



# 数据挖掘

## Data Mining

主讲: 张仲楠 教授



廈門大學  
XIAMEN UNIVERSITY



# 数据基础



01

**数据类型**

---

02

**数据质量**

---

03

**数据预处理**

---

04

**相似性和相异性的度量**

---

# 1. 数据类型

## 1. 属性

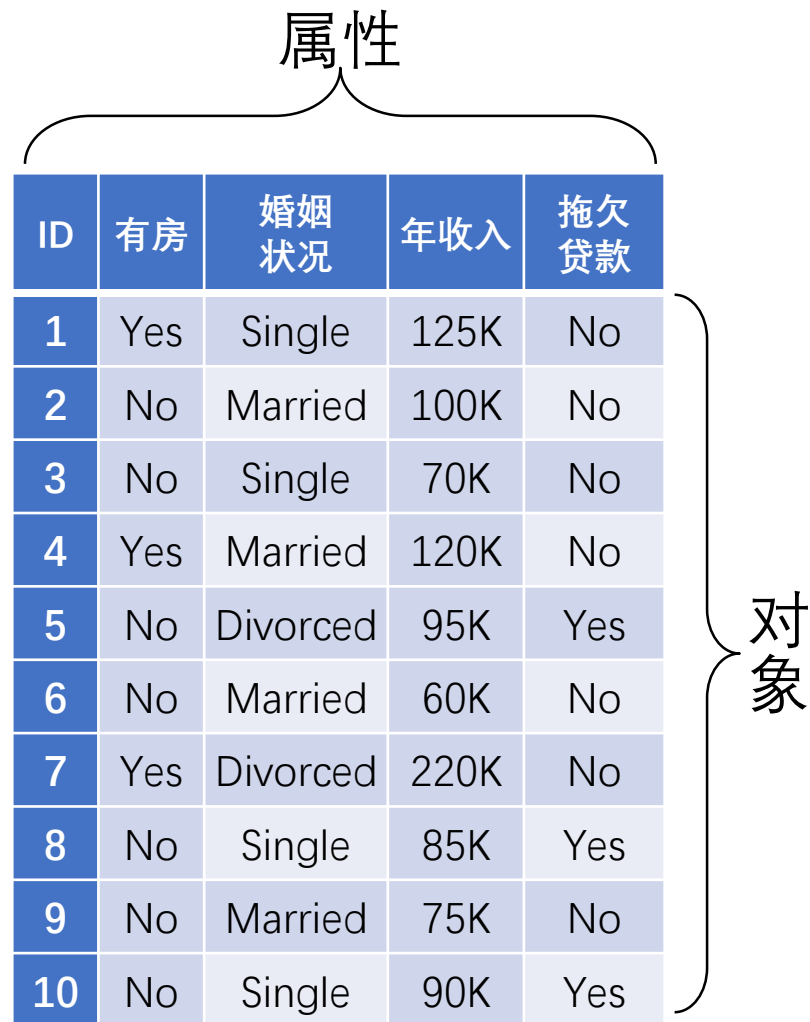
■ 属性：对象的性质或特性，因对象而异，或随时间而变化

■ 也称为变量、特性、字段、特征或维

■ 一组属性用于描述对象

■ 对象也称为记录、点、事例、示例、实体或实例

■ 数据集：数据对象的集合



ID	有房	婚姻状况	年收入	拖欠贷款
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# 1. 数据类型

## 2. 属性值

■ 属性值是指定给特定对象属性的数值或符号

■ 数值：可以具有无穷多个值，比如：距离

■ 符号：具有少量的值，比如：{红, 黄, 蓝}

■ 属性和属性值之间的区别

■ 相同的属性可以映射到不同的属性值，比如性别用{M,F}或{0,1}

■ 不同的属性可以映射到同一组值，比如属性ID和年龄都可以用整数

■ 属性的性质可以不同于用于表示属性值的性质

■ 年龄属性有最大值，而整数没有

# 1. 数据类型

## 3. 属性类型

■ 可以定义四种属性类型：标称、序数、区间、比率

- 标称(Nominal): 标称属性的值是一些符号或实物的名称，每个值代表某种类别、编码或状态。比如：ID，性别。
- 序数(Ordinal): 序数属性可能的取值之间具有有意义的序，但相继值之间的差是未知的。比如：排名，等级。
- 区间(Interval): 区间属性可能的取值之间是有序的，不同取值之间的差异是有意义的。比如：如日期、温度。
- 比率(Ratio): 具有相等单位绝对零点的变量，例如身高、体重等。

# 1. 数据类型

## 3. 属性类型

■ 属性的类型取决于它拥有以下哪些性质/操作

■ 相异性:  $=$ 和 $\neq$

■ 序:  $<$ 、 $\leq$ 、 $>$ 、 $\geq$

■ 差异:  $+$ 和 $-$

■ 比率:  $*$ 和 $/$

■ 标称: 相异性

■ 序数: 相异性、序

■ 区间: 相异性、序、差异

■ 比率: 全具备

# 1. 数据类型

	属性类型	描述	例子	操作
分类的 (定性的)	标称	标称属性值仅用于区分不同对象. (=, ≠)	邮政编码, 雇员编号, 肤色, 性别: {男, 女}	众数、熵、列联相关, $\chi^2$ 检验
	序数	序数属性的值可用于对象的排序. (<, >)	等级 (优、良、中、及格、不及格)、年级 (一年级、二年级…)、职称	中值、百分位、秩相关、游程检验、符号检验
数值的 (定量的)	区间	对于区间属性, 值之间的差异是有意义的 (+, -)	日期, 温度	均值、标准差、皮尔逊相关系数、t 和 F 检验
	比率	对于比率属性, 差和比率都是有意义的. (*, /)	年龄、质量、长度、速度	几何平均、调和平均、百分比变化

## 属性分类



# 1. 数据类型

## 4. 用值的个数描述属性

- 离散：具有有限个值或无限可数个值，常用**整数变量**表示
  - 可以是分类的(如邮政编码或ID号)，也可以是数值的(如计数)
  - **二元**属性是离散属性的**特例**，常用**布尔变量**表示
- 连续：取**实数值**的属性，常用**浮点变量**表示
  - 例如：温度、高度、重量
- 通常，标称和序数属性是二元的或离散的，区间和比率属性是连续的

# 1. 数据类型

## 5. 非对称属性

■ 只有存在**非零属性值**才被视为重要

- 文档中出现的单词

- 客户购买过的商品

■ 非对称的二元属性：只有非零值才重要的二元属性

- 对于关联分析特别重要

■ 也存在离散的/连续的非对称属性

- 比如：文档中单词出现的次数/频率

TID	Beer	Bread	Coke	Diaper	Egg	Milk
1	0	1	1	0	0	1
2	1	1	0	0	0	0
3	1	0	1	1	0	1
4	1	1	0	1	0	1
5	0	0	1	1	0	1

非对称的二元属性

# 1. 数据类型

TID	Beer	Bread	Coke	Diaper	Egg	Milk
1	0	1	1	0	0	1
2	1	1	0	0	0	0
3	1	0	1	1	0	1
4	1	1	0	1	0	1
5	0	0	1	1	0	1

稀疏数据

## 6. 数据集的重要特性

■ 维度：数据对象具有的**属性数**

■ 分析高维数据有时会陷入**维灾难**，主要通过**降维**解决

■ 稀疏性：稀疏的数据对象**大多数属性值都是0**

■ 对于关联分析特别重要

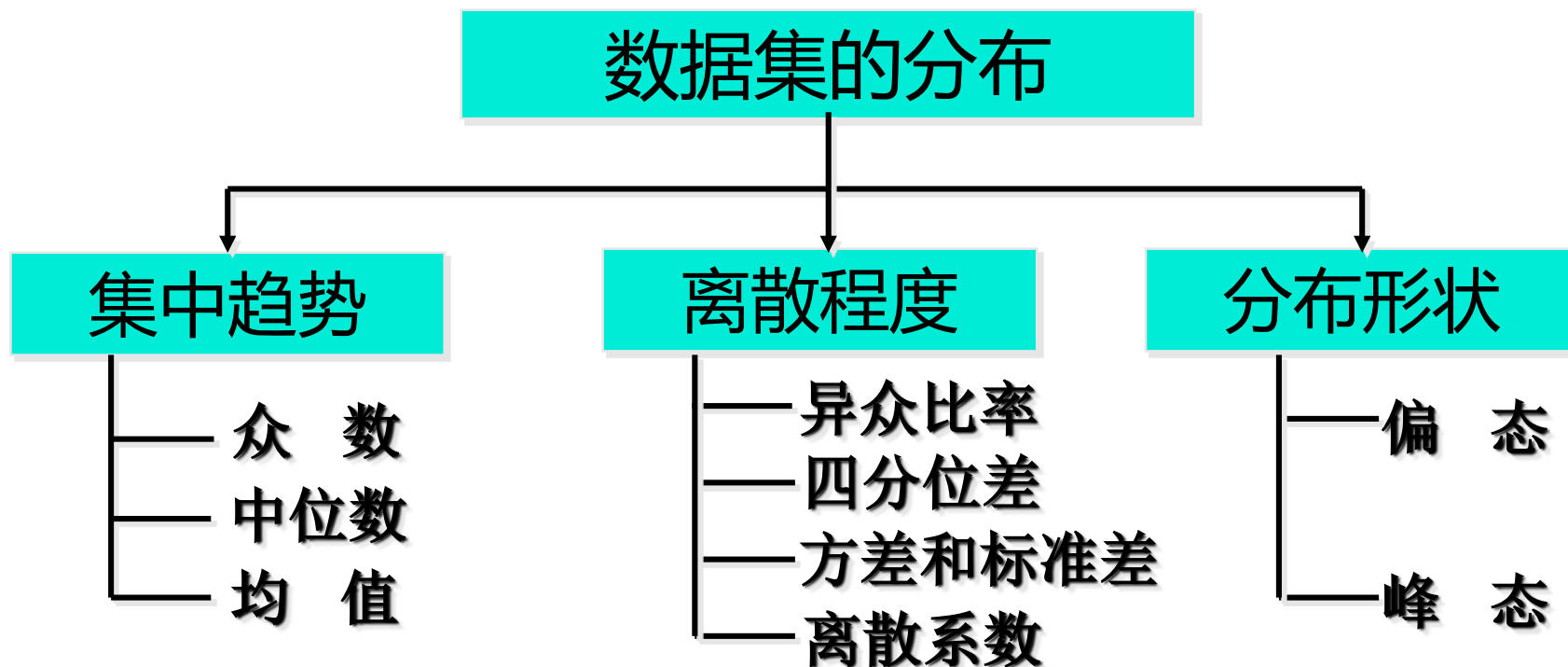
■ 分布：构成数据对象的属性的各种**值或值的集合**出现的**频率**

■ 很多数据的分布**并非标准的统计分布**，比如高斯分布

# 1. 数据类型

## 6. 数据集的重要特性

■ 数据集的分布特征可以从**集中趋势**、**离散程度**及**分布形态**三个方面进行描述



# 1. 数据类型

## 6. 数据集的重要特性

- 集中趋势的主要测度包括三种：众数、中位数、均值
- 众数：在一组数据中，出现次数最多的数据值
  - 不受极端值的影响
  - 一组数据可能没有，也可能有好几个

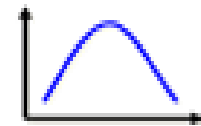
■ 无众数

原始数据: 10 5 9 12 6 8



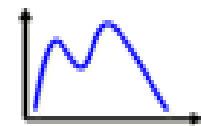
一个众数

原始数据: 6 7 9 8 7 7



多于一个众数

原始数据: 25 28 28 42 36 36



# 1. 数据类型

## 6. 数据集的重要特性

- 集中趋势的主要测度包括三种：众数、中位数、均值
- 中位数：将一组数据按大小依次排列，把处在最中间位置的一个数据(或两个数据的平均数)
  - 不受极端值影响
  - 各变量与中位数的差绝对值之和最小
  - 求解公式：  $\min(\sum_{i=1}^n |x_i - M_e|)$

1, 3, 3, 6, 7, 8, 9  
Median = **6**

1, 2, 3, 4, 5, 6, 8, 9  
Median =  $(4+5) \div 2 = \mathbf{4.5}$

# 1. 数据类型

## 6. 数据集的重要特性

■ 集中趋势的主要测度包括三种：众数、中位数、均值

■ 均值：即一组数据的均衡点所在，是集中趋势的最常用测度值

■ 易受极端值影响

■ 两点性质：

■ 1) 各变量值与均值的离差之和等于零

■ 2) 各变量值与均值的离差平方和最小

■ 几何平均值：均值的另一种表现形式

■ 是n个变量值乘积的 n 次方根，适用于对**比率数据的平均**，主要用于计算**平均增长率**

■ 计算公式： $G_m = \sqrt[n]{x_1 \times x_2 \times x_3 \dots \times x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$

简单均值：
$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

加权均值：
$$\bar{x} = \frac{M_1 f_1 + M_2 f_2 + \dots + M_k f_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^k M_i f_i}{n}$$

# 1. 数据类型

## 6. 数据集的重要特性

■ 离散程度：反映观测变量各个取值之间的**差异**，包括：异众比率、四分位差、方差及标准差、标准分数、离散系数等

■ 异众比率：对**分类数据**离散程度的测度，计算公式  $\frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$

众数组的频数

变量值的  
总频数

■ 四分位数：**顺序数据**离散程度的测度，把数据分布划分成**4个相等的部分**，使得每部分表示数据分布的四分之一，这**三个数据点**就称为四分位数，下四分位数Q1、中位数Q2、上四分位数Q3。

