西南科技大学

计算机实验报告

课程名称：数据挖掘实践

实验名称：K-means算法

学 号：5120201314

学生姓名：王善斌

班 级： 信安2004

指导教师：吴珏

评 分：

实验日期： 2022.9.16

**一，实验目的：**

通过分析K-means聚类算法的聚类原理，利用编程工具编程实现K-means聚类算法，并通过对样本数据的聚类过程，加深对该聚类算法的理解与应用过程。

**二，实验内容：**

1. 分析K-means聚类算法；
2. 分析距离计算方法；
3. 分析聚类的评价标准；
4. 编程完成K-means聚类算法，并 基于相关实验数据实现聚类过程；

**三，实验要求：**

输入：簇的书面k和包含n个对象的数据库

输出：是平方误差准则最小的k个簇

算法流程：

1. 给定需要生成的簇的数目k
2. 随机选择k个对象作为初始的簇中心
3. 余下的对象，根据其与各个簇中心的距离，将他们赋给最近的簇
4. 重新计算每个簇的平均值
5. 以所得到的簇的平均值，返回第3步重新分配对象，直到准则函数收敛。

准则函数E：常常采用平方误准则来度量。

**四，实验步骤：**

K均值聚类算法 K-Means Clustering Algorithm 快速理解：

1.有四个牧师去郊区布道，⼀开始牧师们随意选了⼏个布道点，并且把这⼏个布 道点的情况公告给了郊区所有的居⺠，于是每个居⺠到离⾃⼰家最近的布道点去 K均值聚类算法（附例⼦）

2.之后，⼤家觉得距离太远了，于是每个牧师统计了⼀下⾃⼰的课上所有的 居⺠的地址，搬到了所有地址的中⼼地带，并且在海报上更新了⾃⼰的布道点的 位置。

3.牧师每⼀次移动不可能离所有⼈都更近，有的⼈发现A牧师移动以后⾃⼰还不 如去B牧师处听课更近，于是每个居⺠⼜去了离⾃⼰最近的布道点……就这样， 牧师每个礼拜更新⾃⼰的位置，居⺠根据⾃⼰的情况选择布道点，最终稳定了下来。

**步骤**

1）先定义总共有多少个类/簇（cluster）

2）将每个簇⼼（cluster centers）随机定在⼀个点上

3）将每个数据点关联到最近簇中⼼所属的簇上

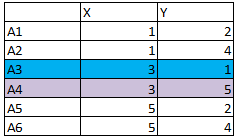
4）对于每⼀个簇找到其所有关联点的中⼼点（取每⼀个点坐标的平均值）

5）将上述点变为新的簇⼼

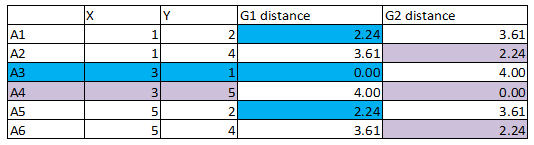
6）不停重复，直到每个簇所拥有的点不变

**例⼦**

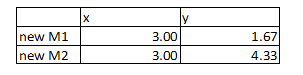
题⽬：有以下6个点，将A3和A4作为两个簇的初始簇⼼。问最后的簇的所属情况



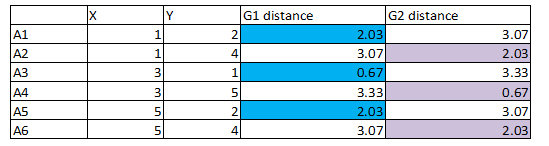
1. 计算每个点到簇⼼距离（根据距离公式），将距离近的点归为⼀类



1. 将蓝⾊每个点，和紫⾊每个点的X,Y值分别求平均。获得新的簇⼼



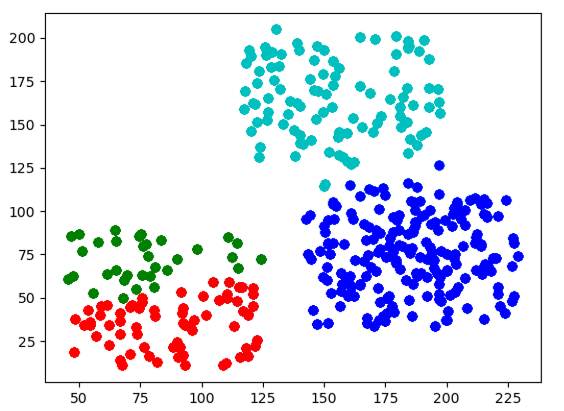
1. 计算每个点到簇⼼的新距离，将距离近的点归为⼀类



4.由于关联点没有变化，所以之后的计算结果不会改变。停⽌计算。

5. 蓝⾊簇:A1,A3,A5。紫⾊簇：A2,A4,A6

6.编程实现结果截图：



**五，实验思考：**

1. 考虑选择不同的距离函数，对结果的影响。

通常情况下，在聚类算法中，样本的属性主要由其在特征空间中的相对

距离离来表示。最常见的距离计算方法：欧式距离，余弦距离，曼哈顿距离；

聚类结果的好坏不仅是看选用的聚类算法如何，合理选取相似度度量函数对于最终聚类结果的好坏也有很大影响

1. 评价函数的计算。

优点：理解容易，聚类效果不错；处理大数据集的时候，该算法可以保证较好的伸缩性和高效率；当簇近似高斯分布的时候，效果非常不错。

缺点：K值是用户给定的，在进行数据处理前，K值是未知的，给定合适的 k 值，需要先验知识，凭空估计很困难，或者可能导致效果很差。对初始簇中心点是敏感的。不适合发现非凸形状的簇或者大小差别较大的簇。特殊值(离群值或称为异常值)对模型的影响比较大。

**六，实验体会：**

通过本次实验，我理解到了K-mens算法的过程以及原理，也知道聚类算法与分类算法最大的区别：聚类算法是无监督的学习算法，而分类算法属于监督的学习算法。

通过编程发现在聚类算法中根据样本之间的相似性，将样本划分到不同的类别中，对于不同的相似度计算方法，会得到不同的聚类结果，常用的相似度计算方法有欧式距离法。

在我们应用K-Means算法进行数据聚类时，往往会用这些分类后的数据进行下一步工作，那么最好的方法就是根据下一步工作的需要来选取聚类数量。