聚类算法

· 无监督学习

目标：给定一组样本，依据相似度或距离，将其归并到若干类或簇(cluster)中。

应用：客户细分和用户画像

层次聚类(hierarchical clustering)

k-means 聚类(K-means clustering)

1. 相似度(similarity)和距离(distance)度量

N个样本 $x_i=(x_{i1}, x_{i2}, \cdots, x_{ip})$

(1) 闵可夫斯基距离(Minkowski Distance)

$$d_{ij}=\left|\sum_{k=1}^{p}|x_{ik}-x_{jk}|^m\right|^{1/m}$$

当 $m=2$ 时，为欧式距离

$m=1$ 时，为曼哈顿距离

$m=\infty$ 时，为切比雪夫距离

(2) 马氏距离

$$d_{ij}=\left[(x_i-x_j)^T S^{-1}(x_i-x_j)\right]^{1/2}$$

，$S\in R^{p\times p}$ 是协方差矩阵，若 $S=I$，与欧式距离一致

(3) 相关系数(Correlation)

$$r_{ij}=\frac{\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)^2\sum_{k=1}^{p}(x_{jk}-\bar{x}_j)^2}$$

，其中 $\bar{x}_i=\frac{1}{p}\sum x_{ik}$

$\bar{x}_j=\frac{1}{p}\sum x_{jk}$

(4) 夹角余弦

$$s_{ij}=\frac{\sum_{k=1}^{p}x_{ik}x_{jk}}{\left[\sum_{k=1}^{p}x_{ik}^2\sum_{k=1}^{p}x_{jk}^2\right]}$$

，常用于文本分析

距离 $(d(x,y))\longrightarrow$ 相似度 $(s(x,y))$：$S(x,y)=\dfrac{1}{1+d(x,y)}$

相似度 $\longrightarrow$ 距离：$d(x,y)=\sqrt{2(1-s(x,y))}$

2. 类或簇(cluster)

(1) 定义

用 G 表示 cluster，用 x_i 与 x_j 表示 G 的样本，$n_G = |G|$，d_{ij} 为距离

定义1：设 T 为给定正数，若 G 中任意两样本 x_i, x_j，有 $d_{ij} \le T$，则记 G 为一个类或簇

2：若 G 中任意相本 x_i，存在另一个样本 x_j，有 $d_{ij} \le T$

3：对 G 中任意 x_i 成立 $\frac{1}{n_G - 1} \sum_{x_j \in G} d_{ij} \le T$

4：$\frac{1}{n_G(n_G - 1)} \sum_{x_i \in G} \sum_{x_j \in G} d_{ij} \le T$

(2) 常用特征

(a) 类的均值(类中心)：$\bar{x}_G = \frac{1}{n_G} \sum_{i=1}^{n_G} x_i \in R^p$

(b) 类的直径(diameter)：$D_G = \max_{x_i, x_j \in G} d_{ij}$

3. 类与类之间的距离(linkage)

类连接：设 G_p 包含 n_p 个样本，G_q 包含 n_q 个样本，类中心为 $\bar{x}_p$ 和 $\bar{x}_q$

(a) 最短距离或单连接(single linkage)

$$D_{pq} = \min\{d_{ij} \mid x_i \in G_p, x_j \in G_q\}$$

(b) 最长距离或完全连接(Complete linkage)

$$D_{pq} = \max\{d_{ij} \mid x_i \in G_p, x_j \in G_q\}$$

层次聚类

聚合聚类 / 分裂聚类

(1) 开始对每个样本各成一类

(2) 将类别最近的两个类合并

(3) 重复直到停止

三个要素：

(1) 距离或相似度

(2) 合并规则

(3) 终止条件 $D = (d_{ij})$

K-均值聚类

目标: 将n个样本划分为K个cluster(提前设置K), K个类 $G_1, G_2, \cdots, G_K$ 是对样本X的划分, $G_i \cap G_j = \phi$ $(i \neq j)$, $\bigcup_{i=1}^{k} G_i = X$

用C表示划分多对一个的函数 $l = C(i)$ $(i=1,2,\cdots,n;\ l=1,2,\cdots,K)$

1. 策略

通过损失函数最小化选取C

$$W(C) = \sum_{l=1}^{K} \sum_{C(i)=l} \| x_i - \bar{x}_l \|^2, \quad \bar{x}_l = \frac{1}{n_l} \sum_{C(i)=l} x_i$$

W(C)表示类内样本的相似程度

$$C^* = \arg\min_{C} W(C)$$

但这是一个NP-hard问题

将n个样本分为K个类 $S(n,k) = \frac{1}{k!} \sum_{l=1}^{k} (-1)^l C(k,l)(l-k)^n$

2. 算法

给定C, $\bar{x}_l = \frac{1}{n_l} \sum_{C(i)=l} x_i$

(1) 给定类中心 $\{m_1, m_2, \cdots, m_K\}$, 寻找 $C(\cdot)$ 使W(C)最小,

$\min\limits_{C} \sum_{l=1}^{K} \sum_{C(i)=l} \| x_i - \bar{x}_l \|^2$, 计算 x_i 与每个类中心的最小值, 将 x_i 分到最近的类中

(2) 给定划分, 求类中心 $\{m_1, m_2, \cdots, m_K\}$, $m_l = \frac{1}{n_l} \sum_{C(i)=l} x_i$

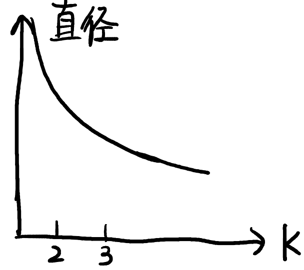
重复(1)-(2)直到达到终止条件

remark:

(1) 收敛性: 不能保证全局最优, 依赖于初值的选取

(2) 初值: 通过层次聚类

(3) 类别数K的选择



聚类性能的度量

①与参考模型(Reference Model)进行比较

②不利用参考模型:内部指标(Internal Index)

外部指标

设聚类给出的划分是 $C=\{C_1, C_2, \cdots, C_k\}$,

参考模型给出 $C^*=\{C_1^*, C_2^*, \cdots, C_k^*\}$,记类别的标签 Y_i 与 Y_i^*

$$a=\#\{(x_i,x_j) \mid Y_i=Y_j, Y_i^*=Y_j^*, i<j\}$$

$$b=\#\{(x_i,x_j) \mid Y_i=Y_j, Y_i^*\neq Y_j^*, i<j\}$$

$$c=\#\{(x_i,x_j) \mid Y_i\neq Y_j, Y_i^*=Y_j^*, i<j\}$$

$$d=\#\{(x_i,x_j) \mid Y_i\neq Y_j, Y_i^*\neq Y_j^*, i<j\}$$

Jaccard 系数: $JC=\dfrac{a}{a+b+c}$

FM 指数: $FMI=\sqrt{\dfrac{a}{a+b}\times\dfrac{a}{a+c}}$

Rnd: $RI=\dfrac{2(a+d)}{n(n-1)}$

内部指标

$$avg(C)=\frac{2}{|C|(|C|-1)}\sum_{1\le i<j\le |C|} dist(x_i,x_j)$$

$$diam(C)=\max_{1\le i<j\le |C|} dist(x_i,x_j)$$

$$d_{min}(C_i,C_j)=\min_{x_i\in C_i, x_j\in C_j} dist(x_i,x_j)$$