第四次作业答案

1、相同：均需要提前设定聚类数和终止条件，并都以样本之间的距离作为分类的依据；

不同：K-means 需要提现指定初始的聚类中心，而层次聚类不用。

2、

（1）各样本距离各初始中心的距离如下：

样本	Center1	Center2	Center3	归属簇
1	0	5	8.062258	1
2	5	0	3.162278	2
3	8.485281	6.082763	7.28011	2
4	3.605551	4.242641	7.211103	1
5	7.071068	5	6.708204	2
6	7.211103	4.123106	5.385165	2
7	8.062258	3.162278	0	3
8	2.236068	4.472136	7.615773	1

第一次聚类结果为：

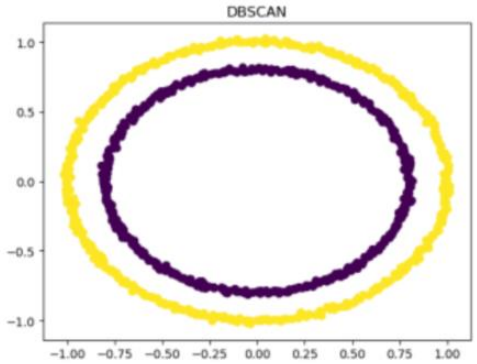
$C_1 = \{x_1, x_4, x_8\}$，$C_2 = \{x_2, x_3, x_5, x_6\}$，$C_3 = \{x_7\}$，

第一次迭代后的 3 个簇的质心：(3.667,9),(5.75,4.5),(1,2)

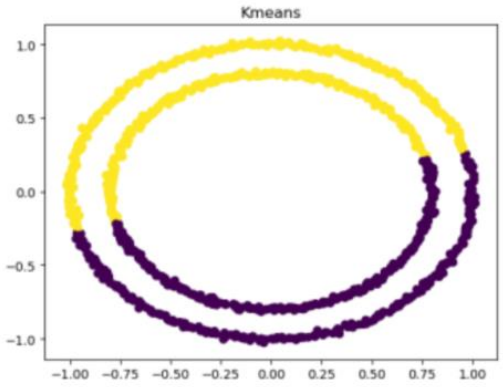
(2)最终的 3 个簇的质心为：(3.667,9),(7,4.333),(1.5,3.5)

