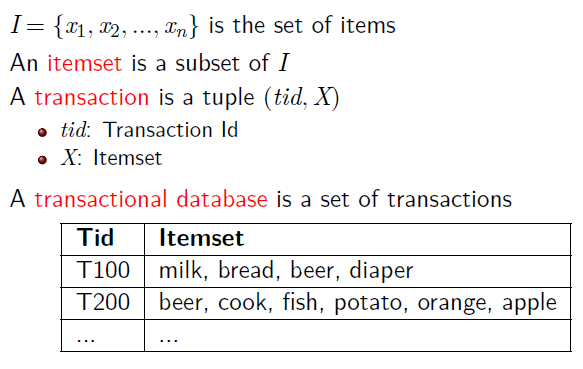
**表结构属性**

**查询访问,SQL**

**Data Warehouse**

**(面向主题，集成，清理的数据集合，用于支持管理层的决策过程)**

**Transactional Database**



**Advanced database applications**

**Data stream**

**Spatial data**

**Text databa**

**Multi-media data**

**Time-series**

**Bio-medical data**

**Network traffic dat**

# Knowledge Discovery(KDD) Process(Core)

|  |
| --- |
|  |

# KDD过程的关键步骤：

1. 应用领域的学习

相应的先验知识和应用的目的

1. 构造目标数据资源
2. 数据清洗和预处理
3. 数据降维和变换
4. 选择挖掘算法搜寻感兴趣的模式
5. 获取知识的用途

**Interesting Patterns’Metrics:**

**A pattern is interesting if it is easily understood by humans,valid on new or test data with some degree of certainty,potentially useful,novel,or validates some hypothesis that a user seeks to confirm.**

**(如果一个模式很容易被人理解，那么这个模式很有意思，它对新的或测试数据有一定的确定性，潜在的有用性，新颖性，或者验证用户想要确认的一些假设。)**

### <1>数据挖掘的研究问题

1. **挖掘方法**
2. **从不同的数据类型挖掘不同类型的数据**
3. **高效性和可扩展性**
4. **并行、分布式、鲁棒式的挖掘方法**
5. **模式评估**
6. **处理噪声点和不完整的数据**
7. **背景知识的合并**
8. **用户交互**
9. **数据挖掘查询语言**
10. **数据挖掘结果的表示和可视化**
11. **应用和社会影响**
12. **特定领域的数据挖掘**
13. **保护数据的安全性、完整性和隐私性**

### <2>在数据挖掘研究上面临的挑战

1. 开发数据挖掘的统一理论
2. 扩展**高维**数据和**高速**数据流
3. 挖掘**序列**数据和**时间序列**数据
4. 从复杂数据中挖掘复杂数据
5. **网络**设置中的数据挖掘
6. **分布式**数据挖掘和挖掘**多智能体**数据
7. 生物和环境的数据挖掘
8. 数据挖掘过程相关问题
9. 安全、隐私和数据完整性
10. 处理**非静态**、**不平衡**和成本**敏感**的数据

# 数据处理

### 预处理数据的原因

1. 不完整性

数据值的不适用性

收集数据和分析数据时有不同的考虑

人为、硬件或者软件的问题

缺失属性

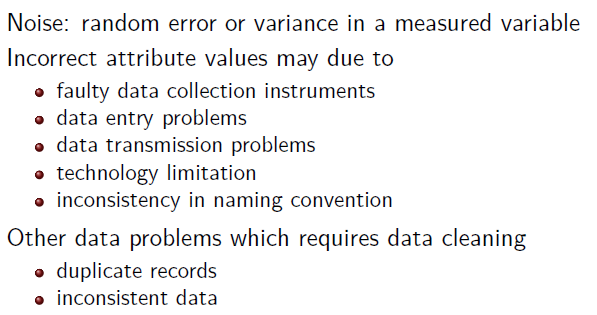
1. 噪点/不正确的值

收集的工具有问题

数据录入时出现人为或计算机的错误

数据传输错误

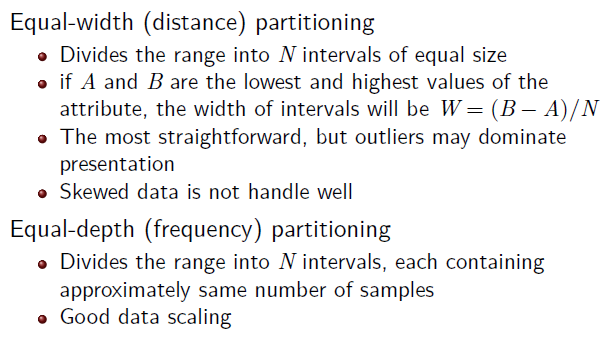
包含错误或奇异值点



如何处理噪声数据



**Simple Discretization Methods: Binning分箱**



**Binning Methods for Data Smoothing**

**Regression**

**Cluster Analysis**

1. 不连贯性

不同的数据来源

违反功能的依赖关系

重复的记录也需要数据清洗

包含差异性

**Note：**

**没有高质量的数据，也就没有好的挖掘结果！**

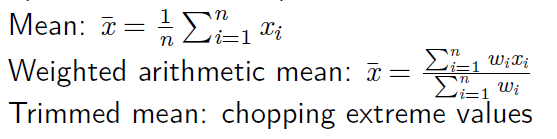
**好的决策必定取决于好的数据！**

**数据提取，清理和转换包括构建数据仓库的大部分工作！**

### 数据的描述汇总

Central tendency

均值、加权平均



分散特征

·中位数、最大值、最小值、分位数、异常值、方差

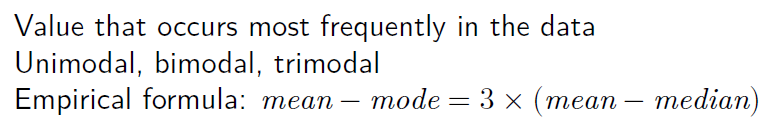
Median(如果是奇数个值，则为中间值，否则为中间两个值的平均值)

数字维度对应于已排序的时间间隔

·数据分散：以多个精度粒度进行分析

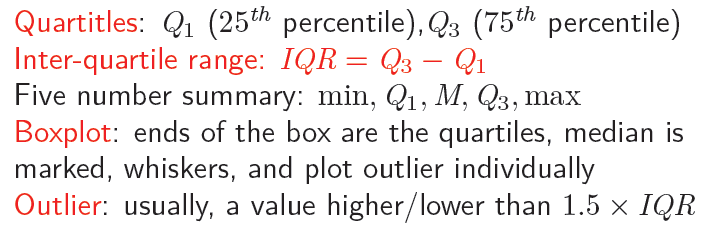
·Boxplot或按排序间隔进行量子分析

Mode



**Measuring the Dispersion of Data(测量数据的色散)**

**四分位数，异常值和箱形图**



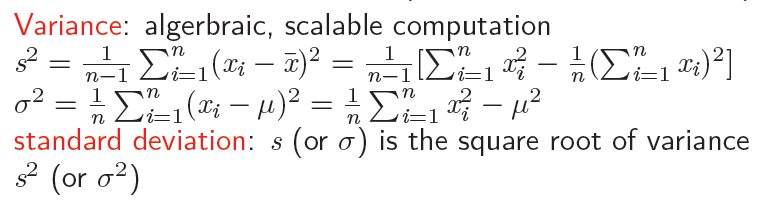
**四分位距离：IQR = Q3 – Q1**

**min; Q1; M; Q3; max**

**Boxplot：(盒子的末端是四分位数，中位数被标记，whiskers，并单独绘制离群值)**

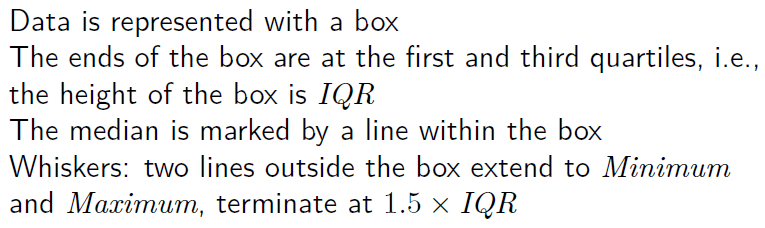
**Outlier：a value higher/lower than 1.5 X IQR**

