第八章作业.

1. kMeans: A_2 B_2 C_2 D_1 E_1 F_2.

对于A.F.应选用DBSCAN, B_1 C_1 D_2 E_2应该是GMM降维结果.

2.

$$X-\mu_1=\begin{bmatrix}-0.3 & 0\\ -1.6 & -0.3\\ 0 & -0.4\\ -1.5 & 0\\ -0.7 & 1\\ -1.2 & -0.2\\ -1.3 & -0.1\\ 0.5 & -0.1\\ -1.1 & 0.6\\ -0.2 & -0.2\end{bmatrix}\quad d_{x\to\mu_1}=\begin{bmatrix}0.3\\ 1.628\\ 0.400\\ 1.500\\ 1.221\\ 1.217\\ 1.304\\ 0.510\\ 1.253\\ 0.283\end{bmatrix}$$

$$X-\mu_2=\begin{bmatrix}-0.7 & -0.5\\ -2 & -0.8\\ -0.4 & -0.9\\ -1.9 & -0.5\\ -1.1 & 0.5\\ -1.6 & -0.7\\ -1.7 & -0.6\\ 0.1 & -0.6\\ -1.5 & 0.1\\ -0.6 & -0.7\end{bmatrix}\quad X-\mu_3=\begin{bmatrix}-0.6 & 0.2\\ -1.9 & 0.1\\ -0.3 & -0.2\\ -1.8 & 0.2\\ -1 & 1.2\\ -1.5 & 0.\\ -1.6 & 0.1\\ 0.2 & 0.1\\ -1.4 & 0.8\\ -0.5 & 0\end{bmatrix}.$$

$$D_{x\to\mu}=\begin{array}{c}\begin{matrix}\quad R & \quad G & \quad B\end{matrix}\\ \begin{bmatrix}\boxed{0.3} & 0.860 & 0.632\\ \boxed{1.628} & 2.154 & 1.903\\ 0.400 & 0.985 & \boxed{0.361}\\ \boxed{1.5} & 1.965 & 1.811\\ 1.221 & \boxed{1.208} & 1.562\\ \boxed{1.217} & 1.746 & 1.5\\ \boxed{1.504} & 1.803 & 1.603\\ 0.510 & 0.608 & \boxed{0.224}\\ \boxed{1.253} & 1.503 & 1.612\\ \boxed{0.283} & 0.922 & 0.5\end{bmatrix}\end{array}$$

各最小值如红框所示.

则 Red类有: $\begin{bmatrix}5.9 & 3.2\\ 4.6 & 2.9\\ 4.7 & 3.2\\ 5.0 & 3.0\\ 4.9 & 3.1\\ 5.1 & 3.8\\ 6.0 & 3.0\end{bmatrix}$

$$G=\begin{bmatrix}5.5 & 4.2\end{bmatrix}\quad B=\begin{bmatrix}6.2 & 2.8\\ 6.7 & 3.1\end{bmatrix}$$

$\mu_1^{(1)}=[5.3,\ 3.175]^T$(红色) $\mu_2^{(1)}=[6.05,\ 3.95]^T$(绿色) $\mu_3^{(1)}=[6.47,\ 2.97]^T$(蓝色)

3. 1. 球形簇假设
2. 一定会收敛
3. 可使用随机选择聚类中心, 且可将中心点得尽量远些.
4. 采用肘部法, 获取 Loss-k 图像, 寻找出率最大处时互.
5. 只能聚类球形簇, 对离群点敏感, 并非常依赖初值选择.
   改进: 采用更好的初值选取方法, 例 kMeans++;
6. 用z表示样本属于哪个高斯分布.
7. M步最大化Q函数与kMeans目标式相同.
   都需要预先确定k值.
8. 外部指标和内部指标. 例如轮廓系数等. JC系数. FMI、RI兰、DBI等.
   都基本思想都是簇间相似度尽量高低. 簇内相似度尽量高.
   外部指标依赖参考模型, 内部指标不需要.

9. 一旦分裂、无法后退或合并进行重新修正.
不具备很好的伸缩性

10. PCA流程：① 去均值化：取样本均值并减掉：$x-\bar{x}$
② 计算协方差矩阵及其特征值与特征向量.
③ 对特征值大小作排序.
④ 选择特征值最大前k个对应特征向量组成映射矩阵
⑤ 对数据进行映射.