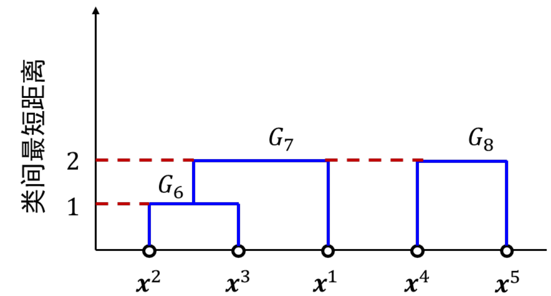
小作业5

一、给定含有5个样本的集合：$X=\begin{bmatrix}0&0&1&5&5\\2&0&0&0&2\end{bmatrix}$

1、应用聚合层次聚类法对这5个样本进行聚类，并在得到两个类时停止；

计算样本之间的欧氏距离矩阵D，

$$D=\left[d_{ij}\right]_{5\times5}=\begin{bmatrix}0&2&\sqrt{5}&\sqrt{29}&5\\2&0&1&5&\sqrt{29}\\\sqrt{5}&1&0&4&2\sqrt{5}\\\sqrt{29}&5&4&0&2\\5&\sqrt{29}&2\sqrt{5}&2&0\end{bmatrix}$$



用5个样本构建5个类，$G_i=\{x^i\}, i=1,2,\ldots,5$。以最短距离为类间距离，则5个类之间的距离矩阵亦为D。可以看出，$D_{23}=D_{32}=1$为最小。把G_2和G_3合并为一个新类，记作$G_6=\{x^2,x^3\}$。计算类间距离矩阵D'，

$$D'=\left[d'_{ij}\right]_{4\times4}=\begin{bmatrix}0&\sqrt{29}&5&2\\\sqrt{29}&0&2&4\\5&2&0&2\sqrt{5}\\2&4&2\sqrt{5}&0\end{bmatrix}$$

可以看出，$D'_{45}=D'_{54}=D'_{16}=D'_{61}=2$为最小。把$G_1$和$G_6$合并为一个新类，记作$G_7=\{x^1,x^2,x^3\}$，同时把$G_4$和$G_5$合并为一个新类，记作$G_8=\{x^4,x^5\}$。

全部样本被聚成两类$\{x^1,x^2,x^3\}$和$\{x^4,x^5\}$，达到终止条件，聚类终止。

小作业5

2、对于上面得到的两个类，从每个类中选取一个与类中心距离最近的点，作为初始类中心，应用k均值聚类算法，将5个样本聚到两个类中；

计算G_7的类中心$\bar{\boldsymbol{x}}_{G_7}=\left(\frac{1}{3},\frac{2}{3}\right)$，$G_8$的类中心$\bar{\boldsymbol{x}}_{G_8}=(5,1)$，在两个类中选择距离它们最近的样本点作为k均值聚类的初始类中心，即：$\bar{\boldsymbol{x}}_{C_1}^{(0)}=\boldsymbol{x}^2=(0,0)$，$\bar{\boldsymbol{x}}_{C_2}^{(0)}=\boldsymbol{x}^4=(5,0)$（选择$\boldsymbol{x}^5=(5,2)$也正确），分别计算各样本与类中心$\bar{\boldsymbol{x}}_{C_1}^{(0)}$、$\bar{\boldsymbol{x}}_{C_2}^{(0)}$的欧氏距离平方。

对 $\boldsymbol{x}^1$，$d\left(\boldsymbol{x}^1,\bar{\boldsymbol{x}}_{C_1}^{(0)}\right)=4$，$d\left(\boldsymbol{x}^1,\bar{\boldsymbol{x}}_{C_2}^{(0)}\right)=29$，将 $\boldsymbol{x}^1$ 分到类 $C_1^{(0)}$；

对 $\boldsymbol{x}^3$，$d\left(\boldsymbol{x}^3,\bar{\boldsymbol{x}}_{C_1}^{(0)}\right)=1$，$d\left(\boldsymbol{x}^3,\bar{\boldsymbol{x}}_{C_2}^{(0)}\right)=16$，将 $\boldsymbol{x}^3$ 分到类 $C_1^{(0)}$；

对 $\boldsymbol{x}^5$，$d\left(\boldsymbol{x}^5,\bar{\boldsymbol{x}}_{C_1}^{(0)}\right)=29$，$d\left(\boldsymbol{x}^5,\bar{\boldsymbol{x}}_{C_2}^{(0)}\right)=4$，将 $\boldsymbol{x}^5$ 分到类 $C_2^{(0)}$。

由第一次聚类得到的 $C_1^{(0)}=\{\boldsymbol{x}^1,\boldsymbol{x}^2,\boldsymbol{x}^3\}$，$C_2^{(0)}=\{\boldsymbol{x}^4,\boldsymbol{x}^5\}$ 计算新的类中心 $\bar{\boldsymbol{x}}_{C_1}^{(1)}=\left(\frac{1}{3},\frac{2}{3}\right)$，$\bar{\boldsymbol{x}}_{C_2}^{(1)}=(5,1)$。分别计算各样本与类中心$\bar{\boldsymbol{x}}_{C_1}^{(1)}$、$\bar{\boldsymbol{x}}_{C_2}^{(1)}$的欧氏距离平方。（注意：前后两次类中心有变化）

