作业5 聚类

12月20日23:55提交

1. K均值聚类

给定8个数据点：。使用 执行K均值聚类，将8个点分组到簇C1和C2。

初始化簇中心分别为和。

1. 一次K均值聚类迭代后C1和C2的成员点是什么？

答：第一次迭代：选定 A1 和 A2 作为 C1 和 C2 的簇中心，A1 和 A2 可以省略不表，则有:



各簇的成员点为: 、

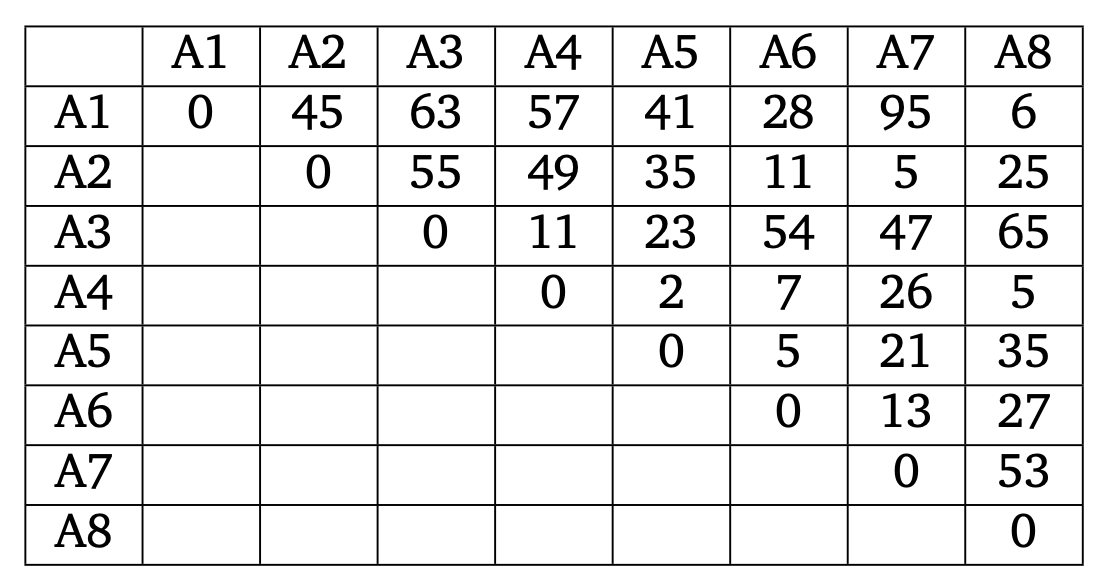
1. 第二次K均值聚类迭代后C1和C2的成员点是什么？

答：第一次迭代后，C1的中心更新为，C2 中心更新为，则有:



1. 基于密度的聚类算法

考虑如下距离矩阵。使用 对8个点采用DBSCAN算法进行聚类。给出迭代过程中每个簇及其成员点。如果有异常点，也请标出。



第一步：找出所有核心点：

第二步：集群连接核心/边界点：

簇 1：

簇 2：

异常值：

1. 聚类算法评估

使用内部评价指标：轮廓系数（Silhouette Coefficient）评估和比较下面3个聚类算法：

1. 具有单链路策略的凝聚层次聚类 （请自行确定适当的切割阈值）
2. DBSCAN
3. K均值聚类
4. 在单个样本层面、簇层面和整体层面，轮廓系数值高意味着什么？

单个样本级别的高轮廓系数表明该样本与其所属的簇匹配良好，而与相邻簇匹配不佳。

在簇级别/整体层面，高轮廓系数意味着簇彼此之间的距离很远，并且可以清楚地区分。

1. 哪种聚类算法提供具有最低和最高轮廓的结果聚类系数？

各种方法的轮廓系数为：

K均值：0.527，

DBSCAN：0.271，

单链路策略的凝聚层次聚类：0.223

1. 哪种聚类得到更均衡的聚类结果（每个簇中的样本数目相似）？

每种方法的聚类是：

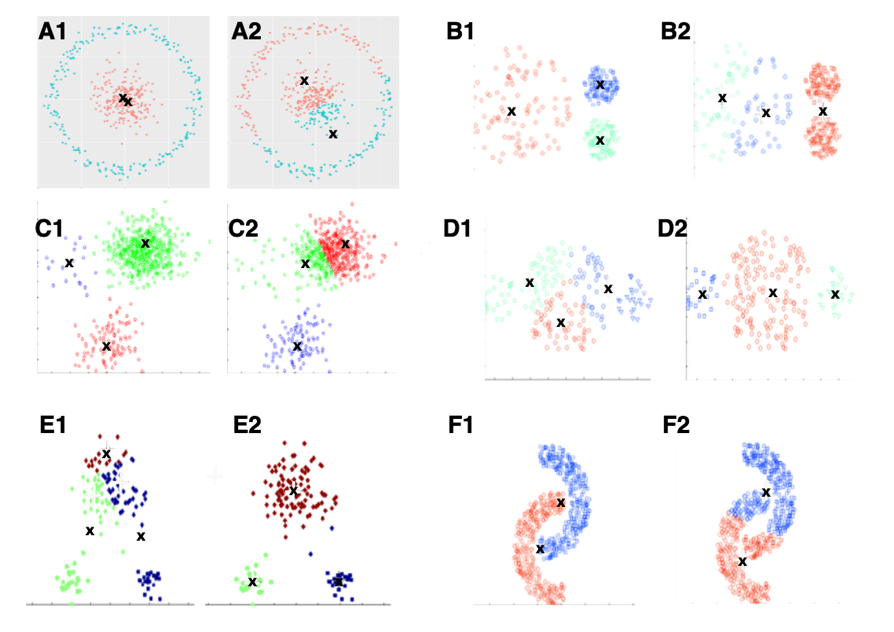
K均值: [1, 0, 0, 0, 0, 0,0, 1],

DBSCAN: [0, 1,‑1, 0, 0, 0, 1, 0] ,

单链路策略的凝聚层次聚类：[0, 0, 1, 0, 0, 0, 0, 0]

在这个例子中，种情况下，DBSCAN产生最均衡的聚类结果。

1. 下图给出了6个数据集A、B、C、D、E、F用两种聚类算法得到的聚类结果，已知其中一种聚类算法是K均值聚类。请问对每个数据集，哪个最可能是K均值聚类的结果。如果K均值聚类结果不够理想，对每个数据集，你建议采用哪种聚类算法？



答：

* A中A2最可能为K-Means的结果;可以考虑采用基于密度的DBSCAN/层次聚类里 MIN 距离的方法。
* B 中 B2 最可能为 K-Means 的结果;可以考虑采用 GMM/层次聚类里 Wards距离的方法。
* C 中 C2 最可能为 K-Means 的结果;可以考虑采用GMM或者层次聚类里 Wards 距离的方法。
* D 中 D1 最可能为 K-Means 的结果;可以考虑采用 GMM 方法。
* E 中 E1 最可能为 K-Means 的结果;可以考虑采用基于密度的 DBSCAN 方法。
* F 中 F2 最可能为 K-Means 的结果;可以考虑采用基于密度的 DBSCAN/层 次聚类里 MIN 距离的方法。