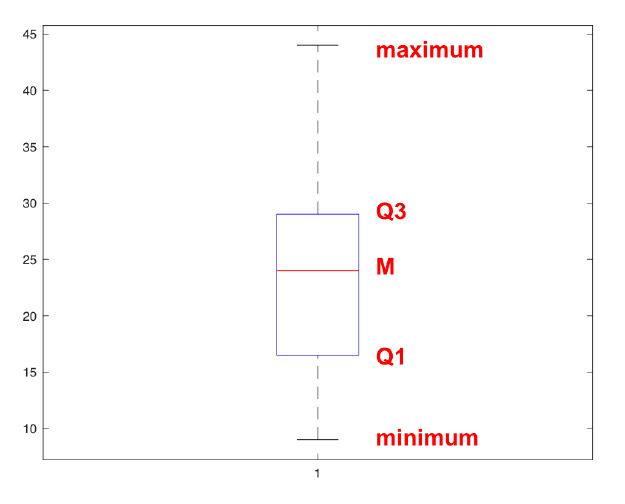
**方差和标准差**



**Boxlpot Analysis**

**Minimum; Q1; M; Q3; Maximum**





**Properties of Normal Distribution Curve**

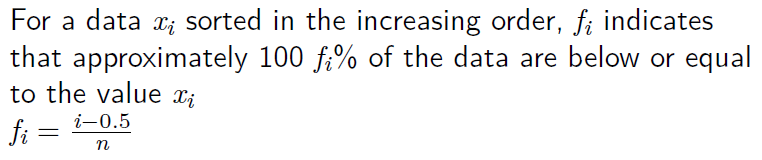
**Histogram Analysis**

**Quantile Plot**

显示所有数据（允许用户评估整体行为和异常事件）

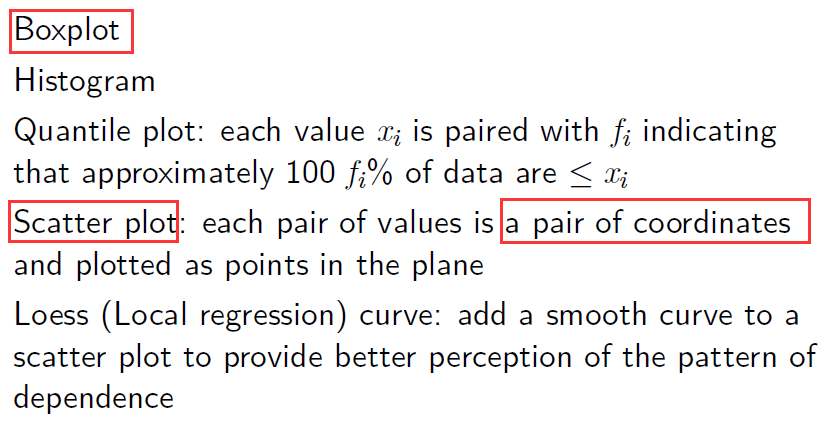
Plot quantile information

对于按递增顺序排序的数据xi，fi表示大约100％的数据低于或等于xi

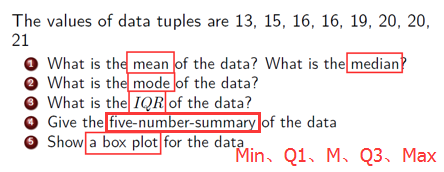


**Scatter Plot**

**Loess Curve**



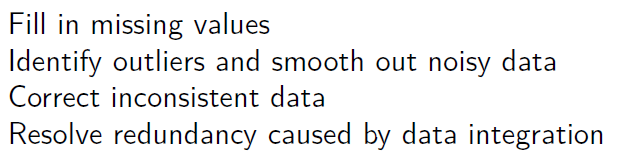
**考点一**



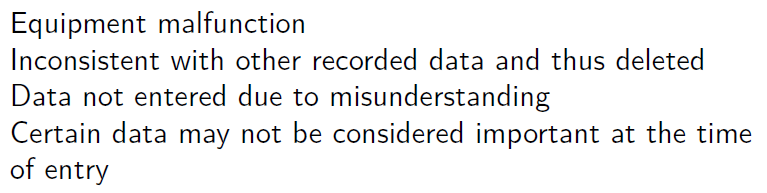
### 数据清洗

填充缺失值，平滑噪声数据，识别或删除异常值，并解决不一致问题。

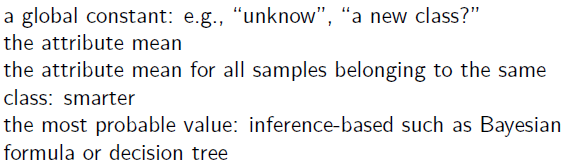
数据清洗任务



丢失数据

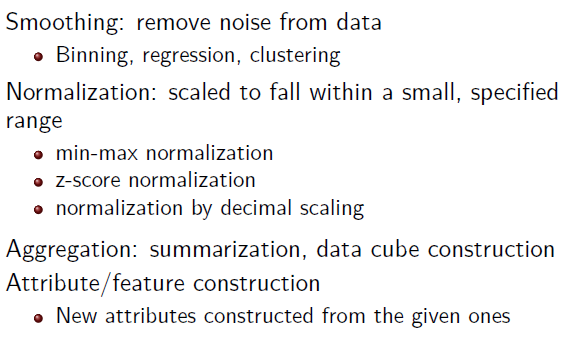


如何处理丢失数据

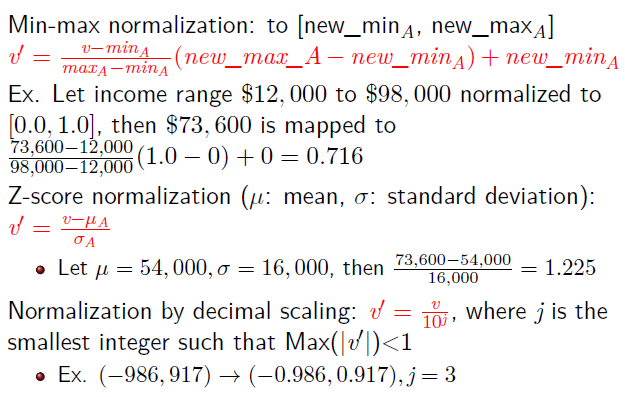


### 数据变换

规范化和聚合。



归一化

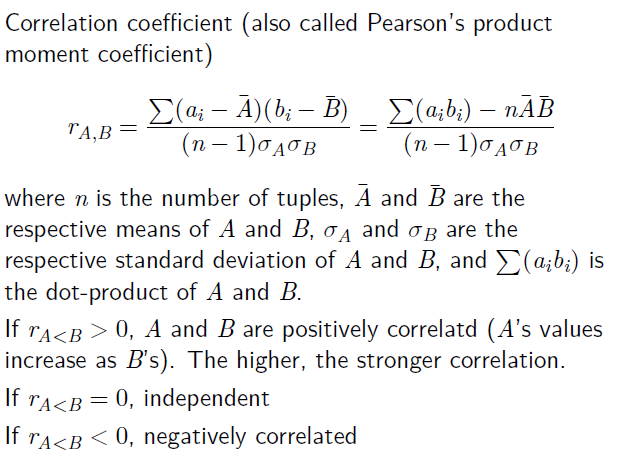


### 数据整合

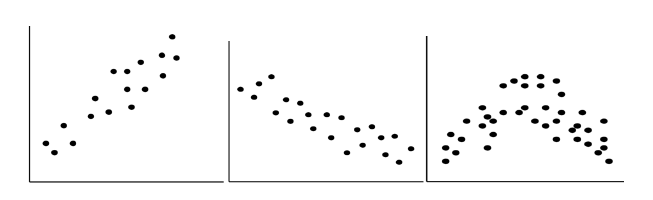
整合多个数据库或文件。



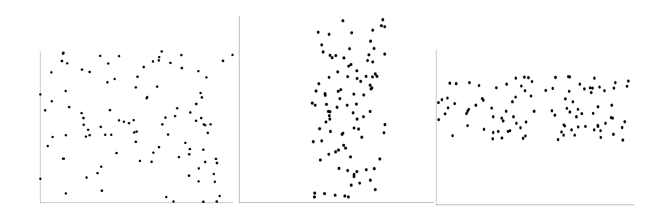
**Handling Redundancy in Data Integration**

**Correlation Analysis (Numerical Data)** 

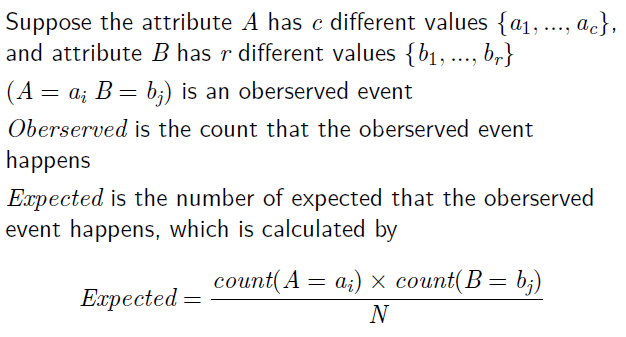
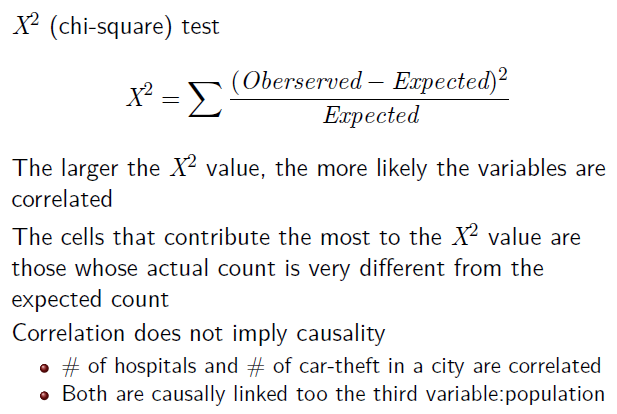
**Positive and Negatively Correlated Data**