3、相对于 K 均值与层次聚类，基于密度聚类方法可以处理不同大小和各种形状的簇，并且不太受噪声和离群点的影响。例如，当簇是圆环形状时如下所示。

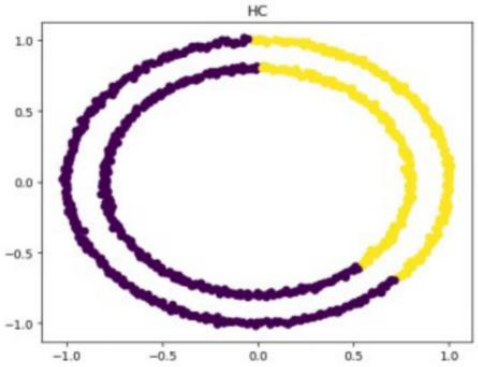
利用 DBSCAN 的聚类结果如下：



利用 K 均值的聚类结果如下：



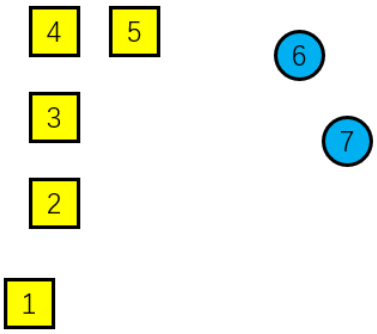
利用层次聚类结果如下：



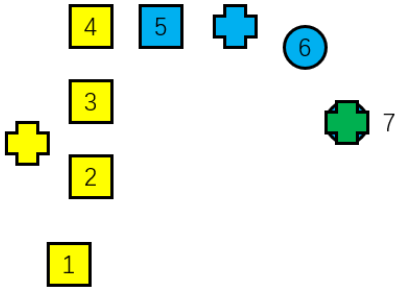
明显的，DBSCAN 更使用于非簇形状的聚类。

4、会出现缺失族的情况，这个问题同聚类过程中产生空簇是-一个问题。原因在于初化中心选择不当。举例说明：

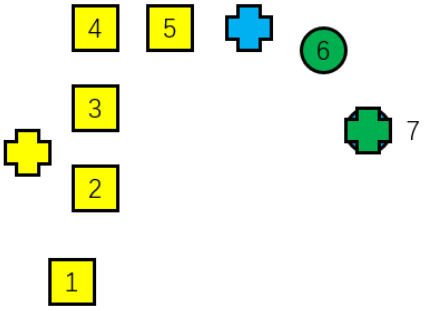
首先，假设有如下数据



聚类数设置为 3，初始中心选择 1，6，7，那么初始分组结果为{1,2,3,4},{5,6},{7}



基于更新的类中心再聚类，可以发现，第二类中的样本为空，形成空簇。无法更新，此时只能返回两个簇。



5、

AGNES：

执行过程

在所给的数据集上运行 AGNES 算法，算法的执行过程如表 5.8 所示，设n = 8，用户输入的终止条件为两个簇。初始簇为{1}，{2}，{3}，{4}，{5}，{6}，{7}，{8}。（采用最小距离计算）

步骤	最近的簇距离	最近的两个簇	合并后的新簇
1	1	{1},{2}	{1,2},{3},{4},{5},{6},{7},{8}
2	1	{3},{4}	{1,2},{3,4},{5},{6},{7},{8}
3	1	{5},{6}	{1,2},{3,4},{5,6},{7},{8}
4	1	{7},{8}	{1,2},{3,4},{5,6},{7,8}
5	1	{1,2},{3,4}	{1,2,3,4},{5,6},{7,8}
6	1	{5,6},{7,8}	{1,2,3,4},{5,6,7,8}

（1）先根据最小距离计算公式，将两两样本点的距离计算出来随机找出距离最小的两个簇，进行合并，最小距离为 1，合并 1、2 点为一个簇。

（2）对上一次合并后的簇进行簇间计算，找出距离最近的两个簇进行合并，合并后 3、4 合并成为一簇。

（3）重复第（2）步的工作，5、6 成为一簇。

（4）重复第（2）步的工作，7、8 成为一簇。

（5）合并{1,2}，{3,4}成为一簇。

（6）合并{5,6}，{7,8}成为一簇，合并后的簇的数目达到终止条件，计算完毕。

DIANA:

执行过程

步骤	具有最大直径的簇	Spliner group	Old party
1	{1,2,3,4,5,6,7,8}	{1}	{2,3,4,5,6,7,8}
2	{1,2,3,4,5,6,7,8}	{1,2}	{3,4,5,6,7,8}
3	{1,2,3,4,5,6,7,8}	{1,2,3}	{4,5,6,7,8}
4	{1,2,3,4,5,6,7,8}	{1,2,3,4}	{5,6,7,8}
5	{1,2,3,4,5,6,7,8}	{1,2,3,4}	{5,6,7,8}终止

在第 1 步中，根据初始簇计算每个簇之间的距离，对簇中的每个点计算平均相异度（假定使用欧式距离）

平均距离如下：

样本	1	2	3	4	5	6	7	8
平均距离	2.96	2.526	2.68	2.18	2.18	2.68	2.526	2.96

挑出平均相异度最大的点 1 放到 Spliner group 中，剩余点放在 Old party 中。

第 2 步，在 Old party 里找出最近的 Spliner group 中的点的距离不大于到 Old party 中最近的点的距离的点，将该点放入 Spliner group 中，改点是 2。

第 3 步，重复第 2 步的工作，在 Spliner group 中放入点 3。

第 4 步,重复第 2 步的工作，在 Spliner group 中放入点 4。

第 5 步，没有新的 old party 中的点分配给 Spliner group，此时分裂的簇数为 2。达到终止条件。如果没有到终止条件，下一阶段还会从分裂好的簇中选一个直径最大的簇按刚才的分裂方法继续分裂。