■ 四分位差：Q=Q3-Q1，其中：Q1的位置=(n+1)/4，Q3的位置=3(n+1)/4

■ 问题：由8人组成的旅游小团队年龄分别为：17、19、22、24、25、28、34、38，求其年龄的四分位差。（12.75）



# 1. 数据类型

## 6. 数据集的重要特性

■ 离散程度：反映观测变量各个取值之间的**差异**，包括：异众比率、四分位差、方差及标准差、标准分数、离散系数等

■ 方差和标准差：数据离散程度的**最常用测度值**；反映了各变量值**与均值的平均差异**

$$s^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

■ 标准分数：也称标准化值，是对某一个值在一组数据中相对位置的度量，计算公式为

$$z_i = \frac{x_i - \bar{x}}{s}$$

■ 离散系数：标准差与其相应的均值之比，**消除了数据水平高低和计量单位的影响**，其

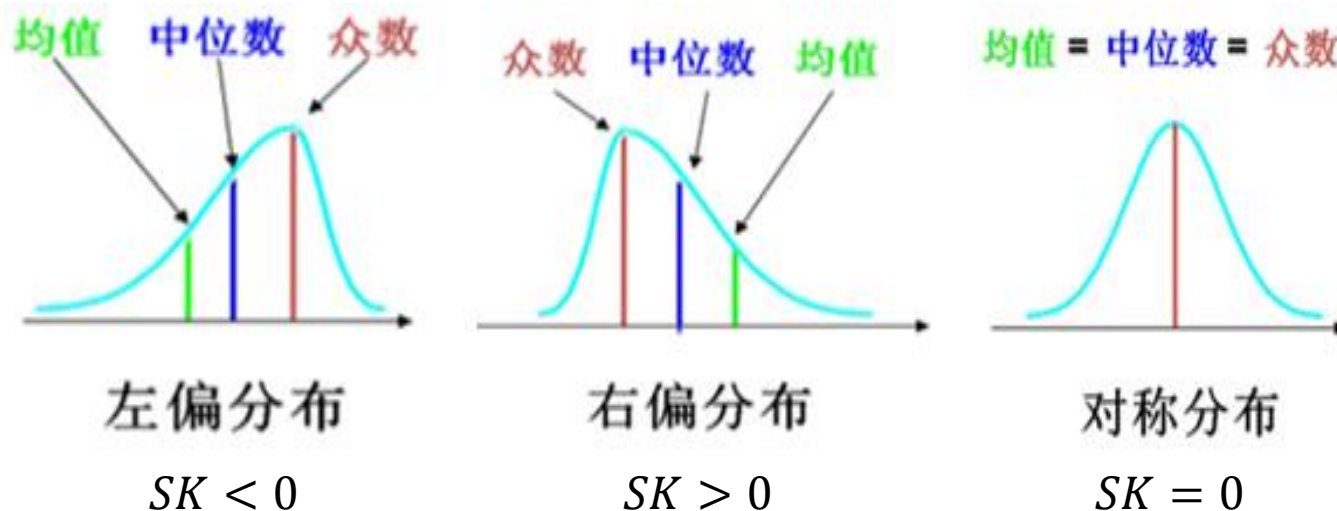
计算公式为  $v_s = \frac{s}{\bar{x}}$

# 1. 数据类型

## 6. 数据集的重要特性

- 数据的分布形状通过偏态和峰态衡量
- 偏态：倾斜度(skewness)，统计数据分布偏斜方向和程度的度量
- 根据原始数据计算偏态系数
- 数据偏态分布

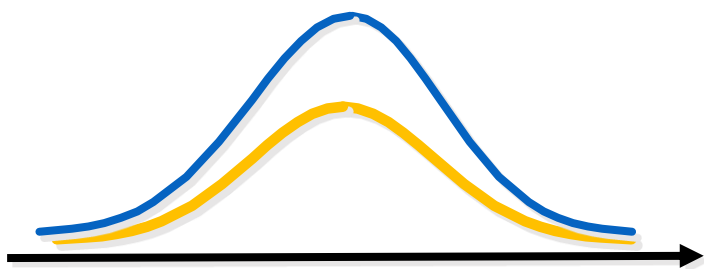
$$SK = \frac{n \sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$



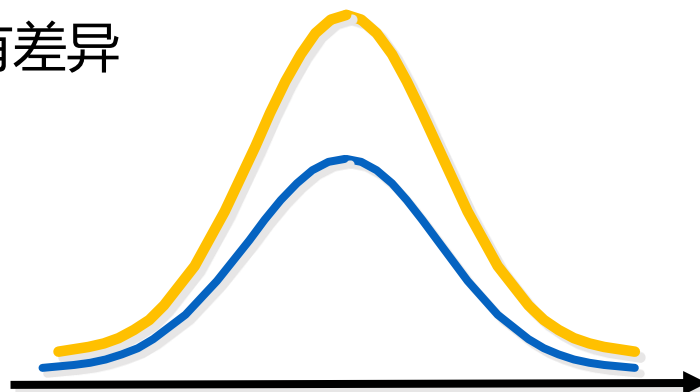
# 1. 数据类型

## 6. 数据集的重要特性

- 数据的分布形状通过偏态和峰态衡量
- 峰态：是反映变量分布陡峭程度的指标
- 通常分为三种情况：标准正态分布、尖峰分布和扁平分布
  - 尖峰分布：数据集中在一个或几个点附近，而其数据则相对较少
  - 扁平分布：样本之间的差异很小，或者没有差异



扁平分布



尖峰分布

# 1. 数据类型

## 6. 数据集的重要特性

- 分辨率(resolution): 经常可以在不同的分辨率下得到数据, 并且在不同的分辨率下数据的性质不同。
  - 例如, 在几米的分辨率下, 地球表面看上去很不平坦, 但在数十公里的分辨率下却相对平坦。
- 数据的模式也依赖于分辨率。如果分辨率太高, 模式可能看不出, 或者掩埋在噪声中; 如果分辨率太低, 模式可能不出现。
  - 例如, 几小时记录一下气压变化可以反映出风暴等天气系统的移动;而在月的标度下, 这些现象就检测不到。

# 1. 数据类型

## 7. 数据集的类型

■ 数据集类型基本可以分成三组：

- 记录数据

- 基于图的数据

- 有序数据

■ 记录数据集：记录的汇集，每个记录具有**固定的、相同的属性集**，记录之间或属性之间没有明显的联系

ID	有房	婚姻状况	年收入	拖欠贷款
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# 1. 数据类型

## 7. 数据集的类型

■ 事务数据：一种特殊类型的记录数据

■ 每个记录涉及一系列项

■ 顾客一次购物所购买的商品的集合就是一个记录，购买的商品就是项

■ 事务数据是项的集合，可以将其视为一组字段为非对称属性的记录

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

TID	Beer	Bread	Coke	Diaper	Egg	Milk
1	0	1	1	0	0	1
2	1	1	0	0	0	0
3	1	0	1	1	0	1
4	1	1	0	1	0	1
5	0	0	1	1	0	1

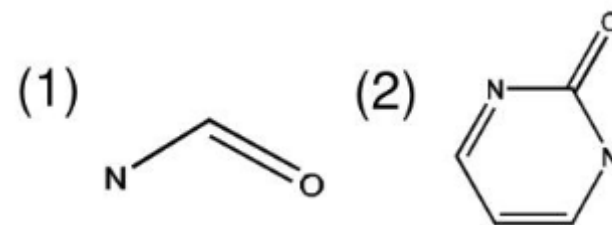
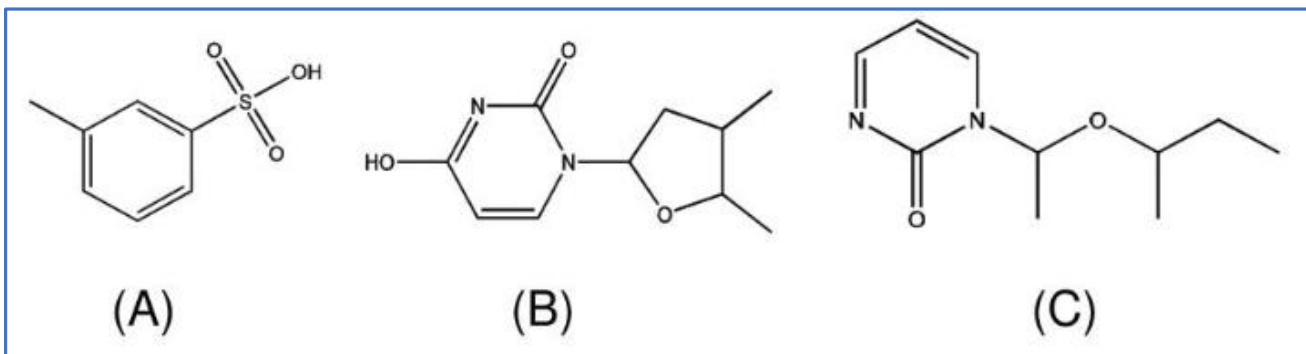
# 1. 数据类型

## 7. 数据集的类型

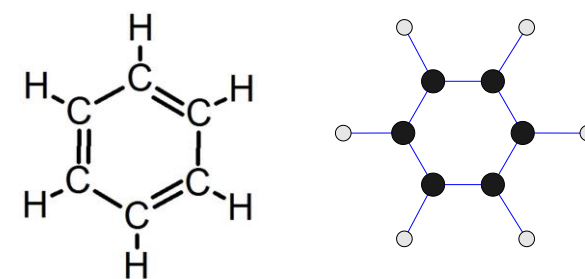
### ■ 基于图的数据

- 图捕获数据对象之间的联系
- 数据对象本身用图表示

### ■ 频繁图挖掘：在图的集合中发现一组频繁出现的子结构



最小支持度=2



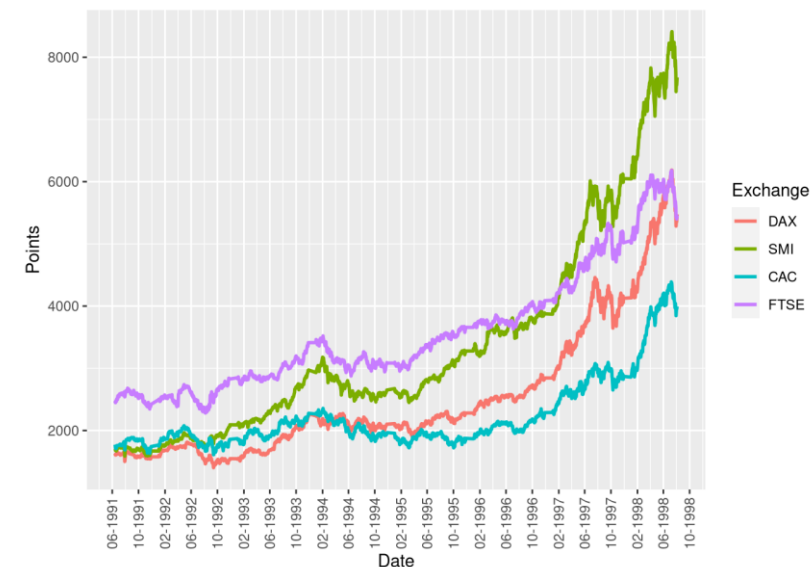
苯 (Benzene)

# 1. 数据类型

## 7. 数据集的类型

- 时序事务数据：事务数据的扩充，每个事务包含一个与之相关联的时间
- 时间序列数据：特殊的有序数据类型，每条记录都是一个时间序列
  - 重要考虑时间自相关性，即如果两个测量的时间很近，则它们的值很相似

