1. 已知8个样本点 $\boldsymbol{x}_1 = (0, 0)^T$, $\boldsymbol{x}_2 = (2, 0)^T$, $\boldsymbol{x}_3 = (0, 2)^T$, $\boldsymbol{x}_4 = (2, 2)^T$, $\boldsymbol{x}_5 = (6, 6)^T$, $\boldsymbol{x}_6 = (8, 6)^T$, $\boldsymbol{x}_7 = (6, 8)^T$, $\boldsymbol{x}_8 = (8, 8)^T$, 利用K-均值算法将上述样本聚为2类，（1）要求选用样本点 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_5$ 分别作为两类中心的初始位置。（2）要求选用样本点 $\boldsymbol{x}_2$ 和 $\boldsymbol{x}_4$ 分别作为两类中心的初始位置。尝试分析初始点的选取对聚类效果的影响？

- 解：（1）以 x_1 和 x_5 作为两个聚类中心的初始位置：

初始聚类：$z_1 = (0,0)^T \quad z_2 = (6,6)^T$ ，

对所有样本进行聚类：

$$\|x_1 - z_1\| = 0 \Longrightarrow x_1 \in z_1$$

$$\|x_2 - z_2\| = \sqrt{52} > \|x_2 - z_1\| = 2 \Longrightarrow x_2 \in z_1$$

$$\|x_3 - z_2\| = \sqrt{52} > \|x_3 - z_1\| = 2 \Longrightarrow x_3 \in z_1$$

$$\|x_4 - z_2\| = \sqrt{32} > \|x_4 - z_1\| = \sqrt{8} \Longrightarrow x_4 \in z_1$$

$$\|x_5 - z_2\| = 0 \Longrightarrow x_5 \in z_2$$

$$\|x_6 - z_1\| = 10 > \|x_6 - z_2\| = 2 \Longrightarrow x_6 \in z_2$$

$$\|x_7 - z_1\| = 10 > \|x_7 - z_2\| = 2 \Longrightarrow x_7 \in z_2$$

$$\|x_8 - z_1\| = 8\sqrt{2} > \|x_8 - z_2\| = 2\sqrt{2} \Longrightarrow x_8 \in z_2$$

更新聚类中心：

$$z_1 = (x_1 + x_2 + x_3 + x_4)/4 = (1,1)^T, z_2 = (x_5 + x_6 + x_7 + x_8)/4 = (7,7)^T$$

根据新聚类中心，更新样本类别：

$$\|x_2 - z_2\| = 7\sqrt{2} > \|x_1 - z_1\| = \sqrt{2} \Rightarrow x_1 \in z_1$$

$$\|x_2 - z_2\| = \sqrt{74} > \|x_2 - z_1\| = \sqrt{2} \Rightarrow x_2 \in z_1$$

$$\|x_3 - z_2\| = \sqrt{74} > \|x_3 - z_1\| = \sqrt{2} \Rightarrow x_3 \in z_1$$

$$\|x_4 - z_2\| = 5\sqrt{2} > \|x_4 - z_1\| = \sqrt{2} \Rightarrow x_4 \in z_1$$

$$\|x_5 - z_1\| = 5\sqrt{2} > \|x_5 - z_2\| = \sqrt{2} \Rightarrow x_5 \in z_2$$

$$\|x_6 - z_1\| = \sqrt{74} > \|x_6 - z_2\| = \sqrt{2} \Rightarrow x_6 \in z_2$$

$$\|x_7 - z_1\| = \sqrt{74} > \|x_7 - z_2\| = \sqrt{2} \Rightarrow x_7 \in z_2$$

$$\|x_8 - z_1\| = 7\sqrt{2} > \|x_8 - z_2\| = \sqrt{2} \Rightarrow x_8 \in z_2$$

更新聚类中心：

$$z_1 = (x_1 + x_2 + x_3 + x_4)/4 = (1,1)^T, z_2 = (x_5 + x_6 + x_7 + x_8)/4 = (7,7)^T$$

第二次迭代，聚类中心无变换，迭代停止。

最终聚类结果：

第一类：$\{x_1,x_2,x_3,x_4\}$ ，第一类：$\{x_5,x_6,x_7,x_8\}$

聚类中心：

$$z_1 = (x_1 + x_2 + x_3 + x_4)/4 = (1,1)^T, z_2 = (x_5 + x_6 + x_7 + x_8)/4 = (7,7)^T$$