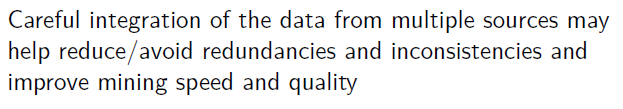
**Independent Data**



**Correlation Analysis (Categorical Data)**

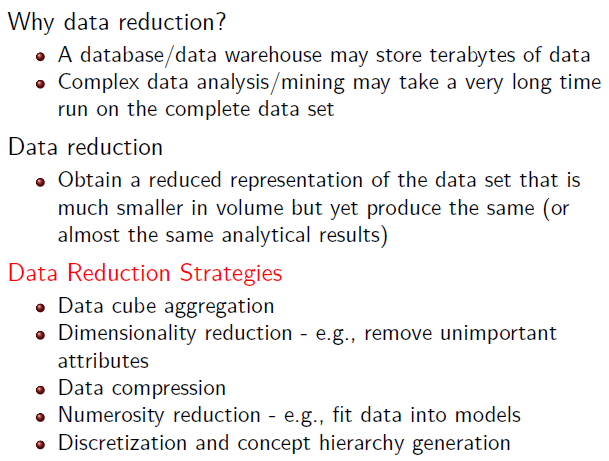


**Handling Redundancy in Data Integration**



### 数据降维

获得体积减少的表示，但产生相同或类似的分析结果。



**Data Cube Aggregation**

**Attribute Subset Selection**

**特征选择Feature selection(子集属性选择)**

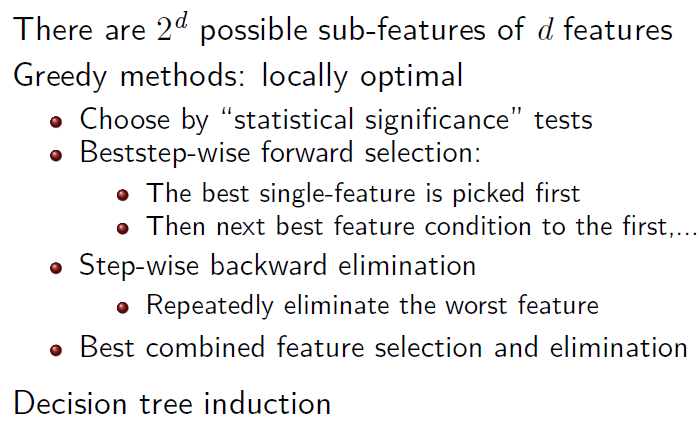
**选择一组最小的特征，使给定这些特征值的不同类别的概率分布尽可能接近原始分布，给定所有特征的值；**

**减少模式中的特征数量，更易于理解。**

**启发式方法**



**特征选择方法**



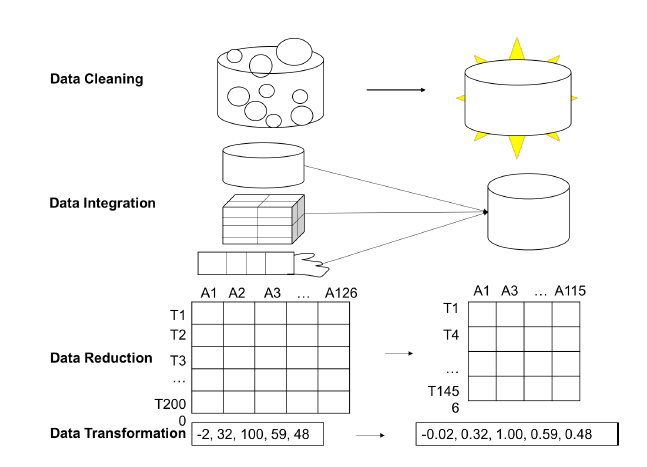
**Forward selection**

**Backward elimination**

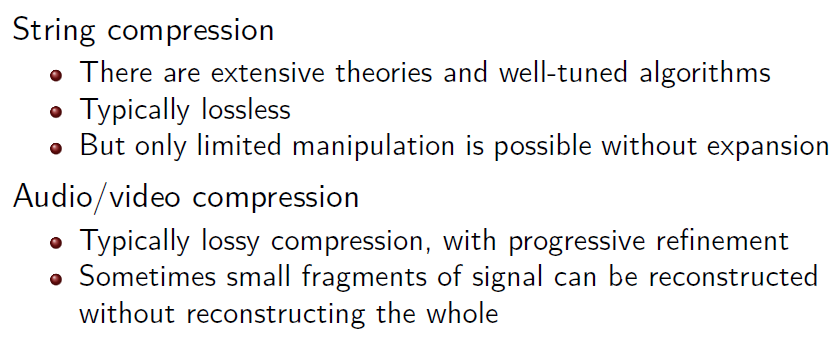
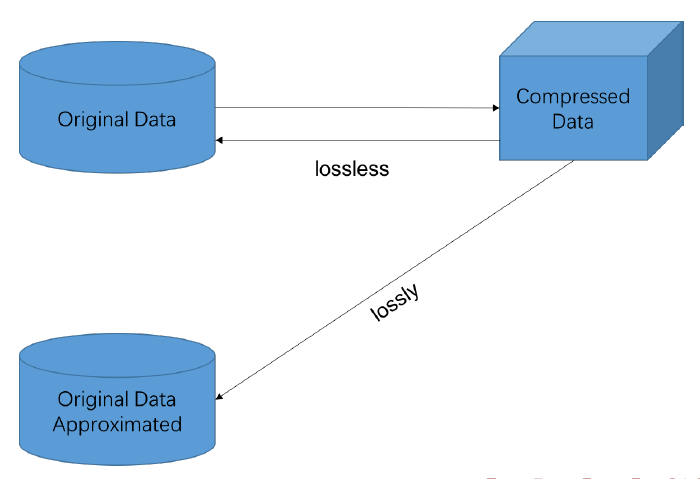
**Decision tree induction**