对 $\boldsymbol{x}^1$，$d\left(\boldsymbol{x}^1,\bar{\boldsymbol{x}}_{C_1}^{(1)}\right)=\frac{17}{9}$，$d\left(\boldsymbol{x}^1,\bar{\boldsymbol{x}}_{C_2}^{(1)}\right)=26$，将 $\boldsymbol{x}^1$ 分到类 $C_1^{(1)}$；

对 $\boldsymbol{x}^2$，$d\left(\boldsymbol{x}^2,\bar{\boldsymbol{x}}_{C_1}^{(1)}\right)=\frac{5}{9}$，$d\left(\boldsymbol{x}^2,\bar{\boldsymbol{x}}_{C_2}^{(1)}\right)=26$，将 $\boldsymbol{x}^2$ 分到类 $C_1^{(1)}$；

对 $\boldsymbol{x}^3$，$d\left(\boldsymbol{x}^3,\bar{\boldsymbol{x}}_{C_1}^{(1)}\right)=\frac{8}{9}$，$d\left(\boldsymbol{x}^3,\bar{\boldsymbol{x}}_{C_2}^{(1)}\right)=17$，将 $\boldsymbol{x}^3$ 分到类 $C_1^{(1)}$；

对 $\boldsymbol{x}^4$，$d\left(\boldsymbol{x}^4,\bar{\boldsymbol{x}}_{C_1}^{(1)}\right)=\frac{200}{9}$，$d\left(\boldsymbol{x}^4,\bar{\boldsymbol{x}}_{C_2}^{(1)}\right)=1$，将 $\boldsymbol{x}^4$ 分到类 $C_2^{(1)}$；

对 $\boldsymbol{x}^5$，$d\left(\boldsymbol{x}^5,\bar{\boldsymbol{x}}_{C_1}^{(1)}\right)=\frac{212}{9}$，$d\left(\boldsymbol{x}^5,\bar{\boldsymbol{x}}_{C_2}^{(1)}\right)=1$，将 $\boldsymbol{x}^5$ 分到类 $C_2^{(1)}$。

得到新的类$C_1^{(1)}=\{\boldsymbol{x}^1,\boldsymbol{x}^2,\boldsymbol{x}^3\}$，$C_2^{(1)}=\{\boldsymbol{x}^4,\boldsymbol{x}^5\}$。由于得到的新的类没有改变，聚类停止。

得到聚类结果：$C_1^*=\{\boldsymbol{x}^1,\boldsymbol{x}^2,\boldsymbol{x}^3\}$，$C_2^*=\{\boldsymbol{x}^4,\boldsymbol{x}^5\}$。

小作业5

3、与讲义中的k均值聚类算法示例比较，试从损失函数的角度，谈谈你对初始类中心选择的看法。

一、选择不同初始类中心可以得到不同的k均值聚类结果

定义损失函数为样本与其所属类中心之间距离的平方和，即$E(\mathcal{C})=\sum_{l=1}^{K}\sum_{\mathcal{C}(\boldsymbol{x}^i)=l}\left\|\boldsymbol{x}^i-\overline{\boldsymbol{x}}_{C_l}\right\|^2$，其中 $\overline{\boldsymbol{x}}_{C_l}$ 为第 l 个类 C_l 的均值或中心。那么，可以计算两种初始类中心得到的两种聚类结果的损失函数值：

当选择$\overline{\boldsymbol{x}}_{C_1}^{(0)}=\boldsymbol{x}^1=(0,2)$，$\overline{\boldsymbol{x}}_{C_2}^{(0)}=\boldsymbol{x}^2=(0,0)$作为初始类中心，最终得到$\mathcal{C}_1$：$C_1^*=\{\boldsymbol{x}^1,\boldsymbol{x}^5\}$，$C_2^*=\{\boldsymbol{x}^2,\boldsymbol{x}^3,\boldsymbol{x}^4\}$。则$E(\mathcal{C}_1)=\left(2.5^2+2.5^2\right)+(4+1+9)=26.5$。

选择$\overline{\boldsymbol{x}}_{C_1}^{(0)}=\boldsymbol{x}^2=(0,0)$，$\overline{\boldsymbol{x}}_{C_2}^{(0)}=\boldsymbol{x}^4=(5,0)$作为初始类中心，最终得到$\mathcal{C}_2$：$C_1^*=\{\boldsymbol{x}^1,\boldsymbol{x}^2,\boldsymbol{x}^3\}$，$C_2^*=\{\boldsymbol{x}^4,\boldsymbol{x}^5\}$。则$E(\mathcal{C}_2)=\left(\frac{17}{9}+\frac{5}{9}+\frac{8}{9}\right)+(1+1)=\frac{16}{3}<E(\mathcal{C}_1)$。