（2）以 x_2 和 x_4 作为两类中心的初始位置：

初始聚类：$z_1 = (2,0)^T \quad z_2 = (2,2)^T$，

对所有样本进行聚类：

$$\|x_1 - z_2\| = 2\sqrt{2} > \|x_1 - z_1\| = 2 \Longrightarrow x_1 \in z_1$$

$$\|x_2 - z_1\| = 0 \Longrightarrow x_2 \in z_1$$

$$\|x_3 - z_1\| = 2\sqrt{2} > \|x_3 - z_2\| = 2 \Longrightarrow x_3 \in z_2$$

$$\|x_4 - z_2\| = 0 \Longrightarrow x_4 \in z_2$$

$$\|x_5 - z_1\| = \sqrt{52} > \|x_5 - z_2\| = \sqrt{32} \Longrightarrow x_5 \in z_2$$

$$\|x_6 - z_1\| = \sqrt{72} > \|x_6 - z_2\| = \sqrt{52} \Longrightarrow x_6 \in z_2$$

$$\|x_7 - z_1\| = \sqrt{80} > \|x_7 - z_2\| = \sqrt{52} \Longrightarrow x_7 \in z_2$$

$$\|x_8 - z_1\| = 10 > \|x_8 - z_2\| = \sqrt{72} \Longrightarrow x_8 \in z_2$$

更新聚类中心：

$$z_1 = (x_1 + x_2)/2 = (1,0)^T, z_2 = (x_3 + x_4 + x_5 + x_6 + x_7 + x_8)/6 = (5, 16/3)^T$$

根据新聚类中心，更新样本类别：

$$\|x_2 - z_2\| = \sqrt{481}/3 > \|x_1 - z_1\| = 1 \Longrightarrow x_1 \in z_1$$
$$\|x_2 - z_2\| = \sqrt{337}/3 > \|x_2 - z_1\| = 1 \Longrightarrow x_2 \in z_1$$
$$\|x_3 - z_2\| = \sqrt{481}/3 > \|x_3 - z_1\| = \sqrt{2} \Longrightarrow x_3 \in z_1$$
$$\|x_4 - z_2\| = \sqrt{325}/3 > \|x_4 - z_1\| = \sqrt{5} \Longrightarrow x_4 \in z_1$$
$$\|x_5 - z_1\| = \sqrt{61} > \|x_5 - z_2\| = \sqrt{13}/3 \Longrightarrow x_5 \in z_2$$
$$\|x_6 - z_1\| = \sqrt{85} > \|x_6 - z_2\| = \sqrt{85}/3 \Longrightarrow x_6 \in z_2$$
$$\|x_7 - z_1\| = \sqrt{89} > \|x_7 - z_2\| = \sqrt{73}/3 \Longrightarrow x_7 \in z_2$$
$$\|x_8 - z_1\| = \sqrt{103} > \|x_8 - z_2\| = \sqrt{145}/3 \Longrightarrow x_8 \in z_2$$

更新聚类中心：

$$z_1 = (x_1 + x_2 + x_3 + x_4)/4 = (1,1)^T, z_2 = (x_5 + x_6 + x_7 + x_8)/4 = (7,7)^T$$

根据新聚类中心，更新样本类别：

$$\|x_2 - z_2\| = 7\sqrt{2} > \|x_1 - z_1\| = \sqrt{2} \Rightarrow x_1 \in z_1$$
$$\|x_2 - z_2\| = \sqrt{74} > \|x_2 - z_1\| = \sqrt{2} \Rightarrow x_2 \in z_1$$
$$\|x_3 - z_2\| = \sqrt{74} > \|x_3 - z_1\| = \sqrt{2} \Rightarrow x_3 \in z_1$$
$$\|x_4 - z_2\| = 5\sqrt{2} > \|x_4 - z_1\| = \sqrt{2} \Rightarrow x_4 \in z_1$$
$$\|x_5 - z_1\| = 5\sqrt{2} > \|x_5 - z_2\| = \sqrt{2} \Rightarrow x_5 \in z_2$$
$$\|x_6 - z_1\| = \sqrt{74} > \|x_6 - z_2\| = \sqrt{2} \Rightarrow x_6 \in z_2$$
$$\|x_7 - z_1\| = \sqrt{74} > \|x_7 - z_2\| = \sqrt{2} \Rightarrow x_7 \in z_2$$
$$\|x_8 - z_1\| = 7\sqrt{2} > \|x_8 - z_2\| = \sqrt{2} \Rightarrow x_8 \in z_2$$

更新聚类中心：

$$z_1 = (x_1 + x_2 + x_3 + x_4)/4 = (1,1)^T, z_2 = (x_5 + x_6 + x_7 + x_8)/4 = (7,7)^T$$