### 离散和概念层次的生成

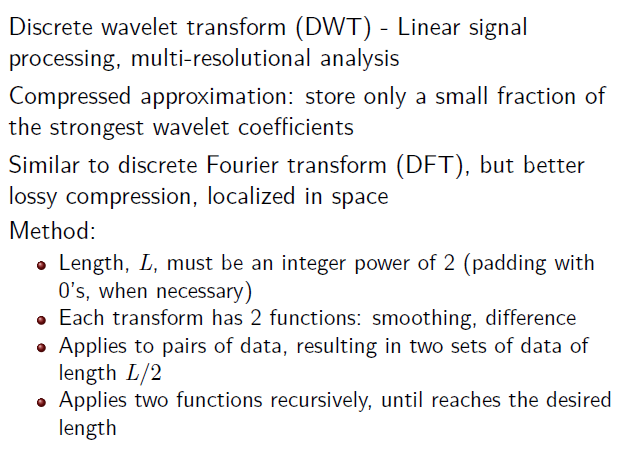
Data discretization：减少数值型数据



### 数据压缩

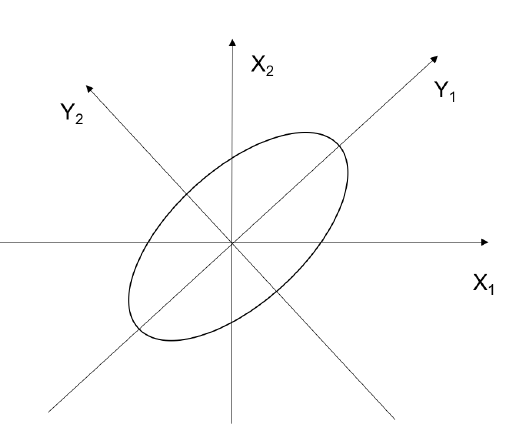
 

**Wavelet Transformation(小波变换)**



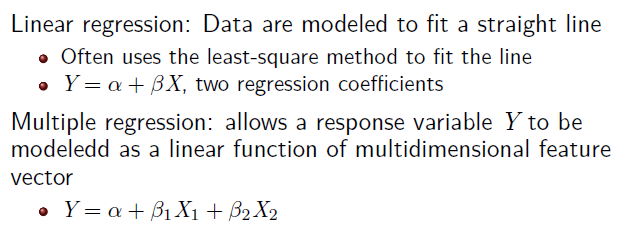
**降维Dimensionality Reduction:**

**Principle Component Analysis(PCA)**



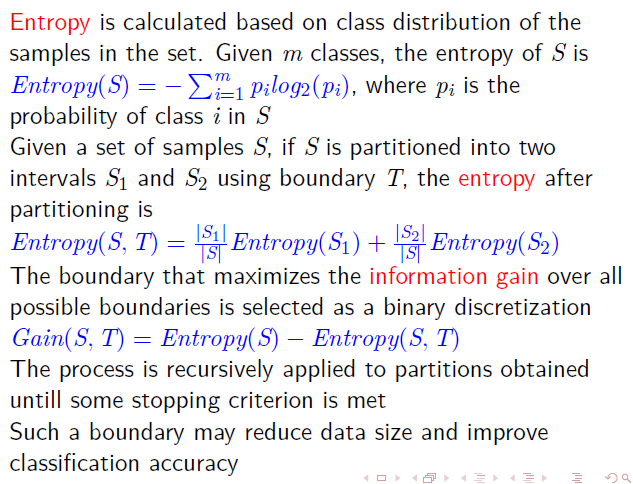
**Data Reduction Method**

1. **Regression Models**



1. **Histograms**
2. **Clustring**
3. **Sampling**

**Entropy-Based Discretization**



# 关联规则(Associated Rules)

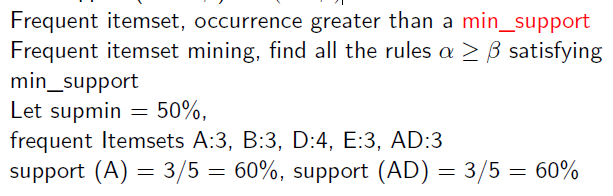
### 概念

关联规则挖掘

发现频繁模式，在交易数据库，关系数据库和其他信息属性的对象集合中找出关联。

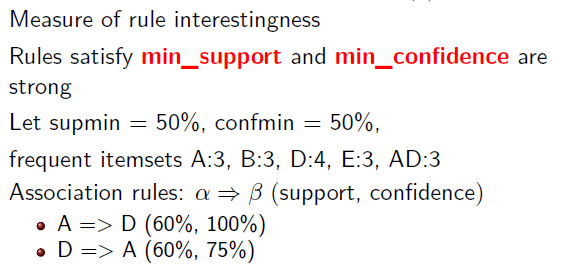
**Support：支持度**

**support(A ==> B) = P(A n B)**

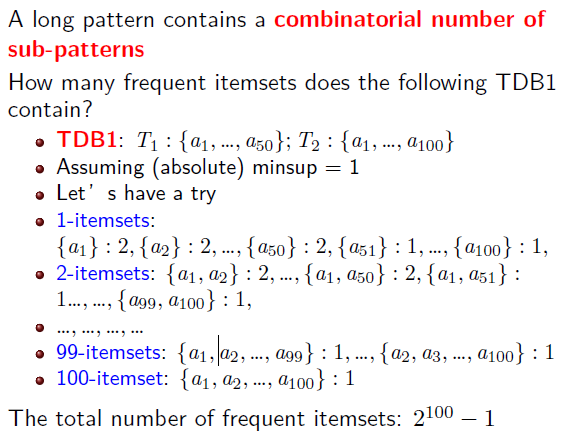


**Confidence：置信度**

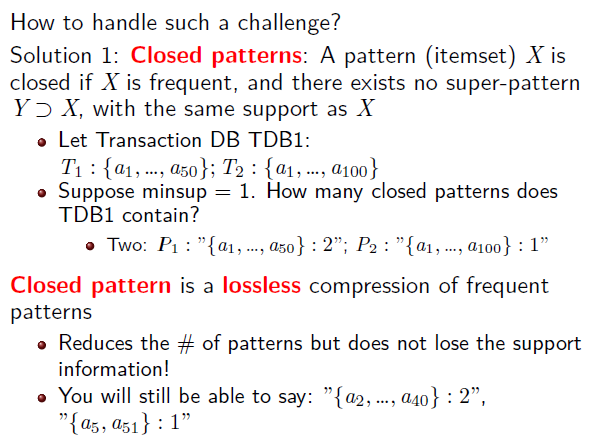
**Confidence(A ==> B) = P(A n B)/P(A)**



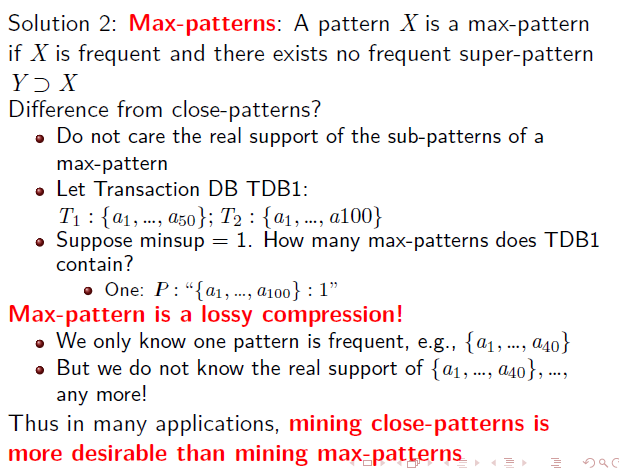
**Too Many Frequent Pattern**



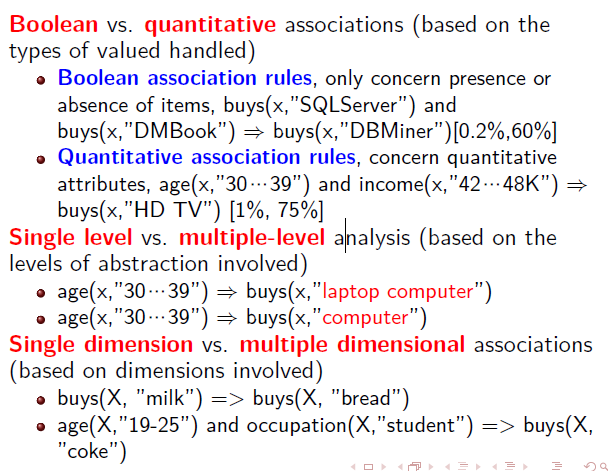
**Closed Pattern**



**Max Pattern**

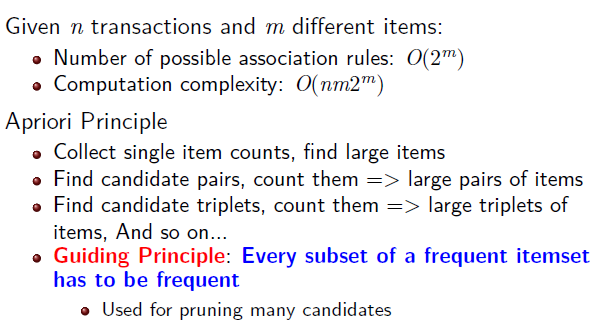


**Association Rule Mining**



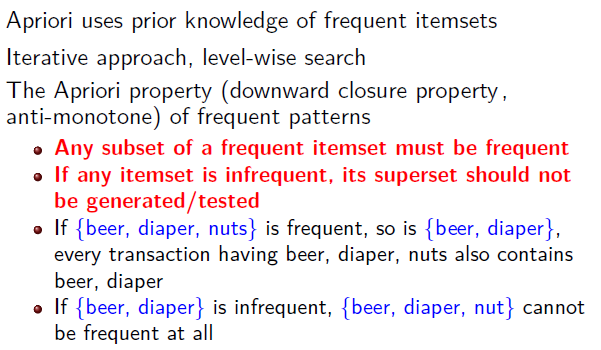
**Mining Frequent Patterns,Association & Correlations**

