第五章作业

1. 聚类算法可分为哪几类？影响聚类算法性能优劣的因素有哪些？

答：

(1) 增类聚类算法（基于分裂的聚类算法）

(2) 减类聚类算法（基于合并的聚类算法）

(3) 动态聚类算法

因素：聚类中心的选择与更新、聚类策略和聚类准则的选择、控制阈值和类别数的设置等

2. 请给出最小张树聚类算法的具体过程，并分析其优缺点。

答：（课本第 218~221 页）

优点：简洁、明快，特别适用于距离较远的两个密集样本点集之间的区分。

缺点：1）当样本集合中存在噪声样本时可能造成错分；2）对于相距较近的两个密集样本点集的区分能力较弱。

3.有以下四个样本：

$$\boldsymbol{X_1=(3,5)^T}，\ \boldsymbol{X_2=(5,1)^T}，\ \boldsymbol{X_3=(1,0)^T}，\ \boldsymbol{X_4=(1,4)^T}$$

初始划分为两类，$\boldsymbol{\omega_1}$：$\{\boldsymbol{X_1}，\boldsymbol{X_2}\}$ 和 $\boldsymbol{\omega_2}$：$\{\boldsymbol{X_3}，\boldsymbol{X_4}\}$，则：

（1）若将$\boldsymbol{X_2}$移到$\boldsymbol{\omega_2}$类中，试计算出转移前后的总的类内离散度矩阵$\boldsymbol{S_w}$，请给出具体计算过程。

（2）若使用$\boldsymbol{S_w}$的行列式作为聚类准则，（1）中的转移是否合适？请给出具体过程。

（3）若使用 C-均值算法的准则函数，（1）中的转移是否合适？请给出具体过程。

答：（1）

类内离散度计算公式：

$$S_w=\sum_{j=1}^{c}P_jS_j$$

$$S_j=\frac{1}{n_j}\sum_{k=1}^{n_j}(X_k^j-m_j)(X_k^j-m_j)^T$$

转移前：

$$m_1=(4,3)^T，\ m_2=(1,2)^T$$

$$S_1=\begin{bmatrix}1 & -2\\ -2 & 4\end{bmatrix}，\ S_2=\begin{bmatrix}0 & 0\\ 0 & 4\end{bmatrix}$$

$$S_w=\sum_{j=1}^{2}P_jS_j=\frac{1}{2}S_1+\frac{1}{2}S_2=\begin{bmatrix}1/2 & -1\\ -1 & 4\end{bmatrix}$$

转移后：

$$m_1 = (3,5)^T,\ m_2 = (7/3, 5/3)^T$$

$$S_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},\ S_2 = \begin{bmatrix} 32/9 & -8/9 \\ -8/9 & 26/9 \end{bmatrix}$$

$$S_w = \sum_{j=1}^{2} P_j S_j = \frac{1}{4} S_1 + \frac{3}{4} S_2 = \begin{bmatrix} 8/3 & -2/3 \\ -2/3 & 13/6 \end{bmatrix}$$

（2）

优化目标：使S_w的行列式，即 $|S_w|$ 越来越小

转移前：

$$|S_w| = 1$$

转移后：

$$|S_w| = \frac{16}{3}$$

综上所述，转移不合适。

（3）

C-均值准则函数：误差平方和准则，越小越好

$$J_e = \sum_{j=1}^{c} \sum_{k=1}^{n_j} ||X_k^j - m_j||^2$$

转移前：

$$J_e = 18$$

转移后：

$$J_e = 19.33$$

综上所述，转移不合适。

4.有以下五个样本：

$$\boldsymbol{X_1} = (\mathbf{0},\mathbf{1},\mathbf{2},\mathbf{1},\mathbf{2},\mathbf{4})^{\boldsymbol{T}},$$
$$\boldsymbol{X_2} = (\mathbf{3},\mathbf{2},\mathbf{3},\mathbf{1},\mathbf{2},\mathbf{1})^{\boldsymbol{T}},$$
$$\boldsymbol{X_3} = (\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{1})^{\boldsymbol{T}},$$
$$\boldsymbol{X_4} = (\mathbf{2},\mathbf{1},\mathbf{0},\mathbf{2},\mathbf{1},\mathbf{2})^{\boldsymbol{T}},$$
$$\boldsymbol{X_5} = (\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{1},\mathbf{0})^{\boldsymbol{T}},$$