第三次迭代，聚类中心无变换，迭代停止。

最终聚类结果：

第一类：$\{x_1,x_2,x_3,x_4\}$ ，第一类：$\{x_5,x_6,x_7,x_8\}$

聚类中心：

$$z_1=(x_1+x_2+x_3+x_4)/4=(1,1)^T, z_2=(x_5+x_6+x_7+x_8)/4=(7,7)^T$$

初始聚类中心的选择对K均值算法的聚类结果有较大的影响，初始点选择可能影响到计算的复杂度，甚至影响聚类结果。

2. 已知5个样本，每个样本5个特征，数据如下：$x_1=(0,3,1,2,0)^T$ $x_2=(1,3,0,1,0)^T$，$x_3=(3,3,0,0,1)^T$，$x_4=(1,1,0,2,0)^T$，$x_5=(3,2,1,2,1)^T$，进行分级聚类，相似性度量采用最小距离准则，最终分为3类。并画出聚类分级树。

- 解：聚类过程：

  第1步：每个样本看作一类

  $$C_1^0=\{x_1\}, C_2^0=\{x_2\}, C_3^0=\{x_3\}, C_4^0=\{x_4\}, C_5^0=\{x_5\}$$

  第2步：计算类间欧式距离，完成第一级聚类

	C_1^0	C_2^0	C_3^0	C_4^0	C_5^0
C_1^0	--	$\sqrt{3}$	$\sqrt{15}$	$\sqrt{6}$	$\sqrt{11}$
C_2^0		--	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{8}$
C_3^0			--	$\sqrt{13}$	$\sqrt{6}$
C_4^0				--	$\sqrt{7}$
C_5^0					--

合并最小距离，第一级聚类结果：

$$C_1^1 = \{x_1, x_2\}, C_2^1 = \{x_3\}, C_3^1 = \{x_4\}, C_4^1 = \{x_5\}$$

第3步：重复第2步过程，完成第二级聚类

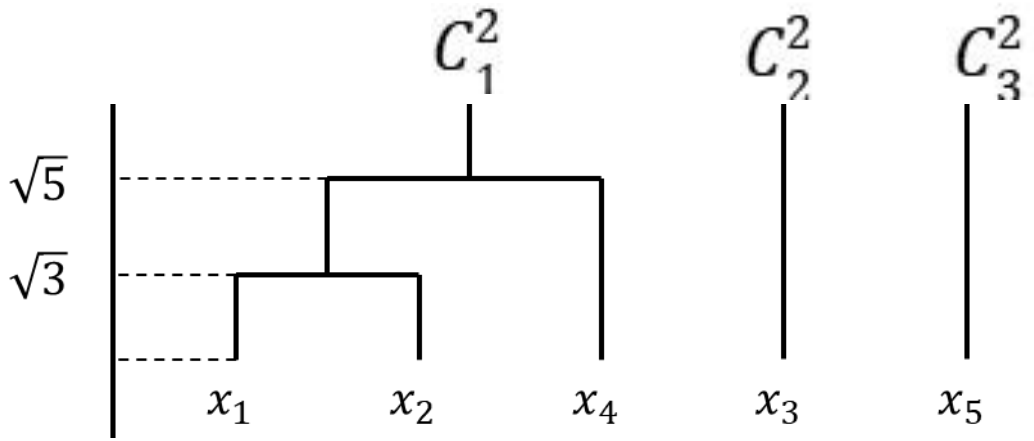
	C_1^1	C_2^1	C_3^1	C_4^1
C_1^1	--	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{8}$
C_2^1		--	$\sqrt{13}$	$\sqrt{6}$
C_3^1			--	$\sqrt{7}$
C_4^1				--

合并最小距离，第二级聚类结果：

$$C_1^2 = \{x_1, x_2, x_4\}, C_2^2 = \{x_3\}, C_3^2 = \{x_5\}$$

聚类数目达到3类，停止合并。

最终聚为三类的分级树：



3. 设有5个四维模式，按最小距离准则和 Tanimoto 测度进行系统分级聚类分析。

$$x_1 = \{1\ 0\ 1\ 0\}$$
$$x_2 = \{0\ 1\ 0\ 1\}$$
$$x_3 = \{0\ 1\ 0\ 0\}$$
$$x_4 = \{0\ 0\ 0\ 0\}$$
$$x_5 = \{1\ 0\ 0\ 0\}$$