**处理指数级别的复杂度**

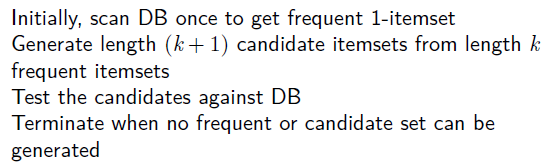


**如果一个项集是非频繁的，那么它的搜友超集也是非频繁的！**

### (二)Apriori：候选生成和测试方法



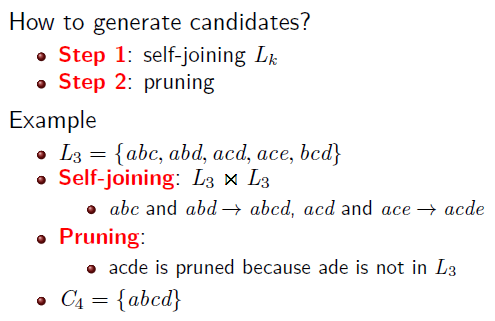
**Apriori 方法:**



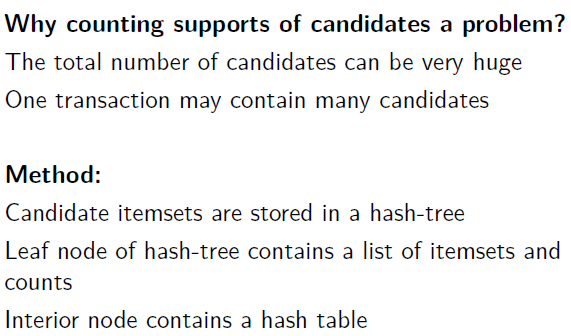
**Apriori算法示例：**



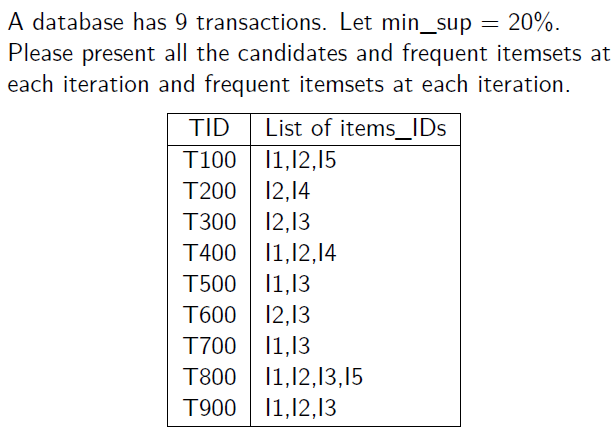
**How to Generate Candidates??**



**How to Count Supports of Candidates??**



**考点二**



**Minsup = 0.2 X 9 = 1.8**

**1st scan: C1🡪( F1)**

**{I1} 6**

**{I1} 7**

**{I1} 6**

**{I1} 2**

**{I1} 2**

**C2->**

**{I1,I2}**

**{I1,I3}**

**{I1,I4}**

**{I1,I5}**

**{I2,I3}**

**{I2,I4}**

**{I2,I5}**

**{I3,I5}**

**2nd scan:C2->**

**{I1,I2} 4**

**{I1,I3} 4**

**{I1,I4} 1**

**{I1,I5} 2**

**{I2,I3} 4**

**{I2,I4} 2**

**{I2,I5} 2**

**{I3,I5} 1**

**F2:**

**{I1,I2} 4**

**{I1,I3} 4**

**{I1,I5} 2**

**{I2,I3} 4**

**{I2,I4} 2**

**{I2,I5} 2**

**3rd scan:C3->**

**{I1,I2,I3}**

**{I1,I2,I4}**

**{I1,I2,I4}**

**{I1,I2,I5}**

**{I2,I3,I5}**

**F3:**

**Itemset sup**

**{I1,I2,I3} 2**

**{I1.I2,I5} 2**

**频繁模式挖掘的挑战**

**Challenges：**

**Multiple scans of transaction database**

**Huge number of candidates**

**Tedious workload of support counting for candidates**

**Apriori的性能提高：**

**Reduce passes of transaction database scans**

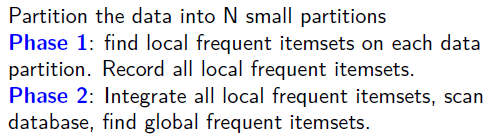
**Shrink number of candidates**

**Facilitate support counting of candidates**

**Partition:Scan Database Only Twice**

**An efficient algorithm for mining association in large databases**

**in VLDB’95**

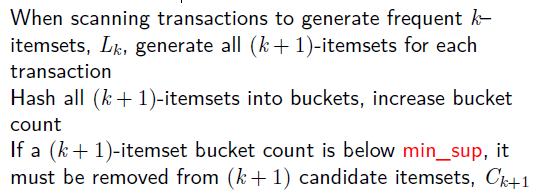


**Reducedn the Number of Candidates**

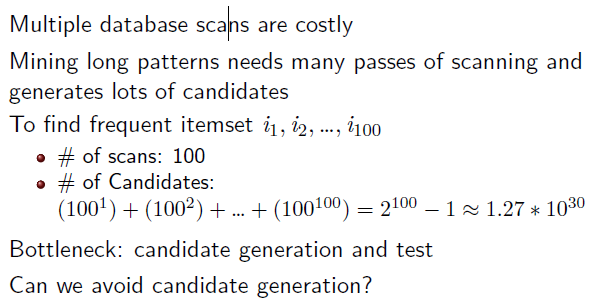
**An effective hash-based algorithm for minning association rules**

**in SIGMOD’95**

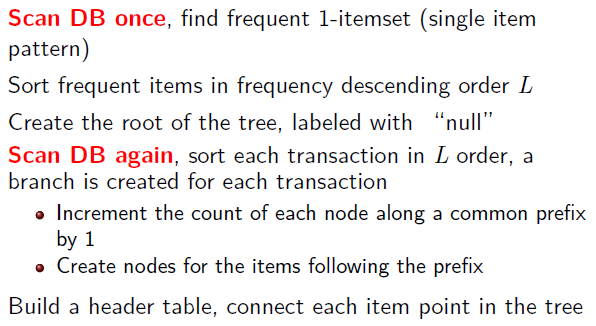
**Hah-based technique:**



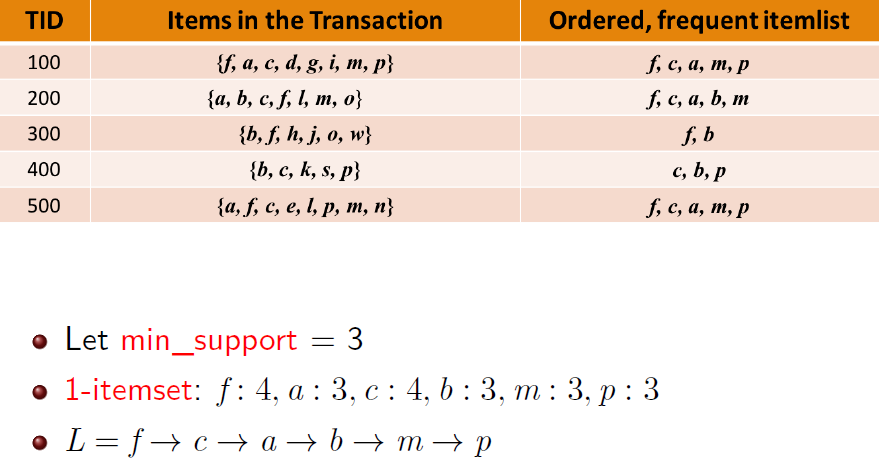
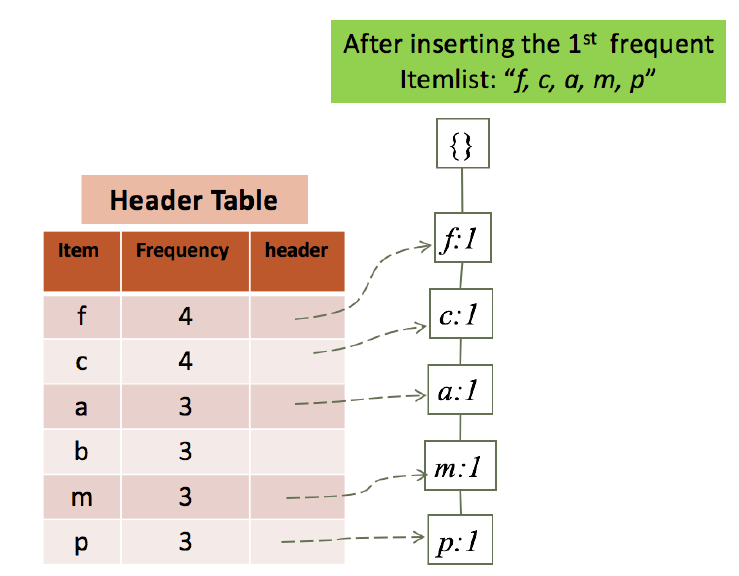
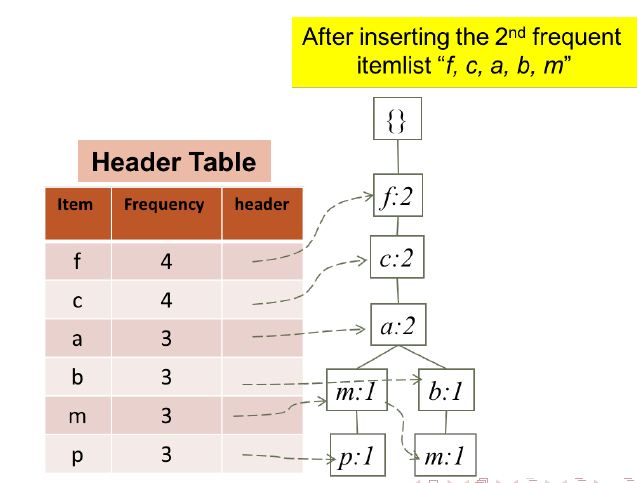
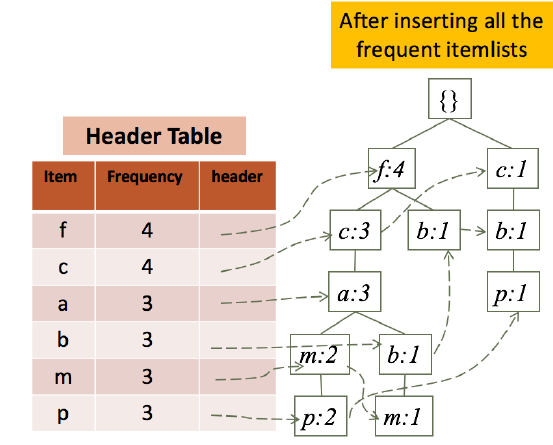
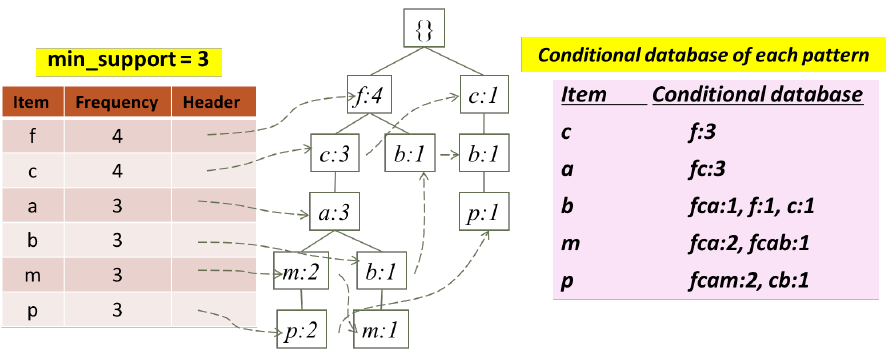
**频繁模式挖掘的瓶颈**