请按照最小距离准则用层次聚类算法进行聚类分析(矩阵 D 直接用根号表示)，并给出层次聚类示意图。

答：Too Simple! 在此省略具体计算过程。

$$\omega_1 = \{X_1\},\ \omega_2 = \{X_2\},\ \omega_3 = \{X_3\},\ \omega_4 = \{X_4\},\ \omega_5 = \{X_5\}$$
$$\downarrow$$
$$\omega_1 = \{X_1\},\ \omega_2 = \{X_2\},\ \omega_3 = \{X_3, X_5\},\ \omega_4 = \{X_4\}$$
$$\downarrow$$
$$\omega_1 = \{X_1\},\ \omega_2 = \{X_2\},\ \omega_3 = \{X_3, X_4, X_5\}$$
$$\downarrow$$
$$\omega_1 = \{X_1, X_3, X_4, X_5\},\ \omega_2 = \{X_2\}$$
$$\downarrow$$
$$\omega_1 = \{X_1, X_2, X_3, X_4, X_5\}$$

5. 已知一个样本集合：

$$\{(\mathbf{0},\mathbf{0})^T,\ (\mathbf{1},\mathbf{1})^T,\ (\mathbf{2},\mathbf{1})^T,\ (\mathbf{4},\mathbf{4})^T,\ (\mathbf{5},\mathbf{3})^T,\ (\mathbf{5},\mathbf{4})^T,\ (\mathbf{6},\mathbf{3})^T,\ (\mathbf{6},\mathbf{5})^T\}$$

试用近邻函数法进行聚类分析，请给出具体过程和最终的聚类结果。

答：

计算距离矩阵：

$$D=\begin{bmatrix} 0 & \sqrt{2} & \sqrt{5} & \sqrt{32} & \sqrt{34} & \sqrt{41} & \sqrt{45} & \sqrt{61} \\ \sqrt{2} & 0 & 1 & \sqrt{18} & \sqrt{20} & 5 & \sqrt{29} & \sqrt{41} \\ \sqrt{5} & 1 & 0 & \sqrt{13} & \sqrt{13} & \sqrt{18} & \sqrt{20} & \sqrt{32} \\ \sqrt{32} & \sqrt{18} & \sqrt{13} & 0 & \sqrt{2} & 1 & \sqrt{5} & \sqrt{5} \\ \sqrt{34} & \sqrt{20} & \sqrt{13} & \sqrt{2} & 0 & 1 & 1 & \sqrt{5} \\ \sqrt{41} & 5 & \sqrt{18} & 1 & 1 & 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{45} & \sqrt{29} & \sqrt{20} & \sqrt{5} & 1 & \sqrt{2} & 0 & 2 \\ \sqrt{61} & \sqrt{41} & \sqrt{32} & \sqrt{5} & \sqrt{5} & \sqrt{2} & 2 & 0 \end{bmatrix}$$

计算近邻系数矩阵：m_{ij}表示X_i对X_j的近邻系数，即进行列统计

$$M=\begin{bmatrix} 0 & 2 & 2 & 7 & 7 & 7 & 7 & 7 \\ 1 & 0 & 1 & 6 & 6 & 6 & 6 & 6 \\ 2 & 1 & 0 & 5 & 5 & 5 & 5 & 5 \\ 3 & 3 & 3 & 0 & 3 & 1 & 4 & 3 \\ 4 & 4 & 3 & 2 & 0 & 1 & 1 & 3 \\ 5 & 5 & 5 & 1 & 1 & 0 & 2 & 1 \\ 6 & 6 & 6 & 3 & 1 & 3 & 0 & 2 \\ 7 & 7 & 7 & 3 & 4 & 3 & 3 & 0 \end{bmatrix}$$

计算近邻函数值矩阵：$\alpha_{ij}=m_{ij}+m_{ji}-2$

$$A=\begin{bmatrix} 16 & 1 & 2 & 8 & 9 & 10 & 11 & 12 \\ 1 & 16 & 0 & 7 & 8 & 9 & 10 & 11 \\ 2 & 0 & 16 & 6 & 6 & 8 & 9 & 10 \\ 8 & 7 & 6 & 16 & 3 & 0 & 5 & 4 \\ 9 & 8 & 6 & 3 & 16 & 0 & 0 & 5 \\ 10 & 9 & 8 & 0 & 0 & 16 & 3 & 2 \\ 11 & 10 & 9 & 5 & 0 & 3 & 16 & 3 \\ 12 & 11 & 10 & 4 & 5 & 2 & 3 & 16 \end{bmatrix}$$

进行分类：

$$\omega_1=\{X_1,X_2,X_3\},\ \omega_2=\{X_4,X_5,X_6,X_7,X_8\}$$

进行判断：

$$\gamma_{12}=6,\ \max\gamma_1=1,\ \max\gamma_1=2$$

$$\gamma_{12}>\gamma_1,\ \gamma_{12}>\gamma_2$$

因此分类结束，分为两类。

6.查阅基于密度的聚类算法 DBSCAN，了解其相关概念定义，并给出算法的伪代码表达。

答：理解内涵，自由发挥。