解：采用Tanimoto测度

$$S(x,y) = \frac{x^T y}{x^T x + y^T y - x^T y}$$

第1步：每个样本看作一类

$$C_1^0 = \{x_1\}, C_2^0 = \{x_2\}, C_3^0 = \{x_3\}, C_4^0 = \{x_4\}, C_5^0 = \{x_5\}$$

第2步：计算Tanimoto测度，完成第一级聚类

	C_1^0	C_2^0	C_3^0	C_4^0	C_5^0
C_1^0	--	0	0	0	1/2
C_2^0		--	1/2	0	0
C_3^0			--	0	0
C_4^0				--	0
C_5^0					--

合并相似度最高的类，第一级聚类结果：

$$C_1^1 = \{x_1, x_5\}, C_2^1 = \{x_2, x_3\}, C_3^1 = \{x_4\}$$

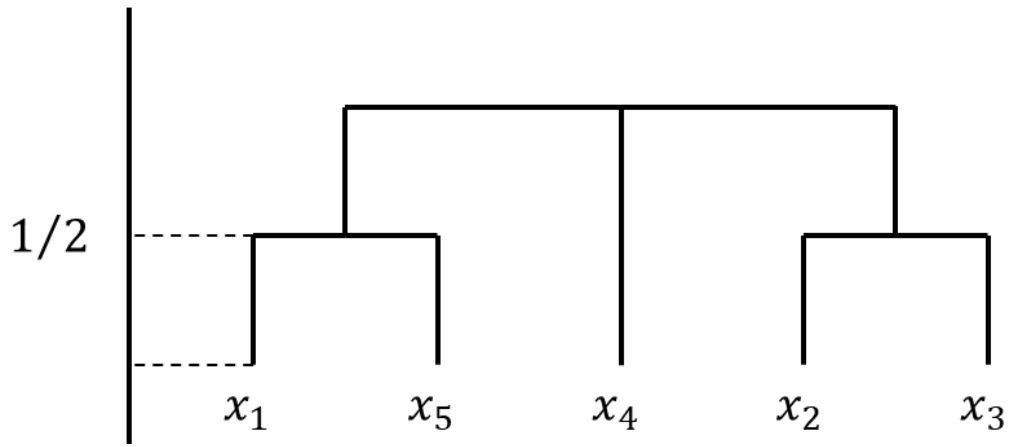
第3步：计算Tanimoto测度，完成第二级聚类

	C_1^1	C_2^1	C_3^1
C_1^1	--	0	0
C_2^1		--	0
C_3^1			--

第二级聚类结果：

$$C_1^2 = \{x_1, x_2, x_3, x_4, x_5\}$$

最终聚为一类的分级树：



4. 给定一组数据

$$x_1=\begin{bmatrix}0\\0\end{bmatrix}\quad x_2=\begin{bmatrix}-1\\0\end{bmatrix}\quad x_3=\begin{bmatrix}1\\-1\end{bmatrix}\quad x_4=\begin{bmatrix}0\\-1\end{bmatrix}\quad x_5=\begin{bmatrix}-2\\0\end{bmatrix}$$

$$x_6=\begin{bmatrix}4\\-5\end{bmatrix}\quad x_7=\begin{bmatrix}3\\-5\end{bmatrix}\quad x_8=\begin{bmatrix}4\\-4\end{bmatrix}\quad x_9=\begin{bmatrix}3\\-4\end{bmatrix}\quad x_{10}=\begin{bmatrix}4\\5\end{bmatrix}$$

采用欧氏距离，设置距离阈值为3，分别求取各数据点的局部密度值 ρ_i 及各点与更高密度值数据的距离 δ_i, 利用基于密度峰值的聚类方法确定这些数据的聚类中心数并实现聚类。

解：第一步：分别求取每一个样本与其他样本的欧式距离，并计算 ρ_i 值。

对于第一个样本 $x_1 = [0,\ 0]^T$

$$||x_1 - x_1|| = \sqrt{(0-0)^2 + (0-0)^2} = 0 < d = 3 \Rightarrow \chi(0-3) = 1$$

$$||x_1 - x_2|| = \sqrt{(0-(-1))^2 + (0-0)^2} = 1 < d = 3 \Rightarrow \chi(1-3) = 1$$

$$||x_1 - x_3|| = \sqrt{(0-1)^2 + (0-(-1))^2} = \sqrt{2} < d = 3 \Rightarrow \chi(2-3) = 1$$

.....

$$||x_1 - x_{10}|| = \sqrt{(0-4)^2 + (0-5)^2} = \sqrt{41} < d = 3 \Rightarrow \chi\left(\sqrt{41}-3\right) = 0$$