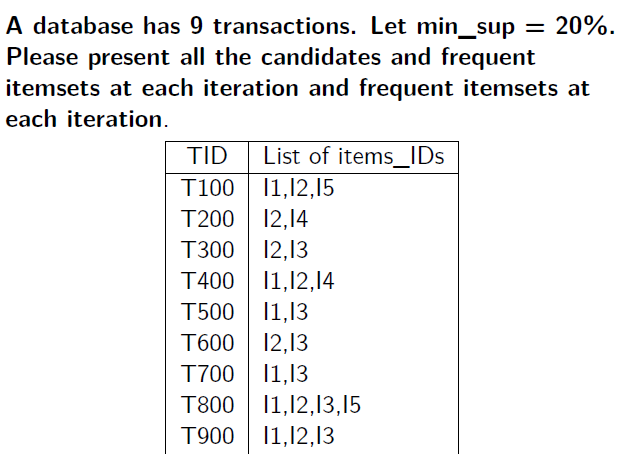
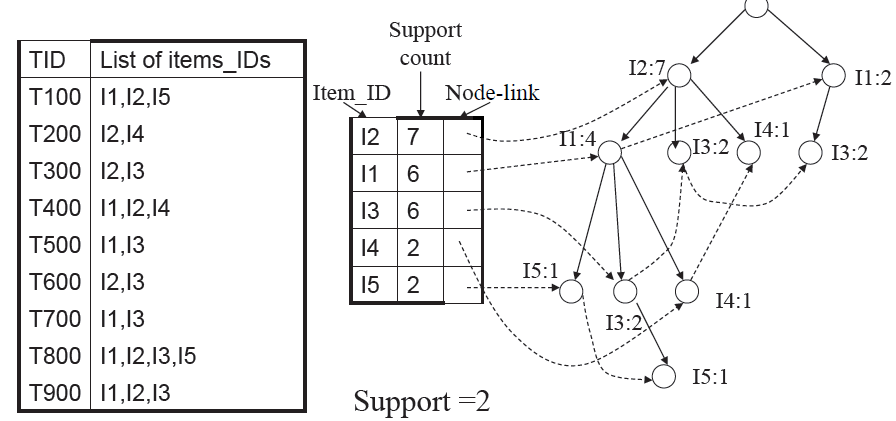
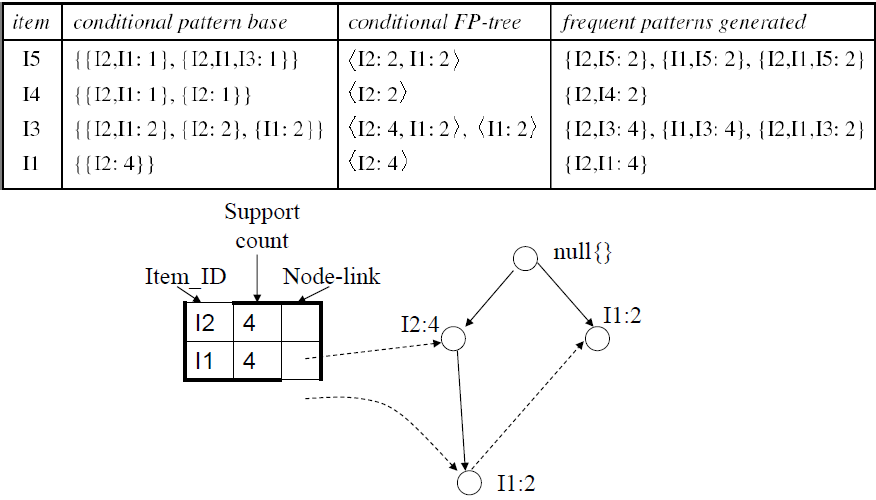
### （三）构建FP-Tree



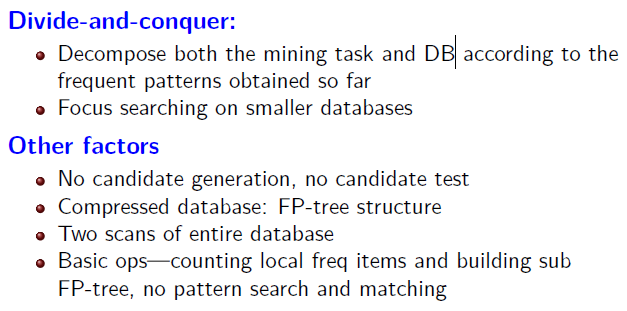
**Construct the FP-Tree**

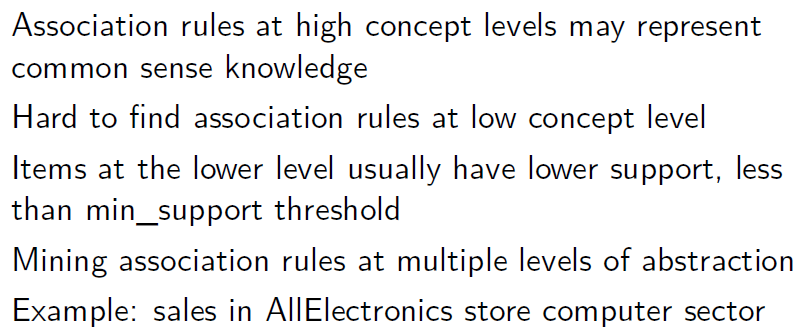
**考点三**

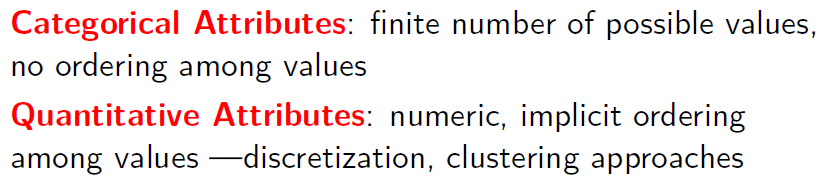
**FP优势:**



### （四）挖掘多级关联规则



### （五）挖掘多维关联规则



**挖掘数量关联**

**Static discretization**

**Based on predefined concept hierarchies.**

**Dynamic discretization**

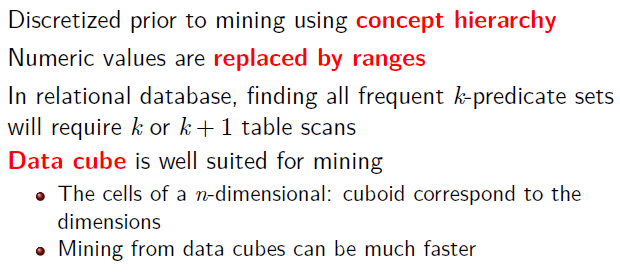
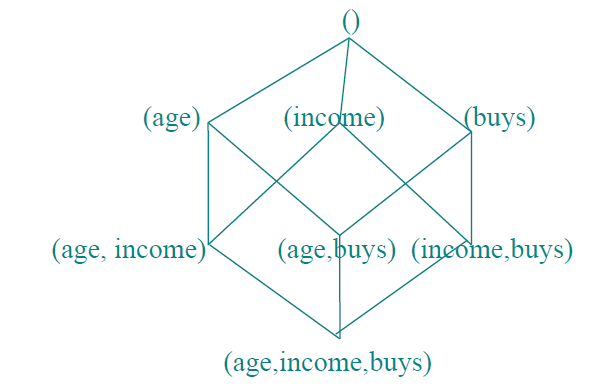
**Based on data distribution.**