时间	顾客	购买的商品
t1	C1	A,B
t2	C3	A,C
t2	C1	C,D
t3	C2	A,D
t4	C2	E
t5	C1	A,E



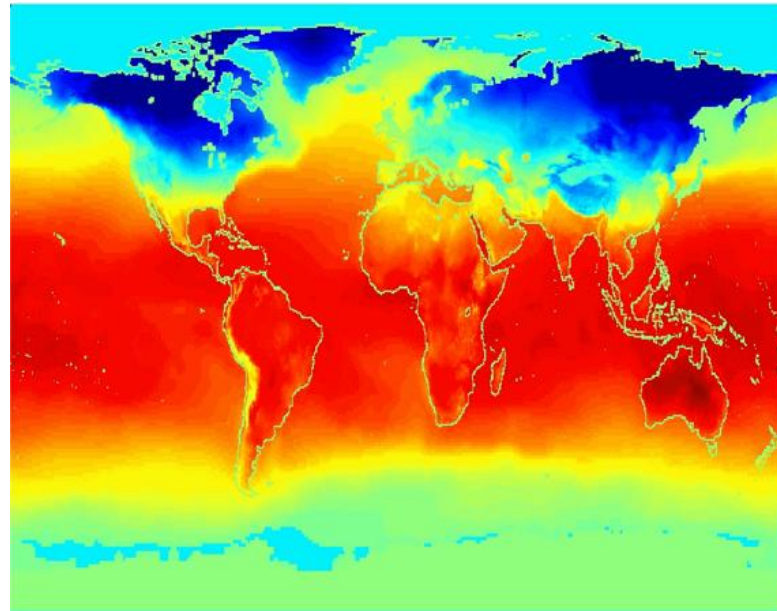


# 1. 数据类型

## 7. 数据集的类型

- 序列数据：一个数据集，包含各个实体的序列，如词或字母的序列
- 空间和时空数据：具有空间属性的数据对象
  - 重要考虑空间自相关性，即物理上靠近的对象趋于在其他方面也相似
  - 空间自相关类似于时间自相关

```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCCGCCCCGCGCCGTC  
GAGAAGGGCCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```





01

**数据类型**

---

02

**数据质量**

---

03

**数据预处理**

---

04

**相似性和相异性的度量**

---

## 2. 数据质量

### 1. 数据质量

- 人类的**错误**、测量设备的**限制**或数据收集过程中的**漏洞**都可能导致**不真实或重复的对象**，也会存在**不一致**
- 数据挖掘着眼于两个方面
  - 数据质量问题的检测与纠正 ---- **数据清理**
  - 使用可以**容忍低质量数据**的算法 ---- **鲁棒算法**

## 2. 数据质量

### 2. 测量和数据搜集问题

- 测量误差：测量过程中产生的问题 ----记录值与实际值不同
  - 误差：对于连续属性，测量值与真实值的差
- 数据搜集错误：遗漏数据对象或属性值，或者不当地包含了其他数据对象等错误
- 这两种问题可能是系统的，也可能是随机的

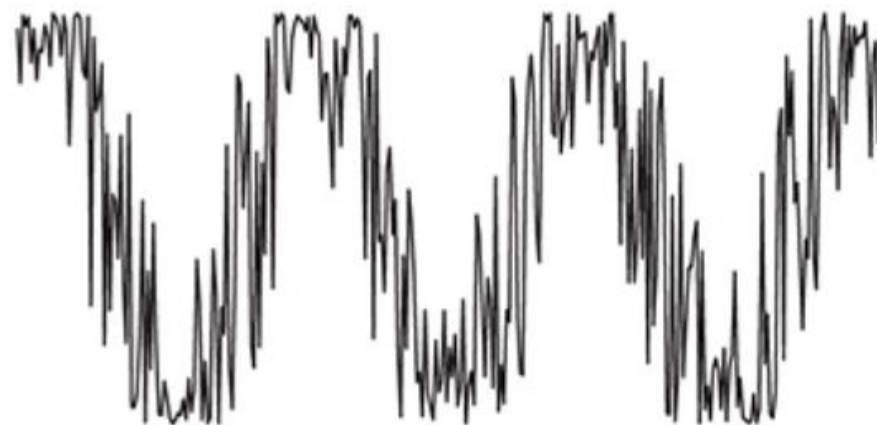
## 2. 数据质量

### 3. 噪声和伪像

- 噪声：测量误差的随机部分，通常涉及值被扭曲或加入了谬误对象
  - 既有时间上的也有空间上的



时序数据



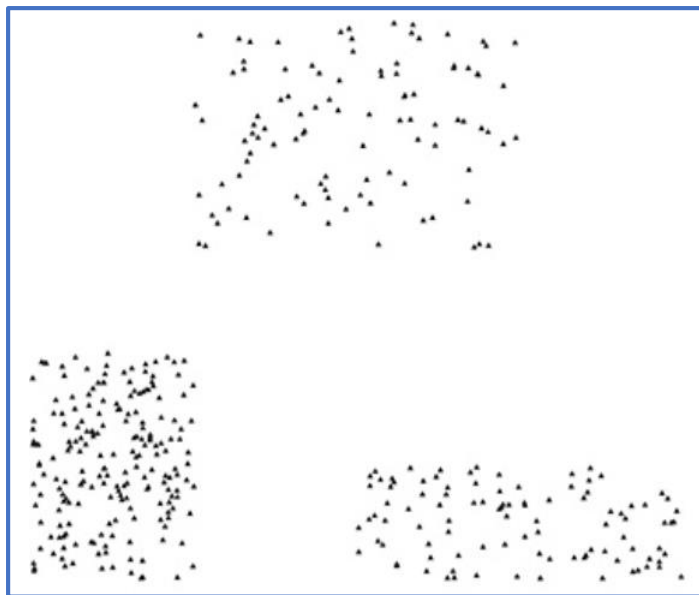
含噪声的时序数据

## 2. 数据质量

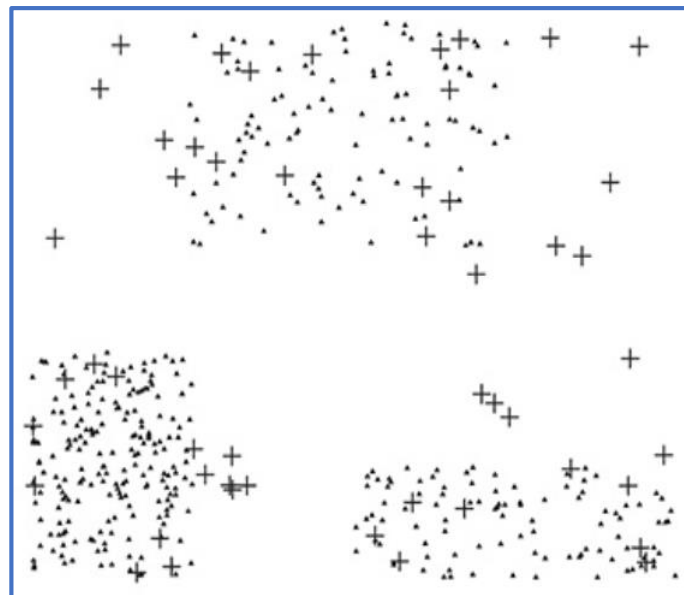
### 3. 噪声和伪像

- 使用信号或图像处理技术降低噪声，从而帮助发现可能“淹没在噪声中”的模式(信号)
- 鲁棒算法：在噪声干扰下也能产生可以接受的结果

空间数据



含噪声的空间数据



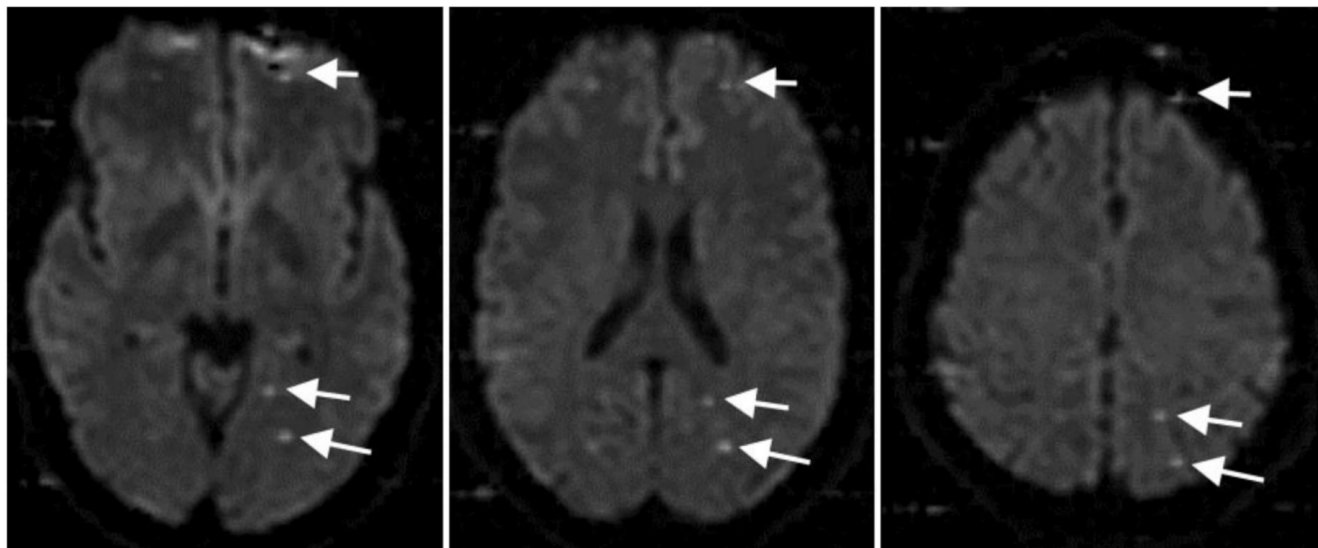
## 2. 数据质量

### 3. 噪声和伪像

■ 数据错误可能是更确定性现象的结果

■ 一组照片在同一地方出现条纹

■ 伪像(artifact): 数据的确定性失真



## 2. 数据质量

### 4. 精度、偏置

■ 测量过程和结果数据用精度和偏置度量

■ 精度(precision): (同一个量的)重复测量值之间的接近程度

■ 用值集合的标准差度量  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

■ 偏置(bias): 测量值与被测量之间的系统的变化

■ 用值集合的均值与被测量的已知值之间的差度量

$$bias^2(\mathbf{x}) = (\bar{f}(\mathbf{x}) - y)^2$$

Avg	1	2	3	4	5	精度	偏置
1.001	1.015	0.990	1.013	1.001	0.986	0.013	0.001
$x_i - \bar{x}$	0.014	-0.011	0.012	0	-0.015		



## 2. 数据质量

### 5. 准确率

- 准确率(accuracy): 被测量的测量值与实际值之间的接近度
- 一个重要方面是有效数字(significant digit)的使用
  - 目标是只使用与数据精度相符的数字来表示测量或计算结果
  - 例如, 对象的长度用最小刻度为毫米的米尺测量, 则我们只能记录最接近毫米的长度数据, 这种测量的精度为 $\pm 0.5\text{mm}$
- 缺乏对数据和结果准确率的理解, 分析者将可能出现严重的数据分析错误

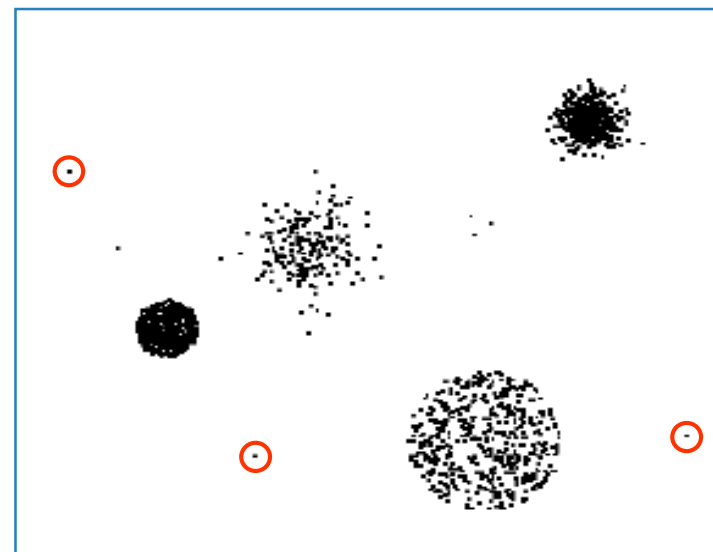
## 2. 数据质量

### 6. 离群点

■ 离群点(outlier): 具有不同于数据集中其他大部分数据对象的特征的数据对象, 或是相对于该属性的典型值来说不寻常的属性值。我们也称其为异常(anomalous)对象或异常值

■ 注意区别噪声和离群点

■ 离群点是合法的数据对象或值



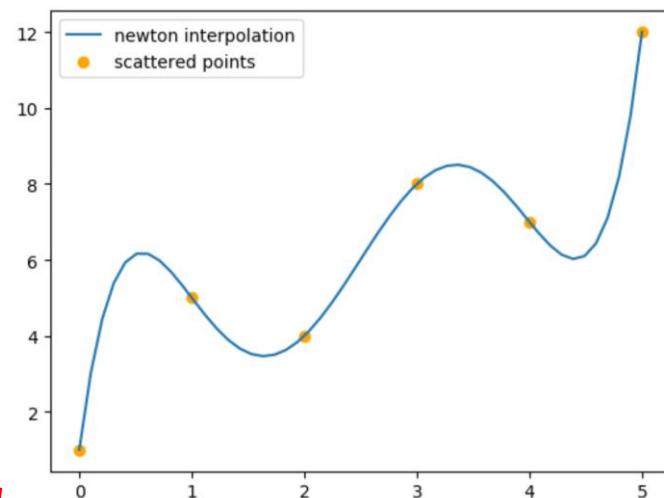
## 2. 数据质量

### 7. 遗漏值

■ 一个对象**遗漏**一个或多个属性值的情况**并不少见**

■ 应对策略:

- 删除数据对象或属性: 删除具有遗漏值的数据对象, 然而有时即便不完整的数据对象也可能包含一些有用的信息。
- 估计遗漏值: 可以采用**插值方法**、**邻近点的属性(平均)值**、**近邻中最常出现的属性值**进行可靠估计。
- 在分析时忽略遗漏值: 在聚类分析时, 需要计算两两数据对象的相似性, 可以采用一个对象或两个对象没有遗漏值的属性来计算相似性。



## 2. 数据质量

### 8. 不一致的值

- 数据可能包含不一致的值
- 无论导致不一致值的原因是什么，重要的是能检测出来并且如果可能的话，纠正这种错误，比如基于“校验”信息
- 检测到不一致后，有时可以对数据进行更正

原始码	奇校验 奇数个1	偶校验 偶数个1
1011000	10110000	10110001
1010000	10100001	10100000
0011010	00110100	00110101
0001000	00010000	00010001
0000000	00000001	00000000

## 2. 数据质量

### 9. 重复数据

- 数据集可以包含重复或几乎重复的数据对象
- 去重复：
  - 对于相同的对象，如果存在属性值不同，要一致化
  - 避免意外地将两个相似但并非重复的数据对象合并
- 在某些情况下，两个或多个对象在数据库的属性度量上是相同的，但是仍然代表不同的对象



01

**数据类型**

---

02

**数据质量**

---

03

**数据预处理**

---

04

**相似性和相异性的度量**

---

## 3. 数据预处理

### 预处理方法

■ **选择**分析所需要的数据对象和属性，以及**创建/改变**属性

- 聚集
- 抽样
- 维归约
- 特征子集选择
- 特征创建
- 离散化和二元化
- 变量变换

## 3. 数据预处理

### 1. 聚集

- 聚集(aggregation): 将两个或多个对象**合并成单个**对象
  - 删除属性的过程
