Homework 5

1. Please give the specific process of the k-means clustering algorithm and analyze its advantages and disadvantages.

A: You will find it on https://en.wikipedia.org/wiki/K-means_clustering.

2.There are four samples:

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

In beginning, they are divided into two classes: ω_1: $\{X_1, X_2\}$ and ω_2: $\{X_3, X_4\}$

(1) If we move the sample X_2 to the class ω_2, please compute the within-class scatter matrix S_w.

(2) If we use the Determinant of S_w as Clustering criteria, please judge whether the movement in (1) is appropriate?

A：(1)

The within-class scatter matrix formula：

$$S_w = \sum_{j=1}^{c} P_j S_j$$

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

Before moving：

$$m_1 = (4,3)^T,\ m_2 = (1,2)^T$$

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

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

After moving：

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

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

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

（2）

Target：to reduce the value of Determinant of S_w (i.e. $|S_w|$)

Before moving：

$$|S_w| = 1$$

After moving：

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

Therefore, it is not appropriate.

3. There are five samples:

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

Please use the hierarchical clustering algorithm to cluster these samples under the minimum distance criterion (distance D can be directly represented by a root number), and give the hierarchical clustering process.

A: It is too simple to do it, what we need to do in each step is just combining two samples which are nearest with each other. Therefore, we do not show the process here, just give the result.

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

4. Consult the density-based clustering algorithm DBSCAN, and give the pseudo-code representation of the algorithm.

A: Just find it on the internet, and try to understand it. It is not the content of class.