作业 4：聚类

1. 请从应用场景、模型学习过程等方面阐述聚类与分类的区别。

   答：聚类是一种无监督学习，基于数据集中样本之间的相似性对样本进行分组，使得同一个组/簇内样本之间相似性大于不同组/簇样本之间的相似性，一般在不知道类别信息的情况下（不知道数据集属于哪些类别），用作为对数据集的初步探索和基本处理。而分类是监督学习，需要足够多的标签数据作为监督信息来训练模型，使其能够较准确的预测未知样本的类别标签。分类用于已知数据集所相关的类别以及有足够多的标记样本的情况下。

2. 对以下数据集进行标准的 k 均值聚类（$d_{ik} = \|x_i - \mu_k\|^2$）。假设 k=2，两个均值$\mu_1$和$\mu_2$分别初始化为第 2、4 个样本，即$\mu_1 = [2，2，1]$，$\mu_2 = [1，0，2]$，给出迭代过程以及收敛时的均值和簇。

序号	属性 1	属性 2	属性 3
1	1	2	1
2	2	2	1
3	0	1	2
4	1	0	2

答：基于初始均值，计算每个样本到两个均值的距离分别为：

$$d_{11} = (1-2)^2 + (2-2)^2 + (1-1)^2 = 1$$
$$d_{12} = (1-1)^2 + (2-0)^2 + (1-2)^2 = 5$$

$argmin_f\ d_{1f} = 1$，*所以第 1 个样本被分到第 1 个类；*

$$d_{21} = (2-2)^2 + (2-2)^2 + (1-1)^2 = 0$$
$$d_{22} = (2-1)^2 + (2-0)^2 + (1-2)^2 = 6$$

$argmin_f\ d_{2f} = 1$，*所以第 2 个样本被分到第 1 个类；*

$$d_{31} = (0-2)^2 + (1-2)^2 + (2-1)^2 = 6$$
$$d_{32} = (0-1)^2 + (1-0)^2 + (2-2)^2 = 2$$

$argmin_f\ d_{3f} = 2$，*所以第 3 个样本被分到第 2 个类；*

$$d_{41} = (1-2)^2 + (0-2)^2 + (2-1)^2 = 6$$
$$d_{42} = (1-1)^2 + (0-0)^2 + (2-2)^2 = 0$$

$argmin_f\ d_{4f} = 1$，*所以第 4 个样本被分到第 2 个类；*

第一次得到划分：C1={1，2}，C2={3，4}

第一次更新后的均值为：

$$\mu_1 = \frac{1}{2}[(1+2),(2+2),(1+1)] = [1.5, 2, 1]$$

$$\mu_2 = \frac{1}{2}[(0+1),(1+0),(2+2)] = [0.5, 0.5, 2]$$

计算每个样本到当前均值的距离分别为：

$$d_{11} = (1-1.5)^2 + (2-2)^2 + (1-1)^2 = 0.25$$
$$d_{12} = (1-0.5)^2 + (2-0.5)^2 + (1-2)^2 = 3.5$$

$argmin_f\ d_{1f} = 1$，*所以第 1 个样本被分到第 1 个类；*

$$d_{21} = (1-1.5)^2 + (2-2)^2 + (1-1)^2 = 0.25$$
$$d_{22} = (2-0.5)^2 + (2-0.5)^2 + (1-2)^2 = 5.5$$

$argmin_f\ d_{2f} = 1$，*所以第 2 个样本被分到第 1 个类；*

$$d_{31} = (0-1.5)^2 + (1-2)^2 + (2-1)^2 = 4.25$$
$$d_{32} = (0-0.5)^2 + (1-0.5)^2 + (2-2)^2 = 0.5$$

$argmin_f\ d_{3f} = 2$，*所以第 3 个样本被分到第 2 个类；*

$$d_{41} = (1-1.5)^2 + (0-2)^2 + (2-1)^2 = 5.25$$
$$d_{42} = (1-0.5)^2 + (0-0.5)^2 + (2-2)^2 = 0.5$$

$argmin_f\ d_{4f} = 1$，*所以第 4 个样本被分到第 2 个类；*

第二次划分：C1={1，2}，C2={3，4}，与之前一样，说明已经收敛。

最后的均值为：

$$\mu_1 = [1.5, 2, 1]\ ,\quad \mu_2 = [0.5, 0.5, 2]$$

3. 对以上数据集进行基于平均连接（average-linkage）的凝聚层次聚类。

提示：先计算两两样本之间的距离矩阵（欧式距离），给出每一步更新后的簇间距离矩阵。

答：先计算两两样本之间的（欧式）距离矩阵：

$$R = \begin{bmatrix} & 1.00 & 1.73 & 2.24 \\ 1.00 & & 2.45 & 2.45 \\ 1.73 & 2.45 & & 1.41 \\ 2.24 & 2.45 & 1.41 & \end{bmatrix}$$

初始化每个样本为一个簇：$C_1 = \{x_1\}$，$C_2 = \{x_2\}$，$C_3 = \{x_3\}$，$C_4 = \{x_4\}$

初始化簇间距离D等于R。

从D中找出平均距离最小的两个簇：C_1和C_2进行合并，得到新的划分：$C_1 = \{x_1, x_2\}$，

$C_2 = \{x_3\}$，$C_3 = \{x_4\}$，用平均距离更新D得

$$D = \begin{bmatrix} & 2.09 & 2.35 \\ 2.09 & & 1.41 \\ 2.35 & 1.41 & \end{bmatrix}$$

其中$d_{12} = \frac{1}{2}(r_{13} + r_{23}) = 2.09$，$d_{13} = \frac{1}{2}(r_{14} + r_{24}) = 2.35$，$d_{23} = r_{34}$

从D中找出平均距离最小的两个簇：C_2和C_3进行合并，得到新的划分$C_1 = \{x_1$，

$x_2\}$，$C_2 = \{x_3, x_4\}$，用平均距离更新D得

$$D = \begin{bmatrix} & 2.22 \\ 2.22 & \end{bmatrix}$$

其中$d_{12} = \frac{1}{4}(r_{13} + r_{14} + r_{23} + r_{24}) = 2.22$

最后把C_1和C_2 *合并*得到$C_1 = \{x_1，\ x_2，\ x_3，\ x_4\}$

基于以上过程得到以下树状结构：