  - 压缩特定属性不同值个数的过程
  - 常用于联机分析处理(Online Analytical Processing, OLAP)
- **定量**属性通常通过**求和或求平均值**进行聚集
- **定性**属性可以**忽略**也可以用更高层次的类别来**概括**



### 3. 数据预处理

事务ID	商品	商店位置	日期	价格
⋮	⋮	⋮	⋮	⋮
101123	Watch	Chicago	09/06/04	\$25.99
101123	Battery	Chicago	09/06/04	\$5.99
101124	Shoes	Minneapolis	09/06/04	\$75.00
⋮	⋮	⋮	⋮	⋮

商店位置	日期	销售额
⋮	⋮	⋮
Chicago	09/06/04	\$20000
Minneapolis	09/06/04	\$10000
⋮	⋮	⋮

删除属性

事务ID	商品	商店位置	日期	价格
⋮	⋮	⋮	⋮	⋮
101123	Watch	Chicago	09/06	\$25.99
101123	Battery	Chicago	09/06	\$5.99
101124	Shoes	Minneapolis	09/06	\$75.00
⋮	⋮	⋮	⋮	⋮

压缩属性值个数

## 3. 数据预处理

### 1. 聚集 --- 动机

- 数据归约导致的较小数据集需要较少的内存和处理时间
- 通过高层而不是低层数据视图，聚集起到了范围或标度转换的作用
- 对象或属性群的行为通常比单个对象或属性的行为更加稳定
  - 平均值、总数等聚集量具有较小的变异性
- 聚集的缺点是可能丢失有趣的细节

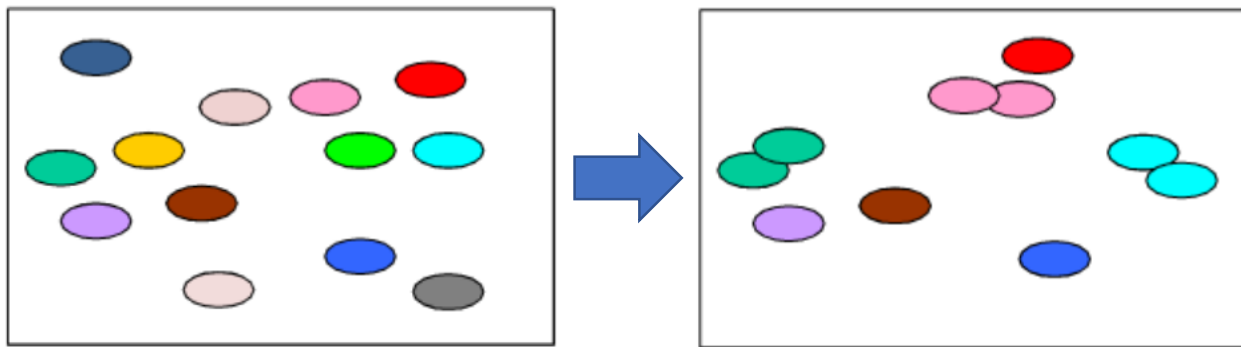
### 3. 数据预处理

#### 2. 抽样

- 抽样是一种选择数据对象子集进行分析的常用方法
- 使用抽样的算法可以压缩数据量，以便可以使用更好但开销较大的数据挖掘算法
- 若样本近似地具有与原数据集相同的(感兴趣的)性质，则称样本是有代表性的
- 如果样本是有代表性的，则使用样本与使用整个数据集的效果几乎一样

### 3. 数据预处理

## 2. 抽样



■ 选择一个**抽样方案**，以确保以很高的概率得到**有代表性的**样本

■ 简单随机抽样：选取任何特定项的**概率相等**

■ 无放回抽样——每个选中项立即从构成总体的所有对象集中删除

■ 有放回抽样——对象被选中时不从总体中删除

■ 有放回抽样(bootstrap sampling)

■ 相同的对象可能被多次抽出

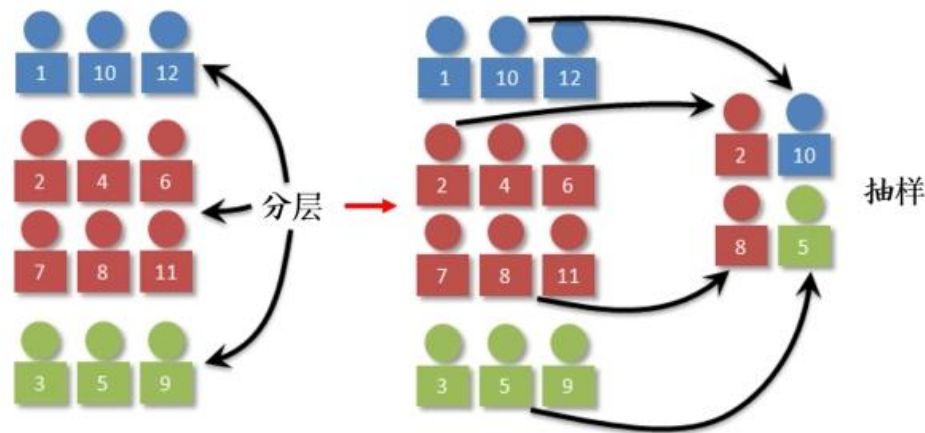
■ 每个对象被选中的概率保持不变

约有**38.6%**的样本不出现

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n \mapsto \frac{1}{e} \approx 0.368$$

### 3. 数据预处理

#### 2. 抽样



- 当总体由不同类型的对象组成并且每种类型的对象数量差别很大时，简单随机抽样不能充分地代表不太频繁出现的对象类型
- 需要提供具有不同频率的感兴趣的项的抽样方案
- 分层抽样(stratified sampling) ---- 类型抽样
  - 从每组抽取的对象个数相同 (如：从100名学生中男、女各抽取10人)
  - 从每一组对象抽取的样本数量正比于该组的大小 (如：每组各取1/3)

### 3. 数据预处理

#### 2. 抽样 --- 蓄水池采样

■ 问题：给定一串很长的数据流，对该数据流中数据只能访问一次，使得数据流中所有数据被选中的概率相等

■ 假设需要采样的数量为 $k$

- 1. 首先构建一个 $k$ 个元素的数组，将序列的前 $k$ 个元素放入数组中
- 2. 对于从第 $j$ 个元素( $j > k$ )，以 $\frac{k}{j}$ 的概率来决定该元素是否被替换到数组中，数组中的 $k$ 个元素被替换的概率是相同的
- 3. 当遍历完所有元素之后，数组中剩下的元素即为采样样本

假设  $k = 2$

数据流: a x c y z k c d e g...

当  $j=5$ , 前5个元素中的每一个被采样的概率都是  $2/5$

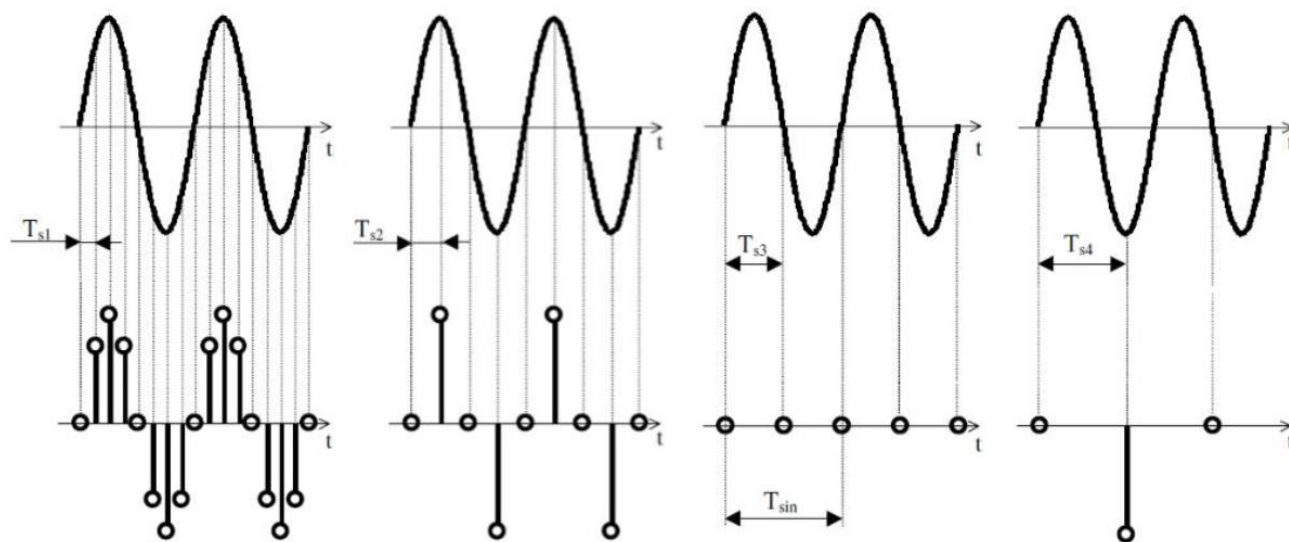
当  $j=7$ , 前7个元素中的每一个被采样的概率都是  $2/7$

### 3. 数据预处理

## 2. 抽样

### ■ 选择样本容量

- 较大容量增大了样本具有代表性的概率，但也抵消了抽样带来的许多好处
- 较小容量的样本，可能丢失模式或检测出错误的模式



### 3. 数据预处理

#### 2. 抽样

- 由于可能很难确定合适的样本容量，因此有时需要使用自适应(adaptive)或渐进抽样(progressive sampling)方法
  - 从一个小样本开始，然后增加样本容量直至得到足够容量的样本
  - 预测模型的准确率随样本容量的增加而增加，但是在某一点准确率的增加趋于稳定
  - 掌握模型准确率随样本逐渐增大的变化情况，估计出当前容量与稳定点的接近程度



## 3. 数据预处理

### 3. 维归约（降维）

- 通过创建新属性或将一些旧属性合并在一起以降低数据集的维度，使得低维表示保留原始数据的一些有意义的特性
  - 删除不相关的特征并降低噪声，许多数据挖掘算法的效果就会更好
  - 可以使模型更容易理解稳定
  - 更容易让数据可视化
  - 降低了数据挖掘算法的时间和内存需求

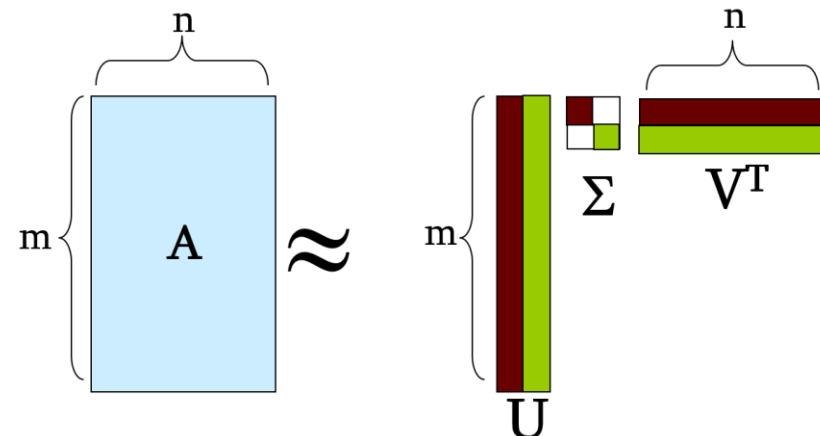
## 3. 数据预处理

### 3. 维归约 (降维)

#### ■ 维灾难

- 随着数据维度的增加，许多数据分析变得非常困难。
- 特别是随着维度增加，数据在它所占的空间中越来越稀疏
- 对于分类，这可能意味着没有足够的数据对象来创建模型，将所有可能的对象可靠地指派到一个类 --- 准确率降低
- 对于聚类，点之间的密度和距离的定义(对聚类是至关重要的)失去了意义 --- 聚类质量下降

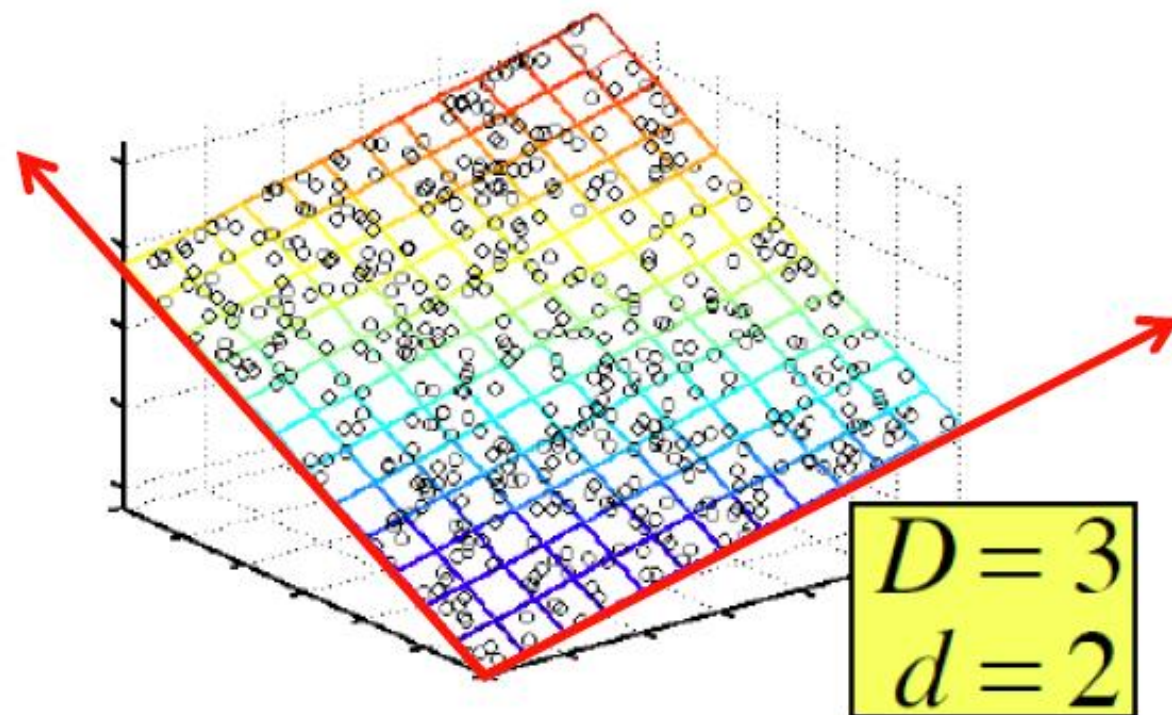
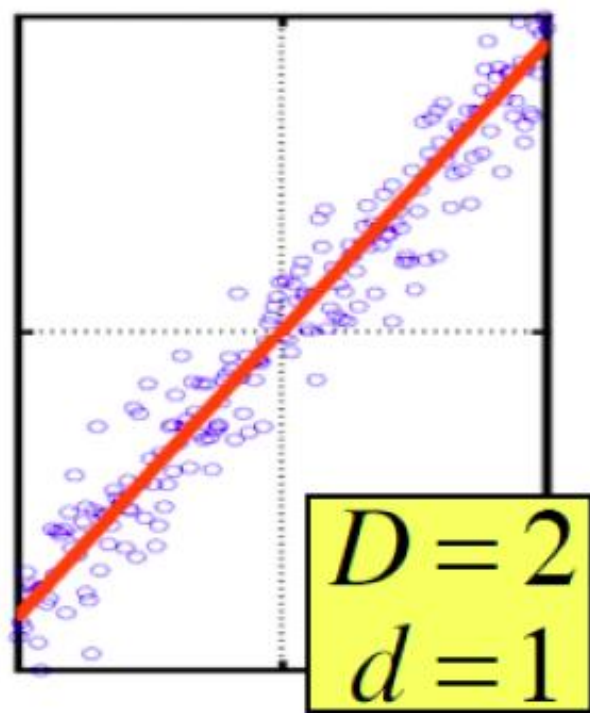
## 3. 数据预处理

$$A \approx U \Sigma V^T$$


### 3. 维归约 (降维)

- 最常用的方法是使用线性代数技术，将数据由高维空间投影到低维空间，特别是对于连续数据
- 主成分分析(Principal Component Analysis, PCA)是一种用于它找出新的属性(主成分)，这些属性是原属性的线性组合，是相互正交的(orthogonal)，并且捕获了数据的最大变差
- 奇异值分解(Singular Value Decomposition, SVD)是一种线性代数技术，它与PCA有关，并且也用于维归约

### 3. 数据预处理



降 维