$\rho_i = \sum_j \chi(d_{ij} - d_c)$ 于是有：

$$\rho_1 = 1+1+1+1+1+0+0+0+0+0 = 5$$

类似地，求取剩余样本的密度值，得：

$$\rho = \{5,5,4,5,4,4,4,4,4,1\}$$

对应的距离为：

0	1	$\sqrt{2}$	1	2	$\sqrt{41}$	$\sqrt{34}$	$4\sqrt{2}$	5	$\sqrt{41}$
1	0	$\sqrt{5}$	$\sqrt{2}$	1	$5\sqrt{2}$	$\sqrt{41}$	$\sqrt{41}$	$4\sqrt{2}$	$5\sqrt{2}$
$\sqrt{2}$	$\sqrt{5}$	0	1	$\sqrt{10}$	5	$2\sqrt{5}$	$3\sqrt{2}$	$\sqrt{13}$	5
1	$\sqrt{2}$	1	0	$\sqrt{5}$	$4\sqrt{2}$	5	5	$3\sqrt{2}$	$\sqrt{52}$
2	1	$\sqrt{10}$	$\sqrt{5}$	0	$\sqrt{61}$	$5\sqrt{2}$	$\sqrt{52}$	$\sqrt{41}$	$\sqrt{61}$
$\sqrt{41}$	$5\sqrt{2}$	5	$4\sqrt{2}$	$\sqrt{61}$	0	1	1	$\sqrt{2}$	10
$\sqrt{34}$	$\sqrt{41}$	$2\sqrt{5}$	5	$5\sqrt{2}$	1	0	$\sqrt{2}$	1	$\sqrt{101}$
$4\sqrt{2}$	$\sqrt{41}$	$3\sqrt{2}$	5	$\sqrt{52}$	1	$\sqrt{2}$	0	1	9
5	$4\sqrt{2}$	$\sqrt{13}$	$3\sqrt{2}$	$\sqrt{41}$	$\sqrt{2}$	1	1	0	$\sqrt{82}$
$\sqrt{41}$	$5\sqrt{2}$	$3\sqrt{5}$	$\sqrt{52}$	$\sqrt{61}$	10	$\sqrt{101}$	9	$\sqrt{82}$	0

第二步：根据 ρ 值计算 δ 值

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

对于第一个样本$\rho_1 = 5$，$\delta_1 = \sqrt{41}$

对于第二个样本$\rho_2 = 5$，$\delta_2 = 5\sqrt{2}$

对于第三个样本$\rho_3 = 4$，密度大于 4 的样本有：$\{x_1, x_2, x_4\}$，从上表查对应的距离为：$\{\sqrt{2}, \sqrt{5}, 1\}$，最小距离为 1，于是：$\delta_3 = 1$

类似地求取每一个样本的δ值，有：

$$\delta_i = \{\sqrt{41}, 5\sqrt{2}, 1, \sqrt{52}, 1, 4\sqrt{2}, 5, 5, 3\sqrt{2}, \sqrt{41}\}$$

第三步：基于$\gamma_i = \rho_i \times \delta_i$，

得到$\gamma = \{5\sqrt{41}, 25\sqrt{2}, 4, 10\sqrt{13}, 4, 16\sqrt{2}, 20, 20, 12\sqrt{2}, \sqrt{41}\}$

$\gamma = \{\sqrt{1025}, \sqrt{1250}, 4, \sqrt{1300}, 4, \sqrt{512}, 20, 20, \sqrt{288}, \sqrt{41}\}$

对 γ 按照降序排序：

$$\gamma_{\text{rank}}=\{\sqrt{1300},\sqrt{1250},\sqrt{1025},\sqrt{512},20,20,\sqrt{288},\sqrt{41},4,4\}$$

出现频次为1　　出现频次为2

从20开始，γ_{rank} 频次高于之前排序的数值，因此 $\gamma \leq 20$的样本均视为非聚类中心点。

$\sqrt{1300}$、$\sqrt{1250}$、$\sqrt{1025}$和$\sqrt{512}$共 4 类，对应的类别中心分别是：

$$\{x_4,x_2,x_1,x_6\}$$

依据初始聚类中心间的距离，根据距离阈值对聚类中心进行合并，得到：

$$\{\{x_1,x_2,x_4\},\{x_6\}\}$$

第四步，根据阈值对剩余样本进行分级聚类，对应的聚类结果为2类：

$$\{\{x_1,x_2,x_3,x_4,x_5\},\{x_6,x_7,x_8,x_9\}\}$$

未被分类的样本为：

$$\{x_{10}\}=\{[4,5]^T\}$$