大数据分析第三次作业

（最晚提交时间 2025.05.23,18:00 之前提交到微助教上）

1. 使用图 3.4 的数据，计算如下 4 个哈希函数得到的最小哈希签名矩阵(20 分)。

h1(x) =x+1 mod 5

h2(x) =3x+1 mod 5

h3(x) = 2x + 4 mod 5

h4(x) = 3x − 1 mod 5

Row	S1	S2	S3	S4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

图 3.4　4 个集合的矩阵

2. 对图 7.2，如果采用如下的两个簇之间距离定义，实现自底向上的层次聚类(Agglomerative clustering, or called bottom-up hierarchical clustering)，那么最终的聚类结果会如何改变？注意这儿不要求实现从 12 个簇依次合并到最后一个簇的全部过程，只需要实现从 12 个簇合并到 11 个簇， 再从 11 个簇合并到 10 个簇的过程即可（40 分）。

(a) 两个簇上点之间的最短距离，两个点分别来自不同的簇。

(b) 簇上点对之间的平均距离，两个点分别来自不同的簇。

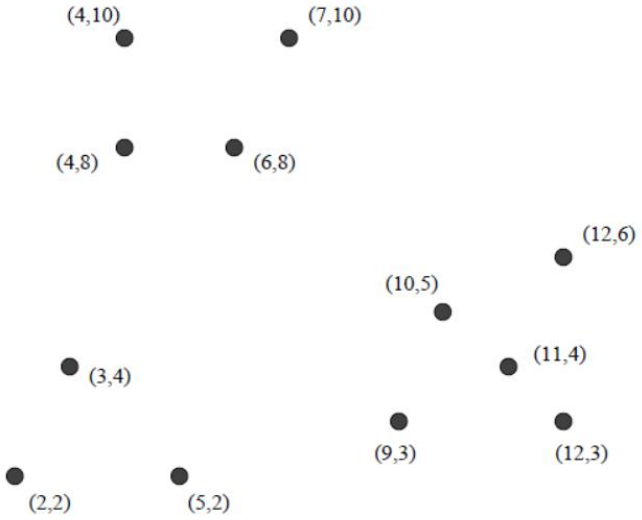

Figure 7.2: Twelve points to be clustered hierarchically

3. 对图 7.8 的点，选择 k-均值算法的三个初始点。要求选择彼此距离尽可能远的点。已知第一个点选择的是（3,4），那么剩下的两个点分别是哪两个点？（20 分）

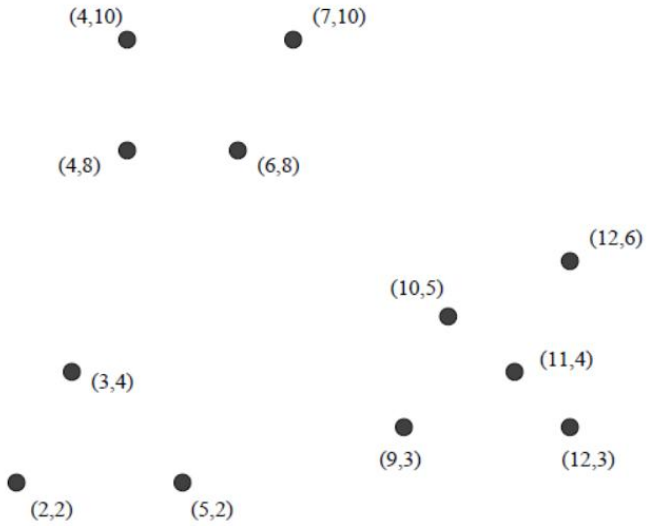

Figure 7.8: Repeat of Fig. 7.2

4. 对图 7.8 给出的三个簇，分别计算如下（20 分）。

（a）利用 BFR 算法中的方式计算所有簇的表示，也即计算簇的 N，SUM 和 SUMSQ 值；

（b）计算每个簇在两个维度中每个维上的方差和标准差。

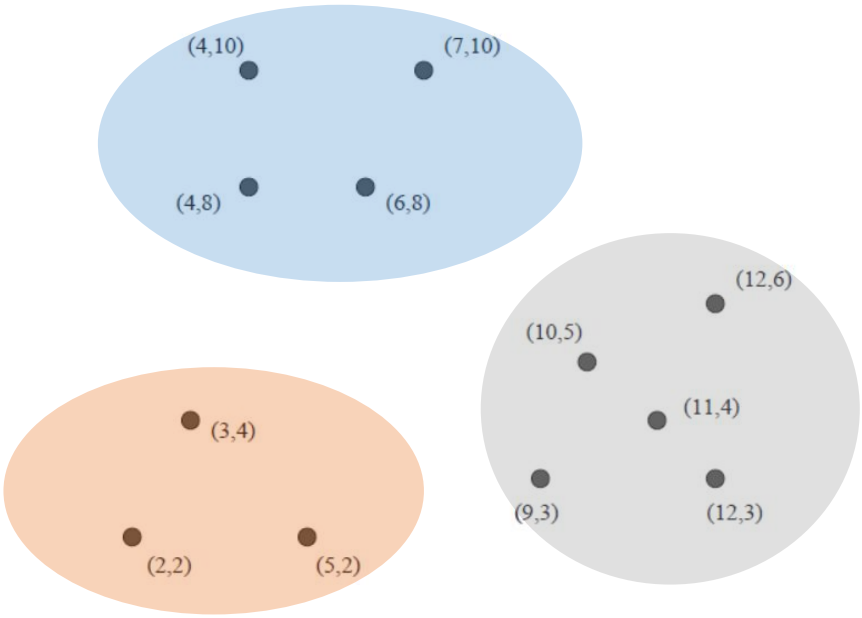

Figure 7.8: Repeat of Fig. 7.2