## 3. 数据预处理

### 4. 特征子集选择

- **冗余特征**: 重复了包含在一或多个其他属性中的许多或所有信息
- **不相关特征**: 包含对于当前的数据挖掘任务几乎完全没用的信息
- 冗余和不相关的特征可能降低分类的准确率, 影响聚类的质量
- 理想选择方法: 将所有可能的特征子集作为感兴趣的数据挖掘算法的输入, 然后选取能产生最好结果的子集
  - 涉及 $n$ 个属性的子集多达  $2^n$  个 --- 行不通!

## 3. 数据预处理

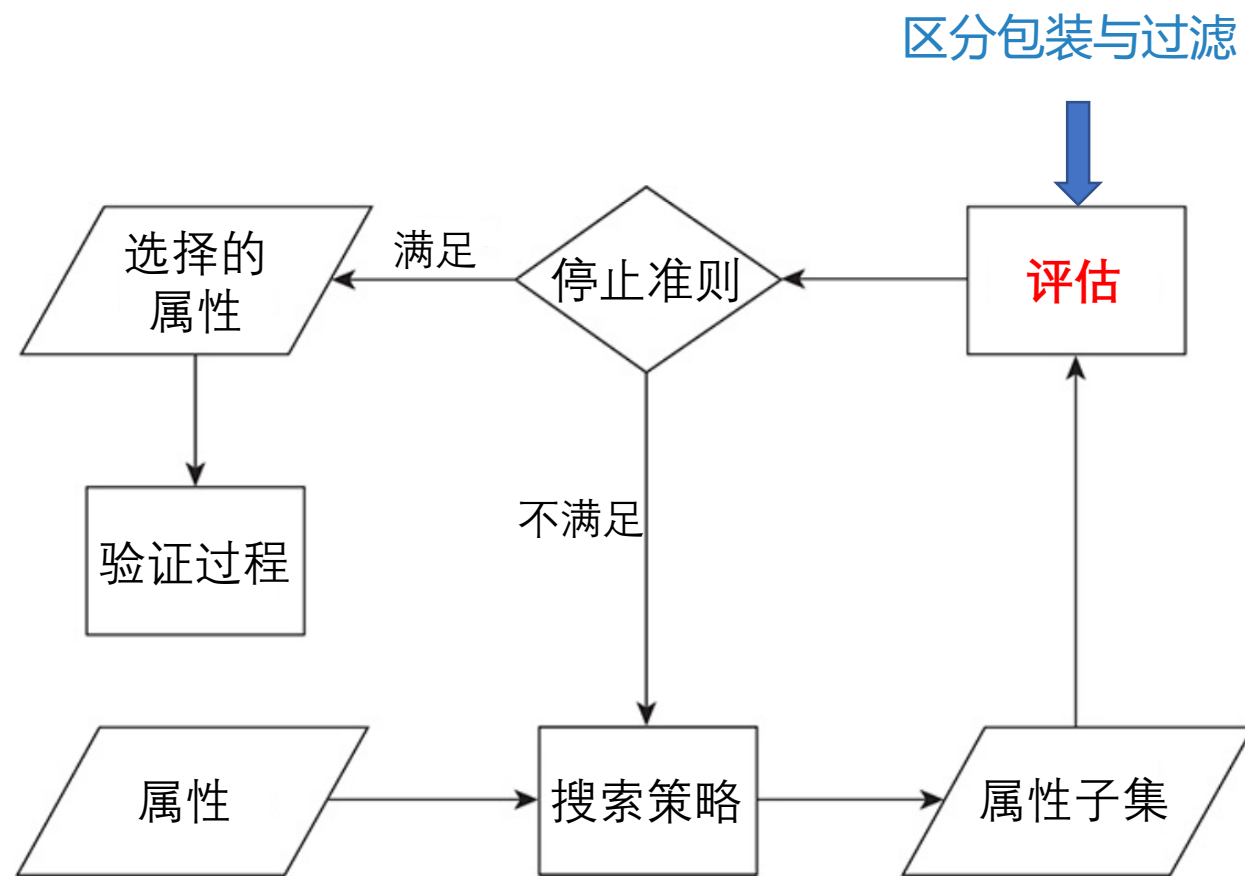
### 4. 特征子集选择

- 嵌入方法(embedded approach): 特征选择作为数据挖掘算法的一部分, 在数据挖掘算法运行期间, 算法本身决定使用哪些属性和忽略哪些属性;
- 过滤方法(filter approach): 使用某种独立于数据挖掘任务的方法, 在数据挖掘算法运行前进行特征选择
- 包装方法(wrapper approach): 直接把最终将要使用的数据挖掘算法的性能作为特征子集的评价准则, 但通常并不枚举所有可能的子集来找出最佳属性子集

### 3. 数据预处理

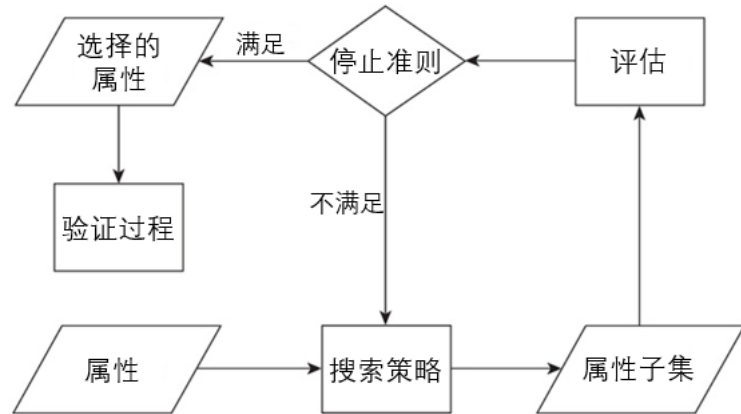
#### 4. 特征子集选择

- 可以将过滤和包装方法放到一个共同的体系结构中
- 特征选择过程可以看作由**四部分**组成：子集评估度量、控制新的特征子集产生的搜索策略、停止搜索判断和验证过程
- 包装：使用目标数据挖掘算法
- 过滤：不同于目标数据挖掘算法



特征子集选择过程流程图

### 3. 数据预处理



#### 4. 特征子集选择

- 特征子集搜索策略的**计算花费应当较低**，并且应当找到**最优或近似最优**的特征子集 --- 很难同时满足，**要折中**
- 评估步骤需要一种**评估度量**，确定属性**特征子集的质量**
- 包装方法：运行**目标数据挖掘算法**，评估函数就是通常用于**度量数据挖掘结果**的评判标准
- 过滤方法：预测**实际数据挖掘算法**在给定的属性集上执行的效果如何



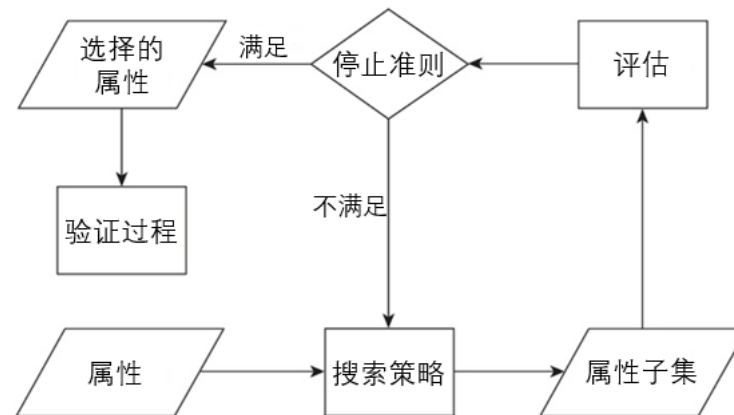
### 3. 数据预处理

#### 4. 特征子集选择

■ 因为子集的数量可能很大，考察所有的子集可能不现实，所以需要某种停止准则

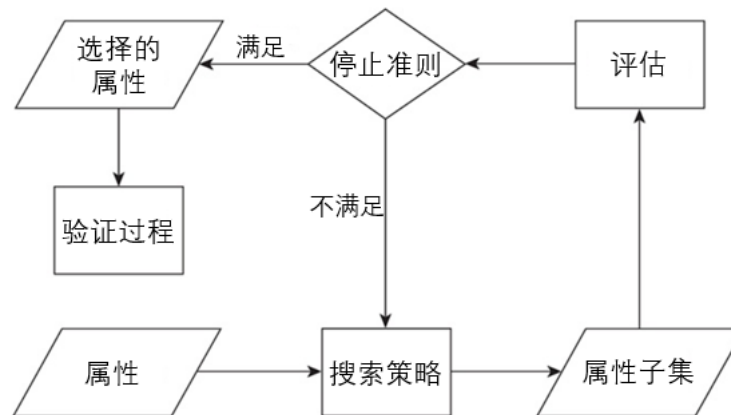
■ 策略通常基于如下一个或多个条件:

- 迭代次数
- 子集评估的度量值是否最优或超过给定的阈值
- 一个特定大小的子集是否已经得到
- 使用搜索策略得到的选择是否可以实现改进



### 3. 数据预处理

#### 4. 特征子集选择



- 一旦选定特征子集，就要验证目标数据挖掘算法在选定子集上的结果。
- 方法1：将使用全部特征得到的结果与使用该特征子集得到的结果进行比较
- 方法2：使用一些不同的特征选择算法得到多个特征子集，然后比较数据挖掘算法在每个子集上的运行结果

## 3. 数据预处理

### 4. 特征子集选择

- 特征加权是另一种保留或删除特征的办法
- 特征越重要，赋予它的权值越大，而对于不太重要的特征，赋予它的权值较小
- 这些权值可以根据特征的相对重要性的领域知识确定，也可以自动确定

## 3. 数据预处理

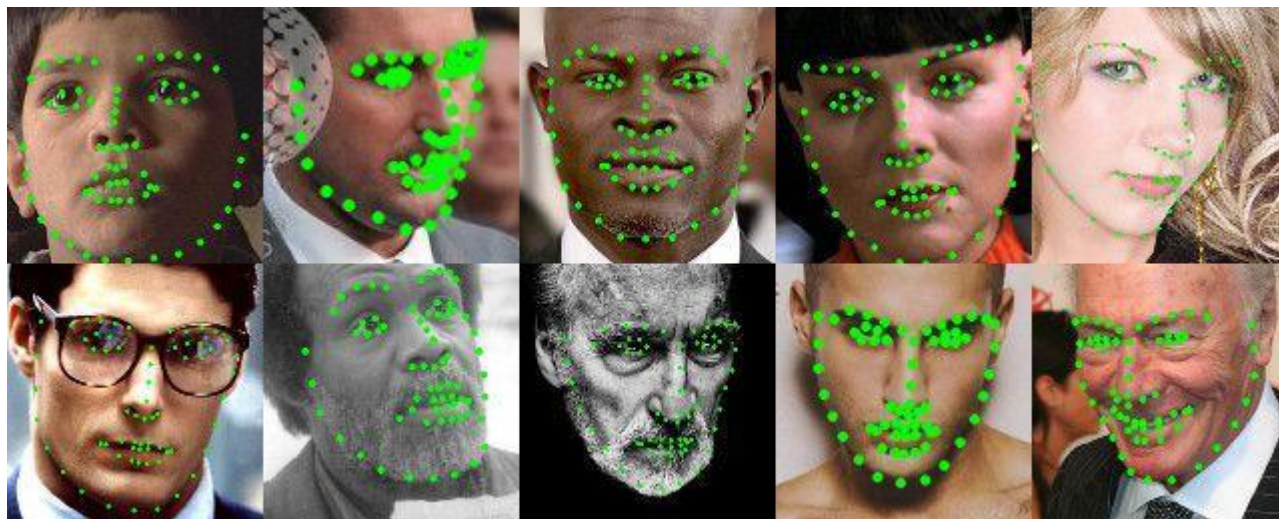
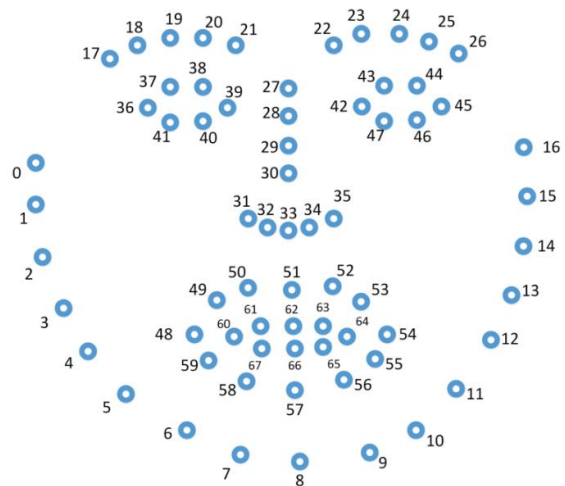
### 5. 特征创建

- 可以由原来的属性**创建新的属性集**，以更有效地捕获数据集中的重要信息
- **新属性的数目**可能比原属性**少**，使得我们可以获得前面介绍的维归约带来的所有好处
- 两种方法：
  - 特征提取
  - 映射数据到新的空间

### 3. 数据预处理

#### 5. 特征创建

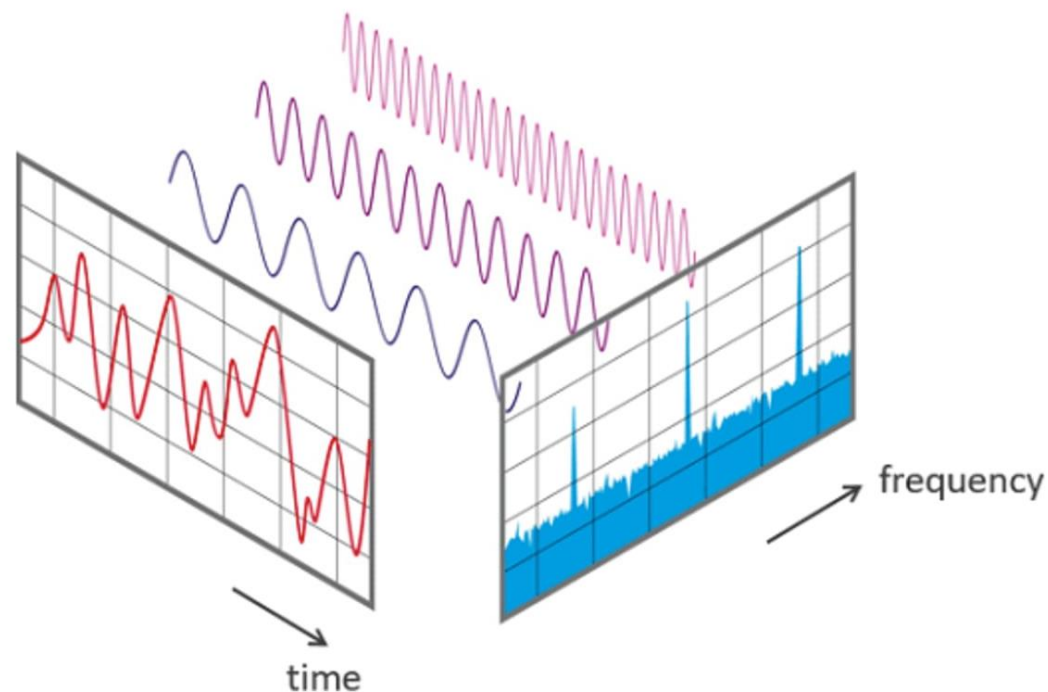
- 特征提取(feature extraction): 由原始数据创建新的特征集
- 将数据挖掘用于一个相对较新的领域, 一个关键任务就是开发新的特征和特征提取方法



### 3. 数据预处理

#### 5. 特征创建

- 映射数据到新空间：使用一种**完全不同的视角**去挖掘数据
- 比如：通过对**时间序列**实施傅里叶变换(Fourier transform)，将它转换成**频率信息**明显的表示，就能检测到这些模式



### 3. 数据预处理

优秀: 85-100  
良好: 70-84  
及格: 60-69  
不及格: <60

## 6. 离散化和二值化

- 某些分类算法, 要求数据是**分类属性形式**
- 发现关联模式的算法要求数据是**二元属性形式**
- 离散化(discretization): 将**连续**属性变换成**分类**属性, 并且连续和离散属性可能都需要变换成一个或多个**二元属性**(二值化, binarization)
- 若一个离散属性有 $m$ 个分类值  $[0, m-1]$ , 则需  $n = \lceil \log_2 m \rceil$  个二元属性

成绩	离散化	成绩
95	→	优秀
80		良好
75		良好
88		优秀
60		及格
55		不及格
68		及格

分类值	整数值	$x_1$	$x_2$
优秀	0	0	0
良好	1	0	1
及格	2	1	0
不及格	3	1	1

二值化

### 3. 数据预处理

#### 6. 离散化和二元化

- 关联分析需要非对称的二元属性，其中只有属性的值为1才是重要的。因此，对于关联问题需要为每一个分类值引入一个二元属性。

分类值	整数值	$x_1$	$x_2$
优秀	0	0	0
良好	1	0	1
及格	2	1	0
不及格	3	1	1



分类值	整数值	$x_1$	$x_2$	$x_3$	$x_4$
优秀	0	1	0	0	0
良好	1	0	1	0	0
及格	2	0	0	1	0
不及格	3	0	0	0	1



### 3. 数据预处理

Temperature:	40	48	60	72	80	90
PlayTennis:	No	No	Yes	Yes	Yes	No
		↑			↑	
		Temp > 54			Temp > 85	

## 6. 离散化和二值化

### ■ 连续属性变换成分类属性涉及两个子任务

- 决定需要多少个分类值 $n$

- 确定如何将连续属性值映射到这些分类值

### ■ 具体步骤:

- 将连续属性值排序后, 通过指定 $n - 1$ 个分割点把它们分成 $n$ 个区间

- 将一个区间中的所有值映射到相同的分类值

### ■ 离散化问题就是决定选择多少个分割点和确定分割点位置的问题

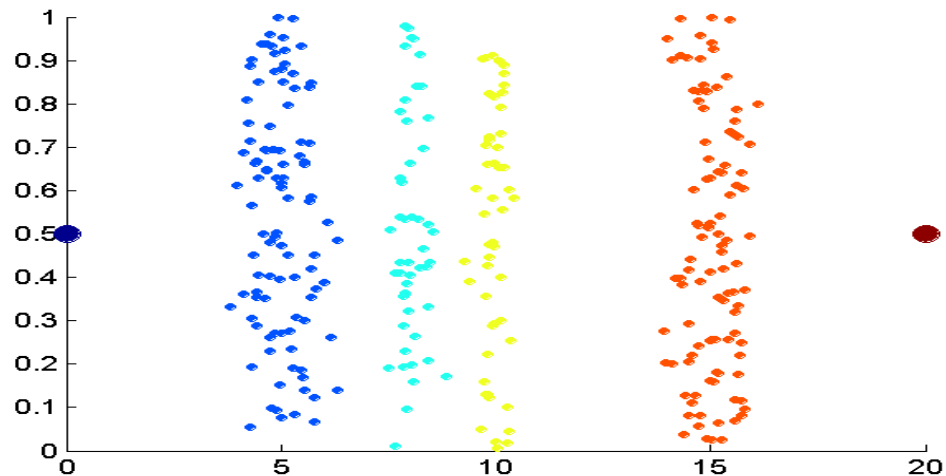
## 3. 数据预处理

### 6. 离散化和二元化

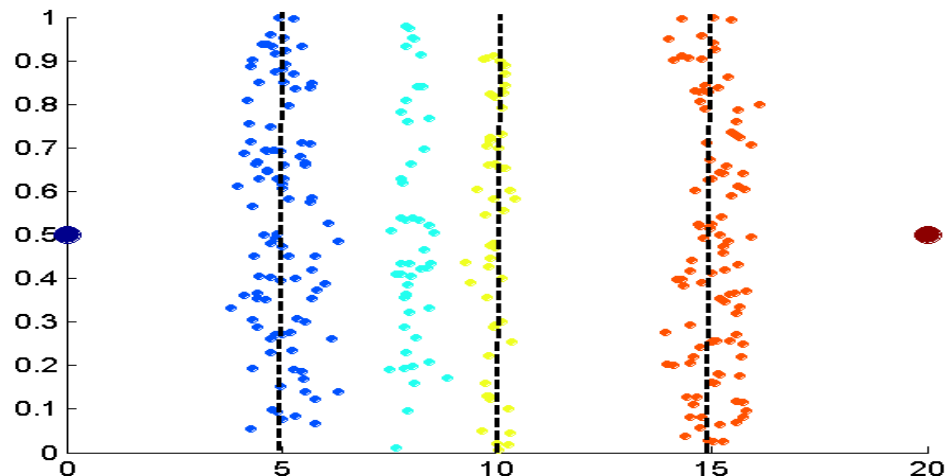