（计算损失函数时取2-范数也对）

可见，选择第二种初始类中心得到的聚类结果的损失函数更小，应采用这种划分结果。

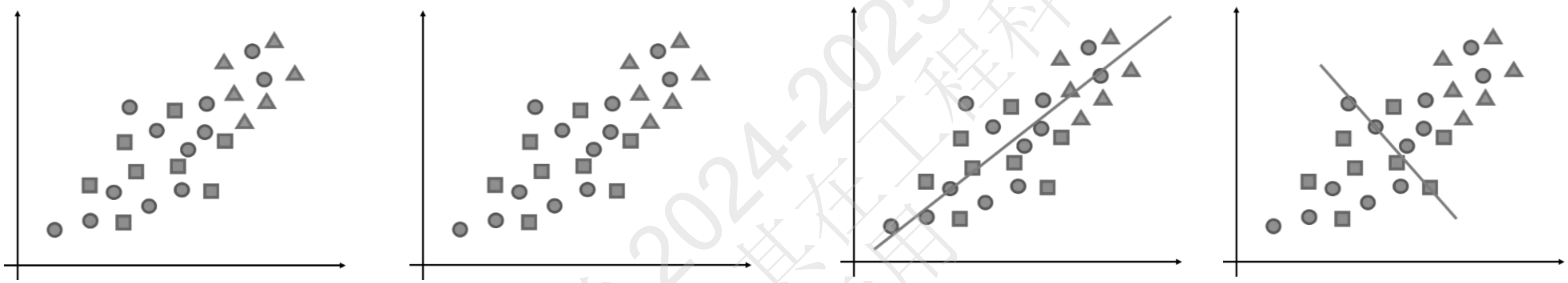
二、初始类中心的选择

在k均值分类中，将样本先经过层次聚类，得到 k 类时停止，然后从每个类中选取一个与类中心距离最近的点作为初始类中心，可以提升k均值聚类的效果。如果同时有多组初始类中心都是最近的，那么，可以尝试多次，选择使损失函数最小的划分为聚类结果。

小作业5

二、回答下列有关PCA和LDA的问题：

1、在左图中画出第一主成分方向，在右图中画出LDA线性判别方向。对线性判别，认为圆形样本代表正例，三角形和正方形样本代表负例。不必给出计算过程，只需画出方向。



a）PCA第一主成分方向　b）LDA线性判别方向

分析：PCA不使用样本的标记（忽略样本是圆形、三角形还是正方形），并要求低维子空间对样本具有最大可分性或最近重构性；LDA 是选择一个最佳的投影方向，使得投影后相同类别的数据分布紧凑，不同类别的数据尽量相互远离，在图b）所示方向上，正例类（圆形）和负例类（三角形和正方形）的类内散度最小。

评论：在该例中，从分类角度来看，PCA和LDA的效果都不好；但总体上，降维仍是机器学习中一种重要的数据预处理技术。

小作业5

2、考虑三个样本的集合：$\boldsymbol{X}=\begin{bmatrix}-1 & 0 & 1\\ 1 & 0 & -1\end{bmatrix}$

I. 确定该样本集的第一、第二主成分方向（写出方向向量）；

II. 将样本投影到第一主成分方向上，求样本的新坐标。

I. 对所有样本进行**中心化**后：$\boldsymbol{x}^1=(-1,\ 1)^{\mathrm{T}}$，$\boldsymbol{x}^2=(0,\ 0)^{\mathrm{T}}$，$\boldsymbol{x}^3=(1,-1)^{\mathrm{T}}$，计算样本的协方差矩阵（忽略常数项的）：$\boldsymbol{X}\boldsymbol{X}^T=\begin{bmatrix}2 & -2\\ -2 & 2\end{bmatrix}$，对$\boldsymbol{X}\boldsymbol{X}^T$做特征值分解，得$\lambda_1=4$，$\lambda_2=0$，对应的特征向量分别为$\boldsymbol{w}_1=(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})^{\mathrm{T}}$，$\boldsymbol{w}_2=(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})^{\mathrm{T}}$。则第一主成分方向为$(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})^{\mathrm{T}}$，第二主成分方向为$(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})^{\mathrm{T}}$。（注意，$\boldsymbol{w}_1=(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})^{\mathrm{T}}$，$\boldsymbol{w}_2=(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})^{\mathrm{T}}$，或者是其他平行向量，也正确）

II. 选取$\boldsymbol{w}_1=(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})^{\mathrm{T}}$为低维子空间的基向量，样本在第一主成分方向上的新坐标$\boldsymbol{z}=\boldsymbol{w}_1{}^{\mathrm{T}}\boldsymbol{X}=[-\sqrt{2}\quad 0\quad \sqrt{2}]$。（注意，一定要用单位向量，即标准化的主方向向量构成的转换矩阵求新坐标，否则，计算出来的坐标是经过缩放的子空间坐标，而非投影后的坐标；$\boldsymbol{z}=[\sqrt{2}\quad 0\quad -\sqrt{2}]$ 也正确）