聚类

Cluster Analysis I

1.1 概念

- 集群：在同一集群中彼此相似的数据对象的集合
- 聚类分析：根据数据中的特征发现数据之间的相似性，并将相似的数据对象分组到集群中
- 无监督学习：没有预定义的类。可作为洞察数据分布的独立工具，作为其他算法的预处理步骤

1.2 应用

模式识别，图像处理

- 金融科技

  市场营销，保险

- 空间数据分析

  土地利用，地震研究，气候

- 信息检索

  文档聚类，生物分类

1.3 什么是好的聚类

- 类内相似度高
- 类间相似度低

- 发现隐藏的模式

1.4 衡量聚类的质量

- 相似性度量

  使用距离函数`d(i,j)`

  对于区间尺度变量、布尔变量、分类变量(categorical variables)、序数比、向量 是不同的，权重与不同的变量关联

  无法定义足够相似或足够好

1.5 注意事项

- 划分的标准

  单层分区与多层分区

  排他性，非排他性（一个物品可属于多个类别）

  基于距离，基于连接（密度，连续性）

  全空间，子空间（高维空间）

1.6 计算

- 平均绝对偏差

  $\frac{1}{n}(|x_{1f}-m_f|+|x_{2f}-m_f|+\cdots+|x_{nf}-m_f|)$

  其中 $m_f=\frac{1}{n}(x_{1f}+x_{2f}+\cdots+x_{nf})$

- 标准化测量(z-score)

  $z_{if}=\frac{x_{if}-m_f}{s_f}$

- 闵可夫斯基距离

  $d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\cdots+|x_{ip}-x_{jq}|^q)}$

  $\lim\limits_{q\to\infty}d(i,j)$为切比雪夫距离，q=1为曼哈顿距离，q=2为欧几里得距离

- 对称二进制变量距离

i/j	1	0	sum
1	a	b	a+b
0	c	d	c+d
sum	a+c	b+d	N

  $d(i,j)=\frac{b+c}{a+b+c+d}=\frac{b+c}{N}$

- 不对称二进制变量距离

  $d(i,j)=\frac{b+c}{a+b+c}$

- jaccard相似度

  $sim_{Jaccard}(i,j)=\frac{a}{a+b+c}$

- 分类变量

  $d(i,j)=\frac{N-m}{N}$

  m是匹配数，N是变量总数

- 比例变量

  $y_{if}=\log(x_{if})$

1.7 k-means

随机选择k个对象作为质心，计算每个点到质心的距离，将点分配到质心，然后更新质心，再迭代

- 优点：高效，$O(tkn)$，n对象，k集群，t迭代
- 缺点：局部最优，对噪声和异常值敏感，不适合非凸形状

1.8 k-medoids

质心是对象

对小数据好，对大数据不够好，对造成和异常值更好

$O(k(n-k)^2)$，n是对象，k是类

Cluster Analysis II

2.1 密度聚类

可发现任意形状

2.1.1 DBSCAN

半径r，MinPts=m，每个对象的圈包含至少m个对象，将这样的对象初始化

然后把圈里满足MinPts>=m的加入并继续扩展，不满足的加入不扩展

2.2 基于网格 STING

将空间划分成矩形单元，多层矩形对应不同分辨率，用置信度来判断分块是不是一起的

2.3 基于模型 EM(期望最大化)

随机k个中心

$P(X_i \in C_k) = p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)}$

重新估计模型参数$m_k = \frac{1}{N}\sum_{i=1}^{n}\frac{X_i P(X_i \in C_k)}{\sum_j P(X_i \in C_j)}$

迭代到收敛