- 无监督离散化：不使用类信息进行离散化
- 等宽(equal width)法：将属性的值域划分成具有相同宽度的区间
  - 区间的个数由用户指定，可能受离群点的影响而性能不佳
- 等频率(equal frequency)或等深(equal depth)：将相同数量的对象放进每个区间
- 诸如K均值等聚类方法

### 3. 数据预处理

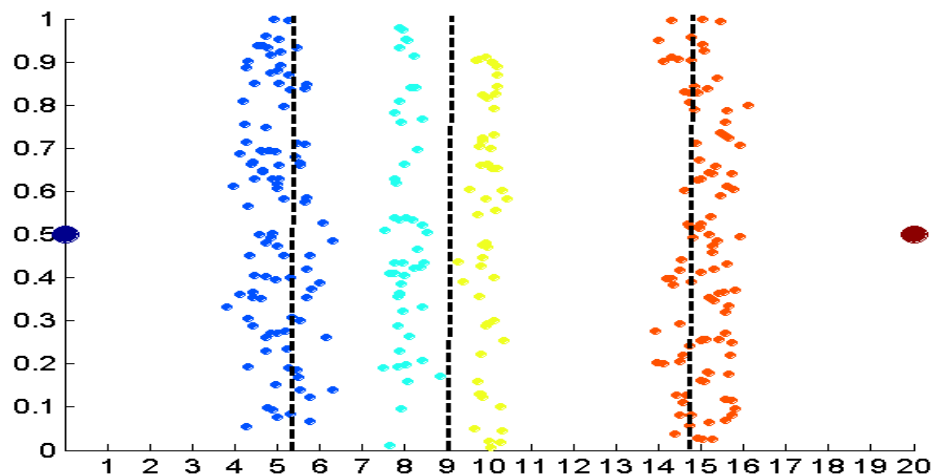
原始  
数据



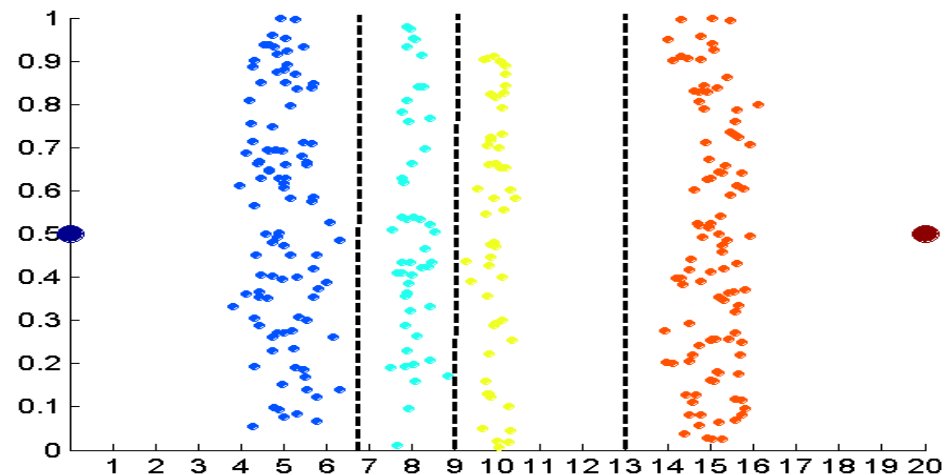
等宽



等频  
(等深)



K均值  
聚类



## 3. 数据预处理

### 6. 离散化和二值化

- **监督**离散化：根据类标对数据进行离散化
- **极大化区间纯度**的方式确定分割点
  - 区间纯度：区间包含单个类别标签的程度
  - 需要人为确定区间的纯度和最小的区间大小
- 一些基于统计学的方法用每个属性值来分隔区间，并通过合并类  
似于根据统计检验得出的相邻区间来创建较大的区间

### 3. 数据预处理

#### 6. 离散化和二元化

■ 基于熵(纯度的度量)的方法是重要的离散化方法之一

■ 熵(entropy): 设 $k$ 是不同的类标号数,  $m_i$ 是某划分的区间 $i$ 中值的个数,  $m_{ij}$ 是区间 $i$ 中类 $j$ 的值得个数, 第 $i$ 个区间得熵是:

$$e_i = -\sum_{j=1}^k p_{ij} \log_2 p_{ij} \quad p_{ij} = \frac{m_{ij}}{m_i} \text{ 第 } i \text{ 个区间中类 } j \text{ 的概率}$$

■ 划分的总熵 $e$ 是每个区间的熵的加权平均

$$e = \sum_{i=1}^n w_i * e_i \quad w_i = \frac{m_i}{m} \text{ 第 } i \text{ 个区间的值的比例}$$

■ 如果一个区间只包含一个类的值(该区间非常纯)则其熵为0; 如果一个区间中的值类出现的频率相等(该区间尽可能不纯), 则其熵最大

## 3. 数据预处理

### 6. 离散化和二元化

- 分类属性有时可能具有太多的值
- 如果分类属性是序数属性，则可以使用类似于处理连续属性的技术，以减少分类值的个数
- 如果分类属性是标称的，就需要使用属性值之间联系的知识，将其合并成较大的组
  - 仅当分组结果能提高分类准确率或达到某种其他数据挖掘目标时，才将值聚集到一起

## 3. 数据预处理

### 7. 变量变换

- 变量变换(variable transformation): 用于变量的所有值的变换
- 一个函数, 它将给定属性的整个值集映射到一组新的替换值, 这样每个旧值都可以用一个新的值来标识
- 简单函数变换
- 规范化或标准化

## 3. 数据预处理

### 7. 变量变换

- 简单函数变换：一个简单的数学函数分别作用于每一个值

- 比如：  $x^k, \log x, e^x, \sqrt{x}, \frac{1}{x}, \sin x, |x|$

- 在统计学中，变量变换(特别是平方根、对数和倒数变换)常用来将不具有高斯(正态)分布的数据变换成具有高斯(正态)分布的数据

- 如果变量具有很大的值域，常用对数变换将其进行压缩

- 使用变量变换时需要小心，因为它们改变了数据的特性

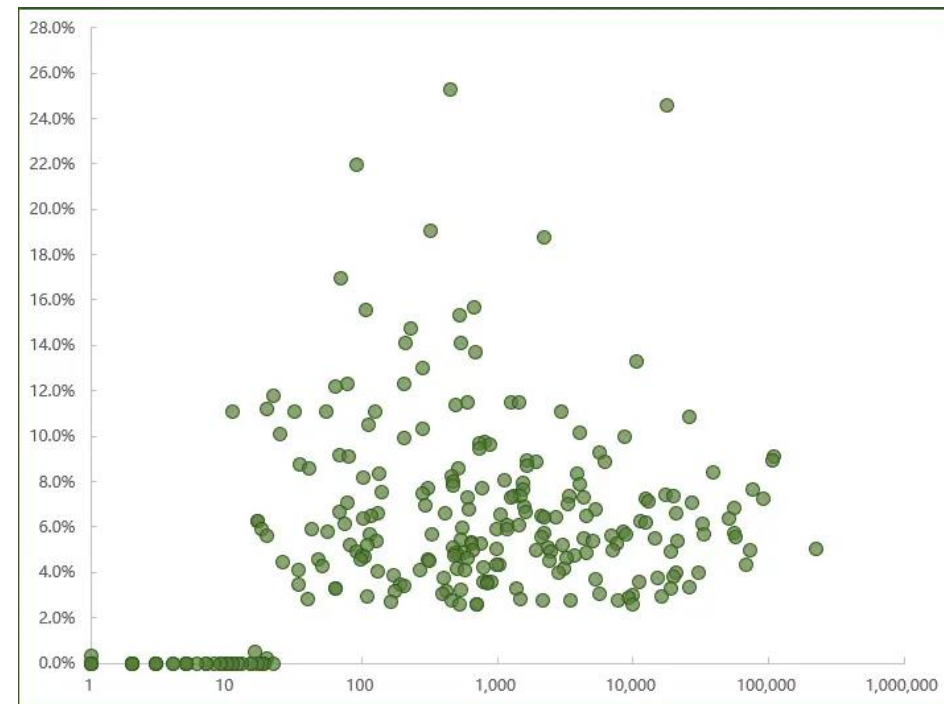
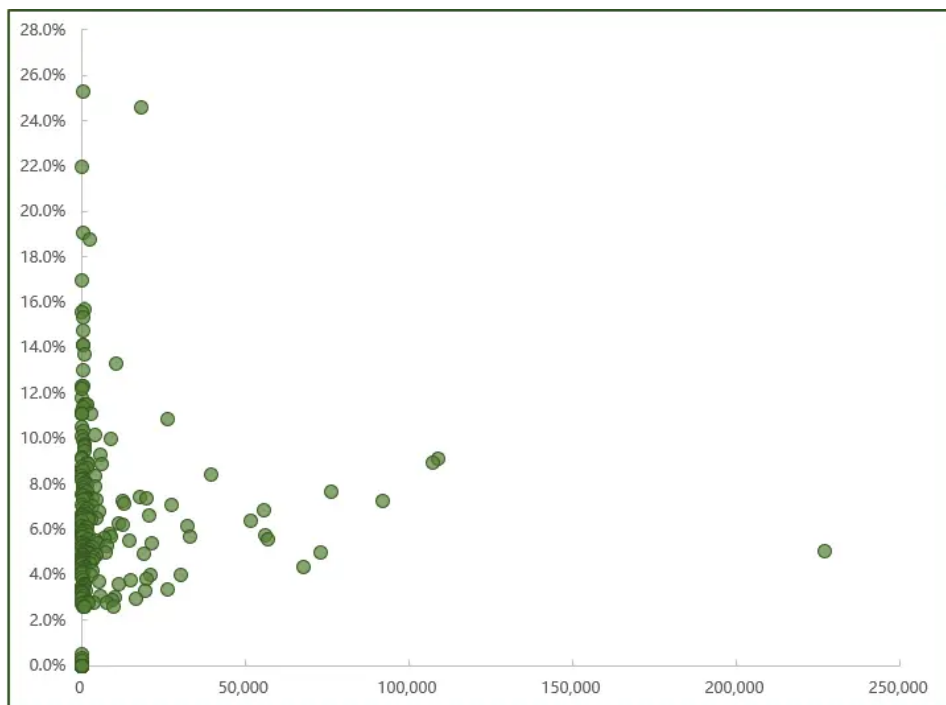
- 新数据特性，是否保序，是否作用0或负值，对0-1间得值有何影响



### 3. 数据预处理

#### 7. 变量变换 --- 对数变换

■ 对数变换：以原数据的对数值作为统计分析的变量值



横坐标的数据差异较大，且集中在较小值的一端

对横坐标取对数后，数据就分散了

## 3. 数据预处理

### 7. 变量变换

- 规范化或标准化：使整个值的集合具有特定的性质
- z分数 (z-score) , 也叫标准分数 (standard score) 是一个数与平均数的差再除以标准差的过程。
- $x' = \frac{(x - \bar{x})}{s_x}$ ,  $\bar{x}$  是属性值的均值(平均值), 而  $s_x$  是它们的标准差, 它具有均值0和标准差1
- 均值和标准差受离群点的影响很大, 可以用中位数取代均值, 用绝对标准差取代标准差

# 目 录

01

数据类型

---

02

数据质量

---

03

数据预处理

---

04

相似性和相异性的度量

---

## 4. 相似性和相异性的度量

### 1. 基础

- 相似度(similarity): 两个对象相似程度的数值度量
  - 非负的
  - 常常在0(不相似)和1(完全相似)之间取值
- 相异度(dissimilarity 距离): 两个对象差异程度的数值度量
  - 对象越类似, 它们的相异度就越低
  - 在区间[0, 1]中取值, 但是0和 $\infty$ 之间取值也很常见
- 邻近度(proximity)表示相似性或相异性

## 4. 相似性和相异性的度量

### 1. 基础

- 相似度到 $[0, 1]$ 区间的变换由如下表达式给出

$$s' = \frac{(s - \min\_s)}{(\max\_s - \min\_s)}$$

$\max\_s$ 和 $\min\_s$ 分别是相似度的最大值和最小值

- 类似的，具有有限值域的相异度也能映射到 $[0, 1]$ 区间

$$d' = \frac{(d - \min\_d)}{(\max\_d - \min\_d)}$$

- 如果相似度(相异度)落在 $[0, 1]$ 区间，则相异度(相似度)可以定义为 $d = 1 - s$ (或 $s = 1 - d$ )

## 4. 相似性和相异性的度量

### 2. 简单属性之间的相似度和相异度

■ 下表显示了两个对象 $x$ 和 $y$ 在单个简单属性方面的相似性和不相似性

属性类型	相异度	相似度
标称	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
序数	$d = \frac{ x-y }{(n-1)}$ 值映射到整数0到 $n-1$ , 其中 $n$ 是值的个数	$s = 1 - d$
区间或比率	$d =  x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d}$ $s = 1 - \frac{(d - \min\_d)}{(\max\_d - \min\_d)}$

## 4. 相似性和相异性的度量

### 3. 数据对象之间的相异度

- 高维空间中两个点 $x$ 和 $y$ 之间的欧几里得距离(Euclidean distance)

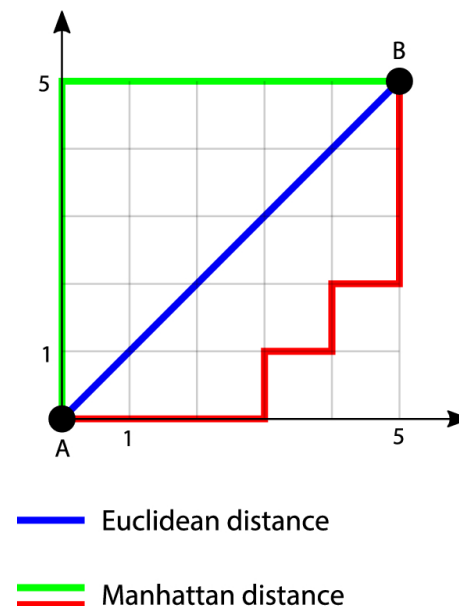
