第二章 数据

1. 不同的属性类型
2. 数据质量问题
3. 数据预处理的主要方法
4. 连续属性离散化
5. 相似性/相关性度量：**距离、余弦、SMC、Jaccard系数**、皮尔森相关系数

1. 不同的属性类型

1. 属性

   属性(attribute): 对象的性质或特性

   属性也称:变量、特性、字段 or 维

2. 测量标度

   测量标度(measurement scale): 将数值和符号值与对象的属性相关联的规则(函数)

3. 属性的类型(测量标度的类型)

   取决于下列4种性质

   - Distinctness(相异性): = ≠
   - Order(序):<>
   - Addition: + -
   - Multiplication(乘法): * /

   结合四种属性,可定义四种属性类型

   1. 分类的(定性的)
      - 标称
      - 序数
   2. 数值的(定量的)
      - 区间
      - 比率

属性类型		描　述	例　子	操　作
分类的（定性的）	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象（=，≠）	邮政编码、雇员ID号、眼球颜色、性别	众数、熵、列联相关、χ^2检验
	序数	序数属性的值提供足够的信息确定对象的序（<，>）	矿石硬度、{好，较好，最好}、成绩、街道号码	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，值之间的差是有意义的，即存在测量单位（+，−）	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、t和F检验
	比率	对于比率变量，差和比率都是有意义的（*，/）	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差

2. 数据质量问题

2.1.2 数据集的类型

1. 数据集的一般特性
    - 维度(Dimensionality)
        - 维度是数据集中的对象具有的属性数目
        - 维灾难(curse of dimensionality)
        - 维归约(dimensionality reduction)
    - 稀疏性(sparsity)
        - 一个对象大部分属性上的值为0
        - 只存储和处理非零值
    - 分辨率(resolution)
        - 数据的模式依赖于分辨率——度量尺度 (scale)
2. 数据集类型
    - 记录数据(record)
        - 数据矩阵(Data Matrix)
        - 文本数据(Document Data)：每篇文档可以表示成一个文档-词矩阵
        - 事务数据(Transaction Dat)
    - 基于图形的数据(graph)
        - World Wide Web
        - 分子结构（Molecular Structures）
    - 有序数据(ordered)
        - 空间数据(Spatial Data)
        - 时间数据(Temporal Data)
        - 序列数据(Sequential Data)

2.2 数据质量

2.2.1 测量和数据收集问题

- 测量误差和数据收集错误
- 噪声和伪像
- 精度、偏倚、准确率
- 离群点
- 遗漏值
- 不一致的值
- 重复的值

2.3 数据预处理

数据预处理的主要方法

- 聚集 (Aggregation):Combining two or more attributes (or objects) into a single attribute (or object)
- 抽样 (Sampling):是一种选择数据对象子集进行分析的常用方法
- 维归约 (Dimensionality Reduction)
- 特征子集选择 (Feature subset selection)
    - emmbedded approach
    - wrapper approach
    - filter approach

- 特征创建 (Feature creation)
  - 特征提取(Feature Extraction)
  - 映射数据到新的空间(Mapping Data to New Space)
  - 特征构造(Feature Construction)
- 离散化与二元化 (Discretization and Binarization)
  - 离散属性二元化
  - 连续属性离散化
- 属性变换 (Attribute Transformation)
  - 简单变换
  - 标准化(standardization)或规范化(normalization)

连续属性二元化:

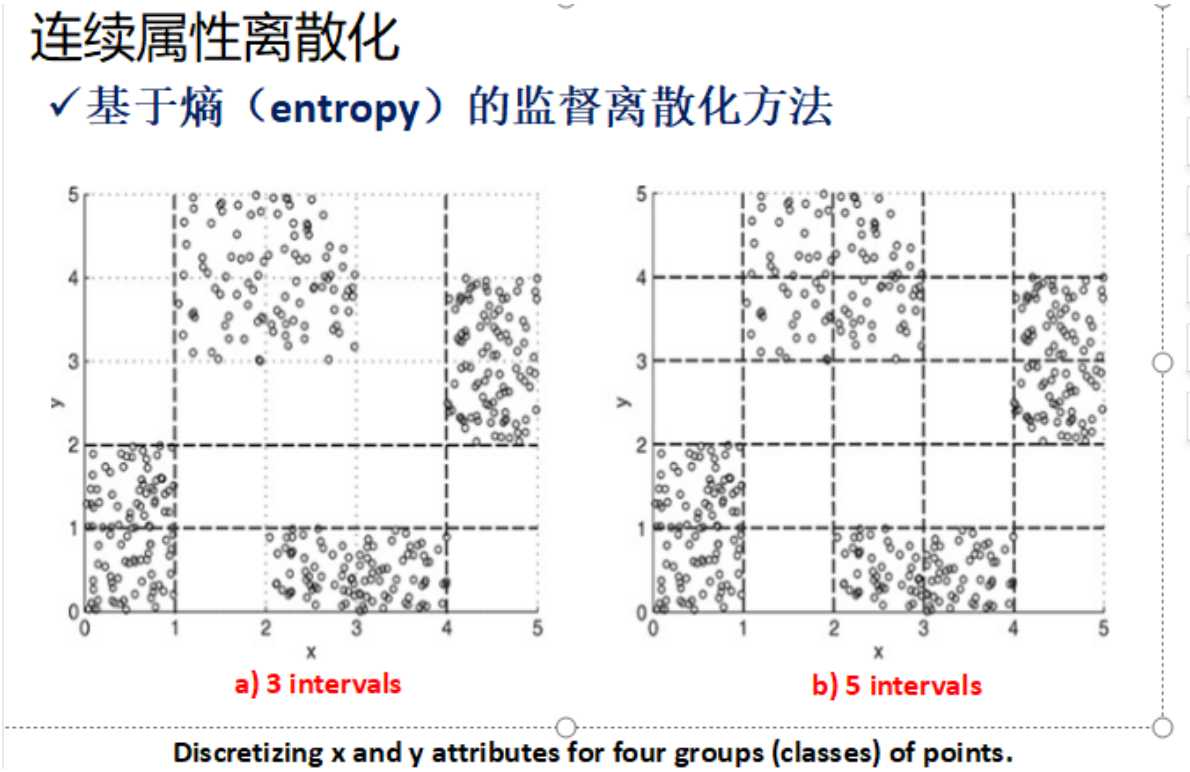
explain:

类信息:

In set theory and its applications throughout mathematics, a class is a collection of sets (or sometimes other mathematical objects) that can be unambiguously defined by a property that all its members share.

使用类信息(supervised)还是不使用类信息(unsupervised)

- 非监督离散化
- 监督离散化



Discretizing x and y attributes for four groups (classes) of points.

Question:

where use the supervised???

2.4 相似性和相异度的度量

2.4.3 数据对象之间的相异度

距离: 具有特定性质的相异度

Euclidean distance(欧式距离)

d(x,y) = $\sqrt {\sum_{k=1}^{n} {(x_k-y_k)^2}}$

欧几里得距离可以用Minkowski distance来推广

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

- r = 1, Manhantann distance
- r = 2, Euclidean distance
- r = ∞, Supremun distance

$d(x,y) =\lim{x \to \infty} \left(\sum{k=1}^{n} {|x_k-y_k|^r}\right)^{\frac{1}{r}} = max{|x_k-y_k|,k=1,2, \dots, n}$

距离的性质

1. 非负性
2. 对称性
3. 三角不等式

满足以上三个性质称为度量(metric)

有些相异度无法满足度量性质

2.4.5 邻近性度量的例子

x,y为两个对象, 都由n个二元属性组成

- f_{00}: x取0 y取0
- f_{01}: x取0 y取1
- f_{10}: x取1 y取0
- f_{11}: x取1 y取1

1. 简单匹配系数(Simple Matching Coefficient)

   该度量对出现和不出现都进行计数

   SMC = $\frac{f{11}+f{00}}{f{01}+f{10}+f{11}+f{00}}$

   SMC可用于是非题就按测回答问题相似学生

2. Jaccard系数(Jaccard Coefficient)

   Jaccard 假定x和y是两个数据对象，代表一个事物矩阵的两行,忽略0-0匹配

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

3. 余弦相似度

   忽略0-0匹配 && 处理非二元向量

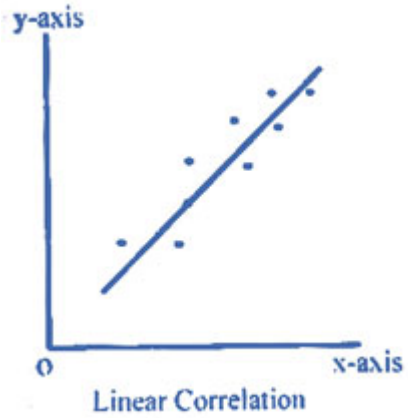
   $\cos(x,y) = \frac{x · y}{||x|| ||y||} = x'y'$

   - $||x||$为向量x的长度
   - $x'=\frac{x}{||x||}$: 长度为1的向量

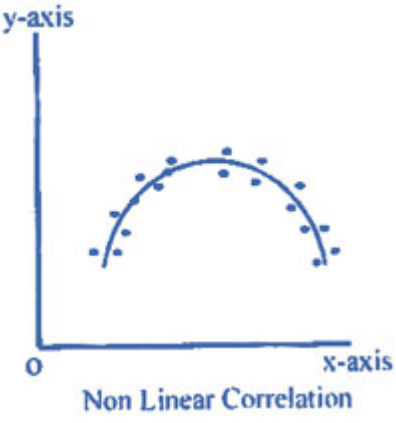
4. 皮尔森相关系数

   explaination:

   线性: Correlation is said to be linear if the ratio of change is constant



Linear Correlation



Non Linear Correlation

对象之间的相关性是对象属性之间**线性**联系的度量

$corr(x,y) = \frac{covariance(x,y)}{standard_deviation(x)× standard_deviation(y)} = \frac{s_{xy}}{s_xs_y}$

$$\text{covariance}(\mathbf{x},\mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) \qquad (2\text{-}11)$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n}x_k \text{ 是 } \mathbf{x} \text{ 的均值}$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n}y_k \text{ 是 } \mathbf{y} \text{ 的均值}$$

---

to be studied
stardard_deviation and covariance