**Clustering**

**Distance-based association**

**Multidimensional Association Rules and Static**

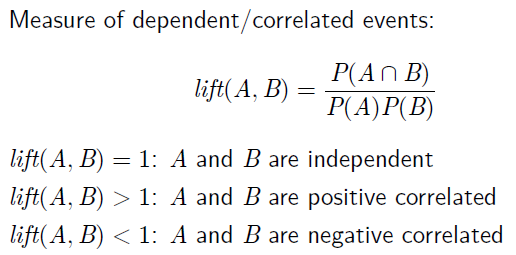
**Discretization of Quantitative Attributes**

**Correlations(Lift)**

**Lift(A,B)=P(A n B)/P(A)P(B)**

**[support,confidence,correlation]**



[**时间序列分析**](file:///F:\InstallDoc\July机器学习课程\1、_0Python数据分析与机器学习实战\资料\唐宇迪-机器学习课程\Python时间序列\时间序列分析.pdf)

**序列模式挖掘**

**序列模式挖掘：给定一组序列，找到完整的频繁子序列集（即满足min\_sup阈值）。**

# 分类(决策树)&聚类

## Classification

**决策树**

**随机森林**

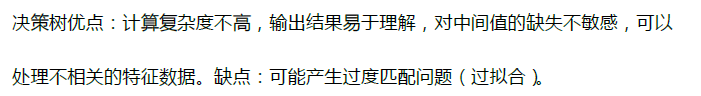
**Adaboost**

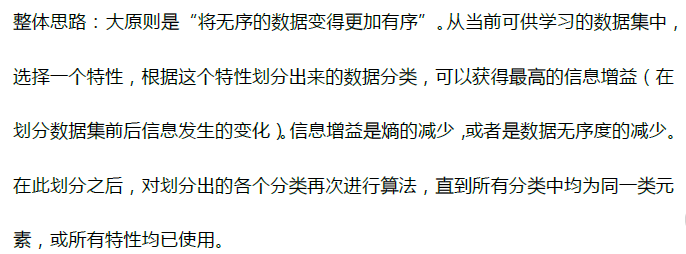
### 决策树(Decision Tree)

**决策树的特性**

**决策树是一种简单但是广泛使用的分类器，通过训练数据构建决策树，可以高校地对未知数据进行分类。决策树有两大优点：**

1. **决策树模型可读性好，具有描述性，有助于人工分析；**
2. **效率高，决策树只需要一次性构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。**



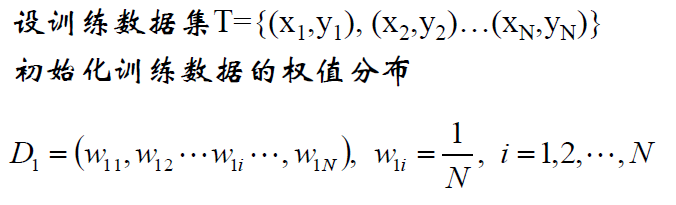


### 随机森林(Random Forest)

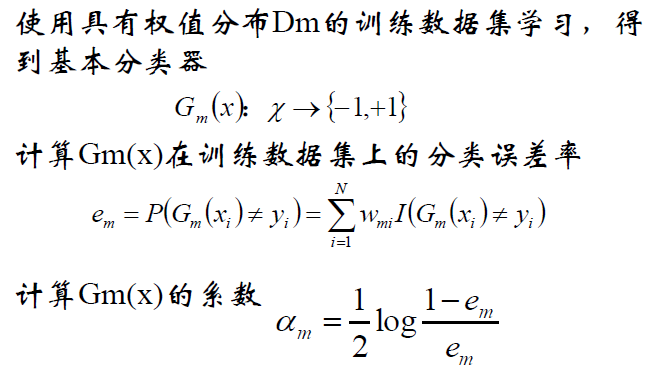
定义：

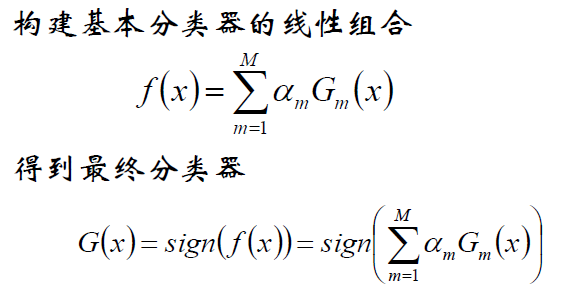
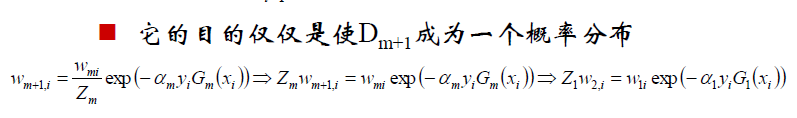
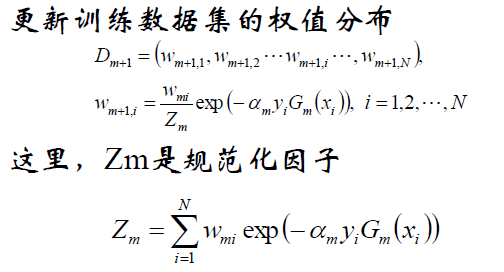
随机森林指的是利用多棵树对样本进行训练并预测的一种分类器。它通过对数据集中的子样本进行训练，从而得到多棵决策树，以提高预测的准确性并控制在单棵决策树中极易出现的过拟合情况。

### Adaboost



**对于m=1,2,…M**





## Cluster

**k-Means**

**层次聚类**

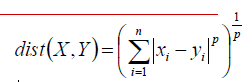
**高斯混合模型**

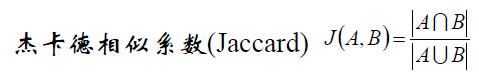
### （一）聚类概念

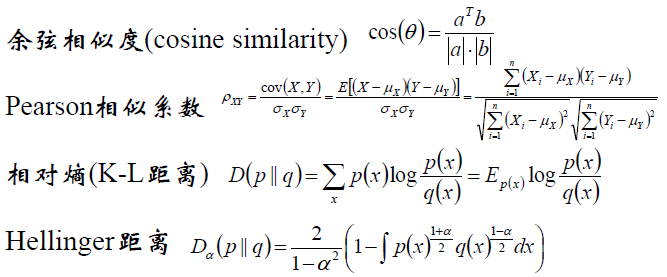
对大量未知标注的数据集，按照数据的内在相似性将数据集划分为多个类别，使得类别内的数据相似度较大而类别间的相似度较小，属于无监督学习的范畴。