$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad n \text{ 是维度, } x_k \text{ 和 } y_k \text{ 分别是 } x \text{ 和 } y \text{ 的第 } k \text{ 个属性}$$

- 将欧氏距离推广可以得到闵可夫斯基距离(Minkowski distance)

$$d(x, y) = (\sum_{k=1}^n |x_k - y_k|^r)^{1/r}$$

- 闵可夫斯基距离的三个特例

- $r=1$ , 城市街区(也称曼哈顿、出租车、 $L_1$ 范数)距离
- $r=2$ , 欧几里得距离( $L_2$ 范数)
- $r=\infty$ , 上确界( $L_{max}$ 或 $L_\infty$ 范数)距离



## 4. 相似性和相异性的度量

### 3. 数据对象之间的相异度

- 距离(如欧几里得距离)具有一些众所周知的性质。如果 $d(x, y)$ 是两个点 $x$ 和 $y$ 之间的距离, 则如下性质成立:
  - 1)非负性。对于所有 $x$ 和 $y$ ,  $d(x, y) \geq 0$ ; 仅当 $x=y$ 时,  $d(x, y) = 0$
  - 2)对称性。对于所有 $x$ 和 $y$ ,  $d(x, y) = d(y, x)$
  - 3)三角不等式。对于所有 $x$ 、 $y$ 和 $z$ ,  $d(x, z) \leq d(x, y) + d(y, z)$
- 满足以上三个性质的测度称为度量(metric)



## 4. 相似性和相异性的度量

### 4. 数据对象之间的相似度

- 对于相似度，三角不等式(或类似的性质)通常不成立，但是**对称性**和**非负性**通常**成立**
- 如果 $s(x, y)$ 是两个点 $x$ 和 $y$ 之间的相似度，则如下性质成立：
  - 1)非负性。  $0 \leq s \leq 1$ ；仅当 $x=y$ 时,  $s(x, y) = 1$
  - 2)对称性。对于所有 $x$ 和 $y$ ,  $s(x, y) = s(y, x)$

## 4. 相似性和相异性的度量

### 5. 邻近度度量的例子

- 两个仅包含二元属性的对象之间的相似性度量也称为相似系数(similarity coefficient), 并且通常在0和1之间取值, 值为1表明两个对象完全相似, 而值为0表明对象完全不相似。
- 设x和y是两个对象, 都由n个二元属性组成。这样的两个对象(即两个二元向量)的比较可生成如下四个量
  - $f_{00}$ : x取0并且y取0的属性个数
  - $f_{01}$ : x取0并且y取1的属性个数
  - $f_{10}$ : x取1并且y取0的属性个数
  - $f_{11}$ : x取1并且y取1的属性个数

x	1	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	1	0	0	1

$f_{00}$	7	$f_{01}$	2
$f_{10}$	1	$f_{11}$	0

## 4. 相似性和相异性的度量

x	1	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	1	0	0	1

$f_{00}$	7	$f_{01}$	2
$f_{10}$	1	$f_{11}$	0

## 5. 邻近度度量的例子

### ■ 简单匹配系数(Simple Matching Coefficient, SMC)

$$SMC = \frac{\text{值匹配的属性个数}}{\text{属性个数}} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

### ■ Jaccard系数(Jaccard Coefficient, J)

SMC	J
0.7	0

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

### ■ Jaccard系数来处理仅包含非对称的二元属性的对象

## 4. 相似性和相异性的度量

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

## 5. 邻近度度量的例子

- 文档用向量表示，向量的每个分量(属性)代表一个特定的词(术语)在文档中出现的频率
- 每个文档向量都是稀疏的，因为它具有相对较少的非零属性值
- 文档的相似性度量不仅应当像Jaccard 度量一样需要忽略0-0匹配，而且还必须能够处理非二元向量

## 4. 相似性和相异性的度量

x	3	2	0	5	0	0	0	2	0	0
y	1	0	0	0	0	0	0	1	0	2

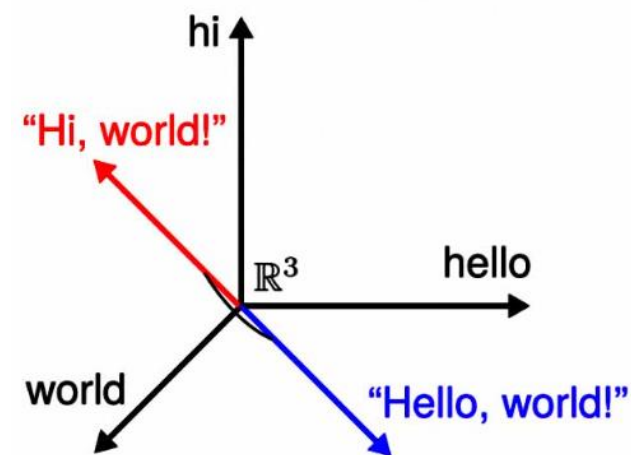
$\cos(x, y)$	0.31
--------------	------

## 5. 邻近度度量的例子

- 余弦相似度(cosine similarity) 是文档相似性最常用的度量之一

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|} = \frac{\sum_{k=1}^n x_k y_k}{\sqrt{\sum_{k=1}^n x_k^2} \sqrt{\sum_{k=1}^n y_k^2}}$$

- 两个向量的内积( $x \cdot y$ )适用于**非对称属性**, 只依赖于两个向量中**非零的分量**
- 两个文档之间的相似性只取决于它们中出现的单词
- 余弦相似度实际上是x和y之间**夹角(余弦)的度量**
- 余弦相似度**不考虑**两个数据对象的**量值**



## 4. 相似性和相异性的度量

### 5. 邻近度度量的例子

#### ■ 广义Jaccard系数(Tanimoto 系数)

x	3	2	0	5	0	0	0	2	0	0
y	1	0	0	0	0	0	0	1	0	2

↓

$T(x, y)$	0.116
-----------	-------

$$T(x, y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y} = \frac{\sum_{k=1}^n x_k y_k}{\sum_{k=1}^n x_k^2 + \sum_{k=1}^n y_k^2 - \sum_{k=1}^n x_k y_k}$$

#### ■ 在二元属性情况下归约为Jaccard系数

$$T(x, y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y} = \frac{f_{11}}{(f_{10} + f_{11}) + (f_{01} + f_{11}) - f_{11}}$$

x	1	1	0	1	0	0	0	1	0	0
y	1	0	0	0	0	0	0	1	0	1

→

$T(x, y)$	0.4
-----------	-----

## 4. 相似性和相异性的度量

### 5. 邻近度度量的例子

- 相关性经常被用来测量两组被观察到的值之间的线性关系
- 相关性可以测量两个变量(高度和重量)之间或两个对象(一对温度时间序列)之间的关系
- 如果两个数据对象中的值来自不同的属性，可以使用相关性来度量属性之间的相似度
- 统计学中的三大相关性系数: Pearson, Spearman, Kendall

## 4. 相似性和相异性的度量

x	-3	6	0	3	-6		$corr(x, y)$
y	1	-2	0	-1	2		-1
x	3	6	0	3	6		$corr(x, y)$
y	1	2	0	1	2		1

## 5. 邻近度度量的例子

■ 皮尔森相关(Pearson's correlation)系数是协方差与标准差的比值

$$corr(x, y) = \frac{covariance(x, y)}{standard_{deviation}(x) \times standard_{deviation}(y)} = \frac{S_{xy}}{S_x S_y}$$

$$S_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad S_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2} \quad S_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

■ 相关度总是在-1到1之间取值

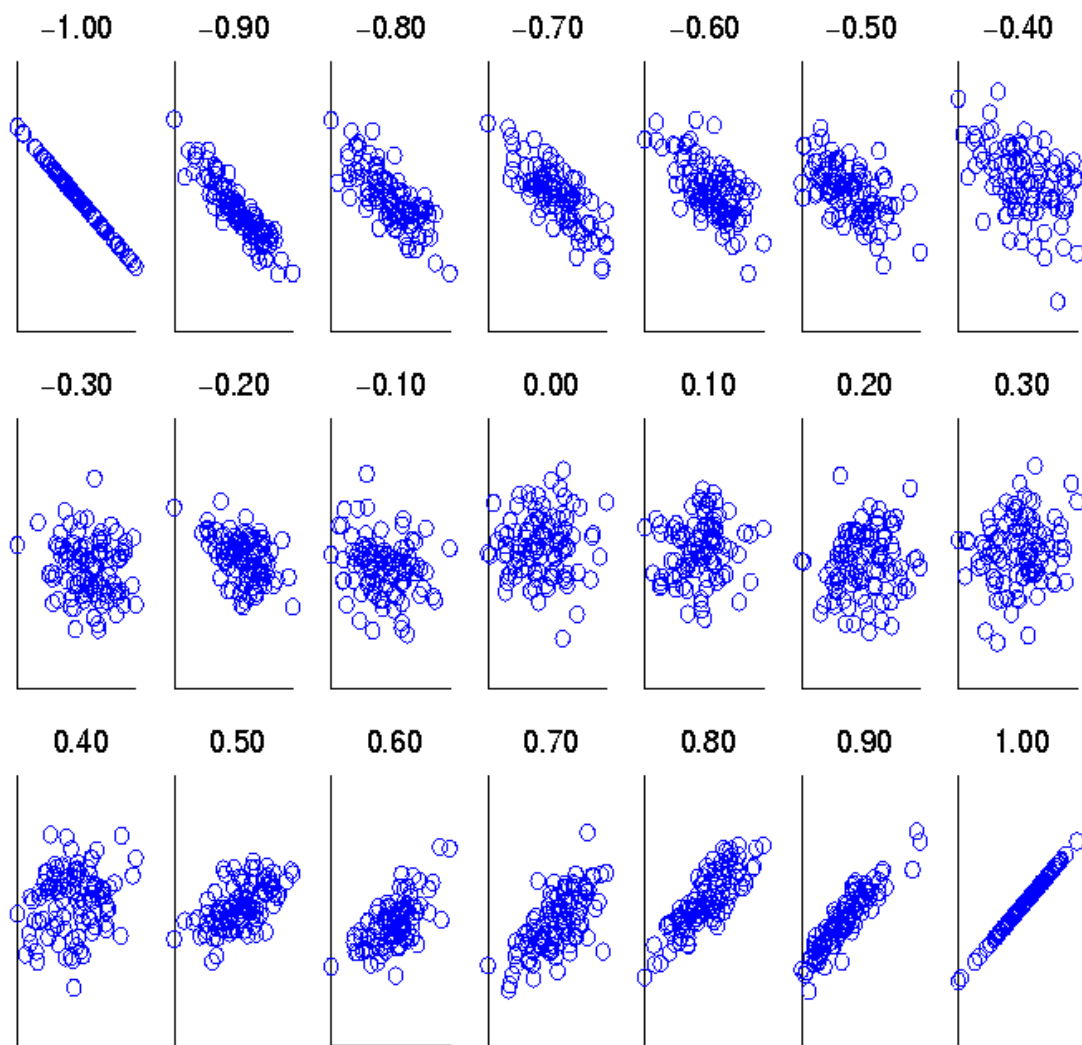
■ 相关度为 1(-1), 意味x和y具有完全正(负)线性关系

■ 相关度为0, 则两个数据对象的属性之间不存在线性关系, 仍然可能存在非线性关系

x	-3	-2	-1	0	1	2	3		$corr(x, y)$
y	9	4	1	0	1	4	9		0



## 4. 相似性和相异性的度量



- $x$ 和 $y$ 具有30个属性
- 图中每个小圆圈代表30个属性中的一个, 其 $x$ 坐标是 $x$ 的一个属性的值, 而其 $y$ 坐标是 $y$ 的相同属性的值
- $x$ 和 $y$ 的相关度从-1到1

## 4. 相似性和相异性的度量

$\mathbf{x} = (1, 2, 4, 3, 0, 0, 0)$ ,  $\mathbf{y} = (1, 2, 3, 4, 0, 0, 0)$   
 $\mathbf{y}_s = \mathbf{y} * 2$  (缩放),  $\mathbf{y}_t = \mathbf{y} + 5$  (平移)

Measure	( $\mathbf{x}$ , $\mathbf{y}$ )	( $\mathbf{x}$ , $\mathbf{y}_s$ )	( $\mathbf{x}$ , $\mathbf{y}_t$ )
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

## 5. 邻近度度量的例子

- 如果对数据对象进行数据变换之后，该邻近度度量方法的值保持不变，则该邻近度度量方法被认为**对数据变换具有不变性**
- 相关性度量对于缩放和平移都有不变性，而余弦度量只对缩放具有不变性。
- 闵可夫斯基距离度量对缩放和平移都是敏感的

性质	余弦	相关性	闵可夫斯基距离
缩放不变(乘法)	是	是	否
平移不变(加法)	否	是	否

## 4. 相似性和相异性的度量

### 6. 邻近度计算中的问题

- (1) 当属性具有不同的尺度(scale)或相关时如何处理?
- (2) 当对象包含不同类型的属性(例如, 定量属性和定性属性)时如何计算对象之间的邻近度?
- (3) 当属性具有不同的权重(即并非所有的属性都对对象的邻近度具有相等的贡献)时, 如何处理邻近度计算?

## 4. 相似性和相异性的度量

### 6. 邻近度计算中的问题

- 当属性相关、具有不同的值域(不同的方差), 并且数据分布近似于高斯(正态)分布时, 马氏(Mahalanobis)距离是有用的
- 两个对象(向量) $x$ 和 $y$ 的马氏距离定义为:

$$M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

- $\Sigma^{-1}$ 是数据协方差矩阵的逆, 协方差矩阵的第 $ij$ 个元素是第 $i$ 个和第 $j$ 个属性的协方差

## 4. 相似性和相异性的度量

$$M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix}$$

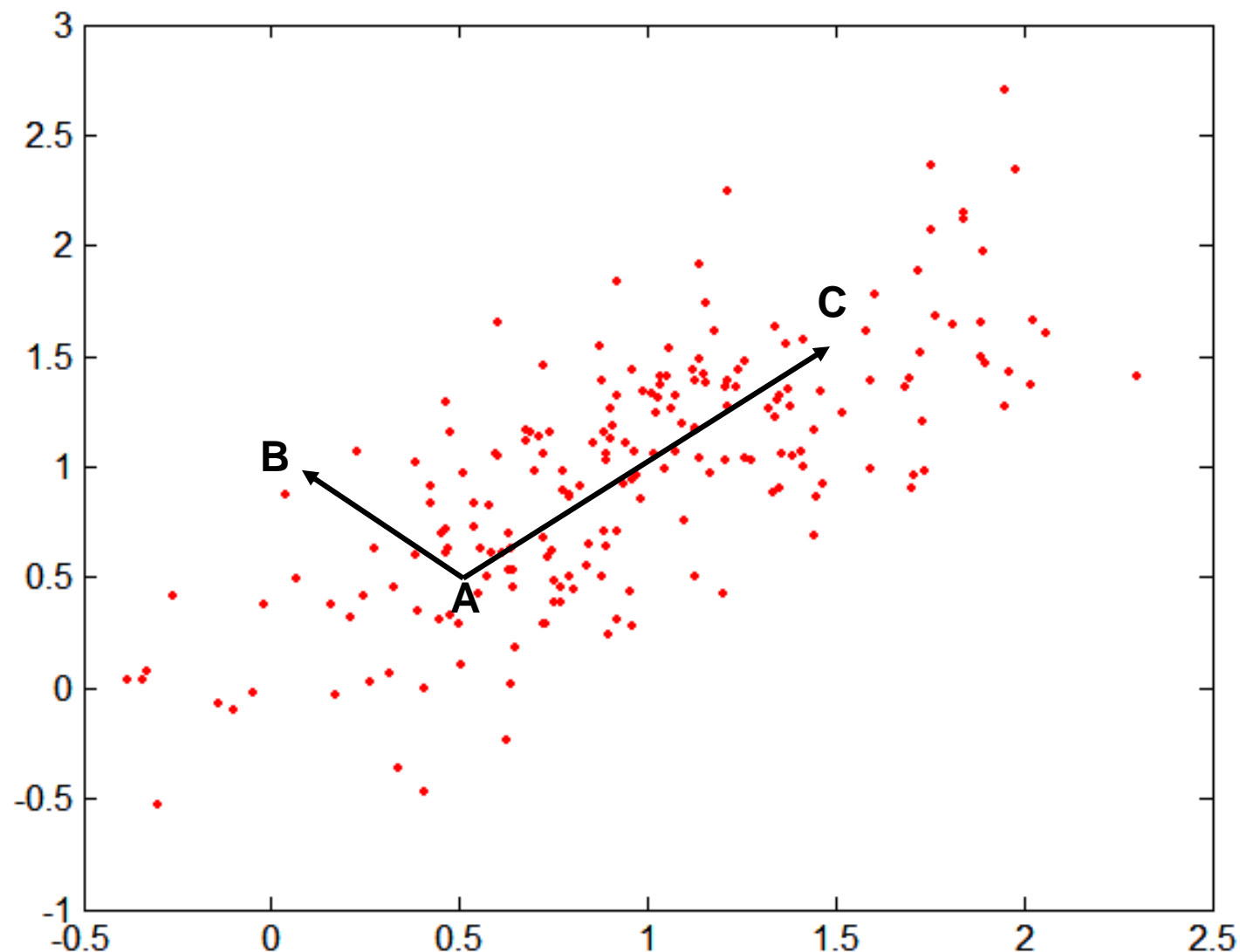
**A: (0.5, 0.5)**

**B: (0, 1)**

**C: (1.5, 1.5)**

**M (A,B) = 5**

**M (A,C) = 4**



## 4. 相似性和相异性的度量

### 6. 邻近度计算中的问题

- 当属性具有不同类型时，直接了当的方法是**分别计算**出每个属性之间的相似度，然后使用一种输出为0和1之间相似度的方法**组合这些相似度**
  - 将总相似度定义为所有属性相似度的平均值
  - 当某些属性是**非对称属性**，如果两个对象在这些属性上的**值都是0**，则在计算对象相似度时**忽略它们**

## 4. 相似性和相异性的度量

### 6. 邻近度计算中的问题

#### 算法2.2 异构对象的相似度

1: 对于第 $k$ 个属性, 计算相似度  $s_k(x, y)$ , 取值范围为区间 $[0, 1]$

2: 对于第 $k$ 个属性, 定义一个指示变量 $\delta_k$ , 如下:

$$\delta_k = \begin{cases} 0, & \text{如果第} k \text{个属性是非对称属性, 并且两个对象在该属性上的值都是0;} \\ & \text{或者如果其中一个对象的第} k \text{个属性含有缺失值} \\ 1, & \text{其他} \end{cases}$$

3: 使用如下公式计算两个对象之间的总相似度:

$$\text{similarity}(x, y) = \frac{\sum_{k=1}^n \delta_k * s_k(x, y)}{\sum_{k=1}^n \delta_k}$$

## 4. 相似性和相异性的度量

### 6. 邻近度计算中的问题

■ 当某些属性对邻近度的定义比其他属性更重要时，可以通过对每个属性的贡献**加权**来修改邻近度公式

■ 属性权重为 $w_k$ 时,  $similarity(x, y) = \frac{\sum_{k=1}^n w_k * \delta_k * s_k(x, y)}{\sum_{k=1}^n w_k * \delta_k}$

■ 闵可夫斯基距离的定义也可以修改为

$$d(x, y) = (\sum_{k=1}^n w_k * |x_k - y_k|^r)^{1/r}$$