### （二）**相似度/距离计算方法**

欧氏距离：







### （三）**聚类的基本思想**

给定一个有N个对象的数据集，构造数据的k个簇，k<=n。满足下列条件：

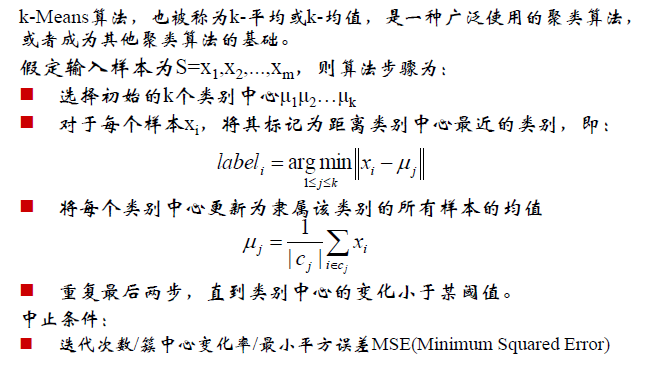
每一个簇至少包含一个对象；

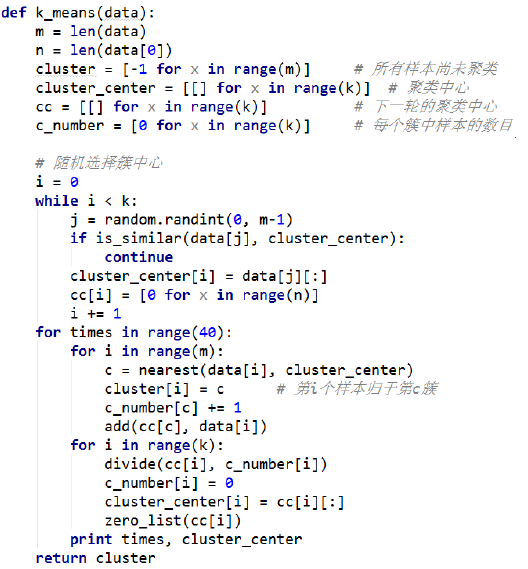
每一个对象属于且仅属于一个簇；

将满足上述条件的k个簇称作一个合理划分。

基本思想；对于给定类别的数目k，首先给出初始划分，通过迭代改变样本和簇的隶属关系，使得每一次改变之后的划分方案都较前一次好。

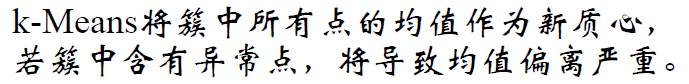
### （四）k-Means算法





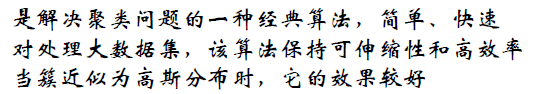
### （五）k-Means存在的隐患

**k-Means是对初值敏感的！！!**

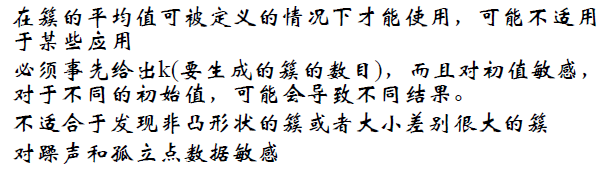


### （六）k-Means聚类算法总结

**优点：**



**缺点：**

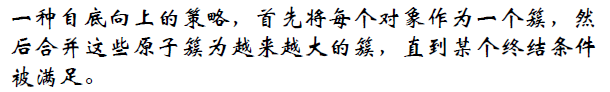


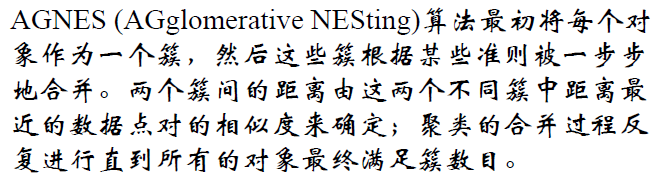
**可作为其他聚类算法的基础：谱聚类**

### 层次聚类方法

定义：层次聚类方法对给定的数据集进行层次的分解，知道某种条件满足为止。具体可分为：

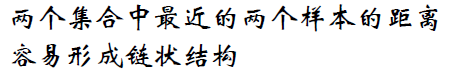
**<I>凝聚的层次聚类:AGNES算法**





**AGNES簇间距离的不同定义**

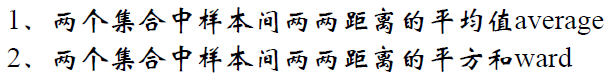
**a.最小距离：**



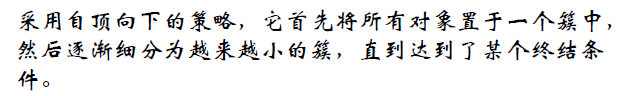
**b.最大距离：**

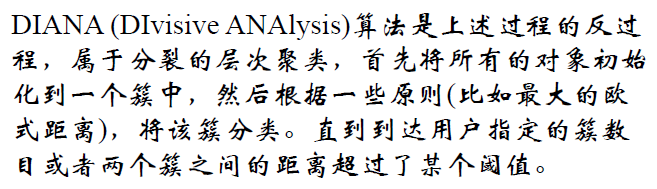


**c.平均距离：**

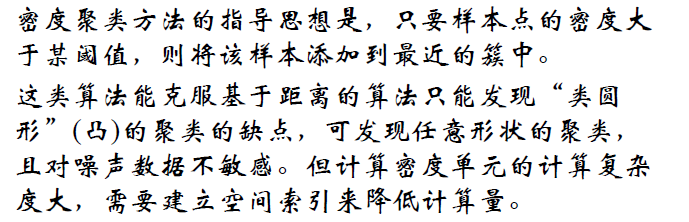


**(II)分裂的层次聚类：DIANA算法**

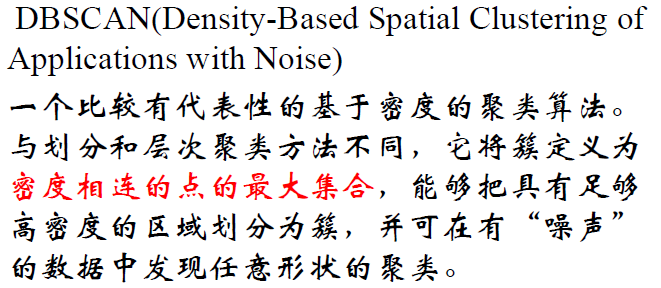


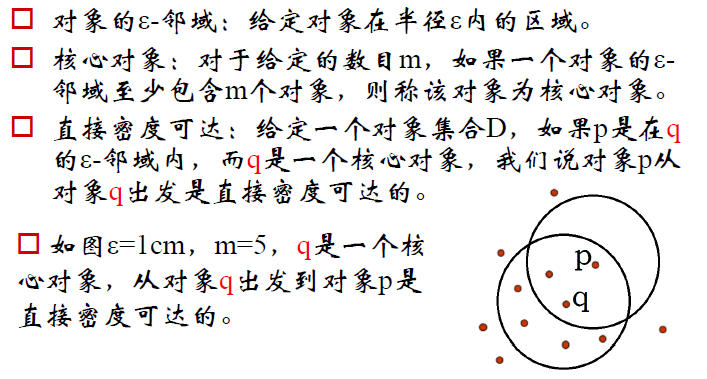


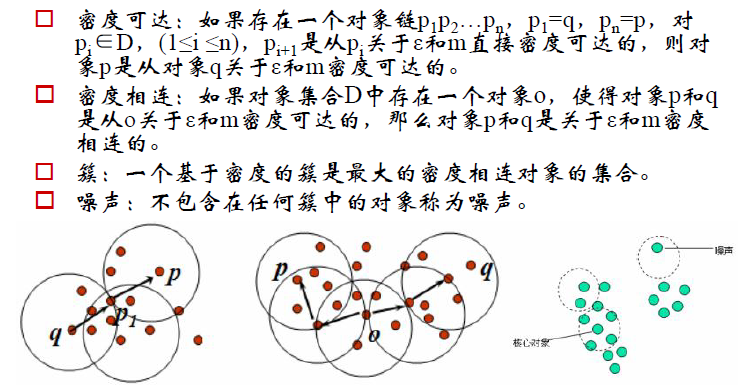
### 密度聚类



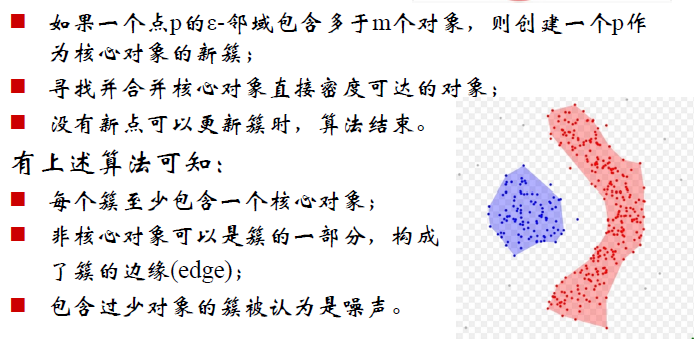
**DBSCAN**



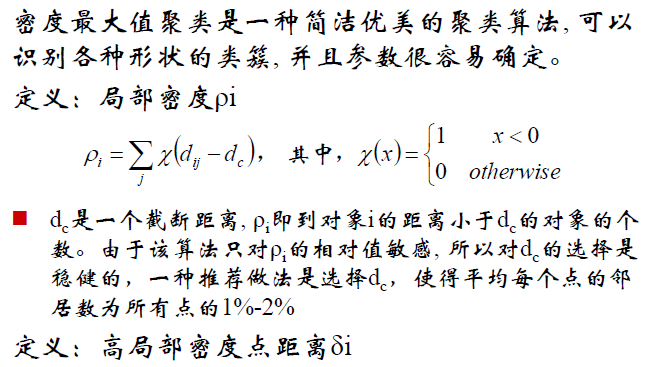




**DBSCAN算法流程：**



**DensityPeak密度最大值聚类**



### 谱聚类

考虑谱聚类和PCA关系

### 半监督LPA算法