模式识别实验（一）

时间：2022.9.13 下午7、8节

地点：文宣楼B311、B312

实验名称：C均值聚类算法

实验目的：掌握动态聚类算法的基本思想；认识类别数、初始类心的选择对聚类结果的影响，认识到不同聚类评价指标对聚类结果的影响；编写能够对实际模式样本正确分类的c-均值算法程序并做出一定改进

实验数据集：

数据集1：iris.csv

Iris数据集中包含了3类鸢尾花特征数据。每一类分别由50个样本，每条样本有4个特征数据（花萼长度，花萼宽度，花瓣长度，花瓣宽度）

数据集2：sonar.csv

Sonar数据集包含了从不同角度返回的声纳信息，并以此来预测目标是岩石还是矿井，R代表岩石，M代表矿井。一共208个样本，60个维度，2个类别

数据集3：Compound.txt

2维人工数据集，第一第二列分别是点在二维平面上的坐标，第三列是所属聚类，非球面数据集

数据集4：threecircles.txt

2维人工数据集，第一第二列分别是点在二维平面上的坐标，第三列是所属聚类，非球面数据集

子实验1：C-均值法的改进 C的调整

实验原理：

将簇间误差平方和看成是类簇数量k的函数。随着k的增加，每个类簇内的离散程度越小，总距离平 方和也就在不断减小，并且减小的程度越来越不明显。极限情况是当k=N时，每个类簇只有一个点，这时总的误差平方和为0。最优聚类数目取值在1~N。

具体过程：

1. 取一个较小的k，计算聚类结果，例如k取为1。
2. 对每个k，计算总的簇间距离平方和。
3. 画出总簇间距离平方和随k值增加的变化趋势。

4. 图中弯曲的“拐点”处对应的k就是最合适的类簇数量

实验要求：

1.探索不同的c值，作出C-J曲线（定义在PPT24页）

2.对每个数据集分析最优聚类数目并给出理由

参考资料：

https://blog.csdn.net/weixin_43624833/article/details/125887812 手肘法

子实验2：改进初始聚类中心的选取

实验原理：

不同的初始聚类中心选取可能影响最终的聚类性能，原始c-means算法最开始随机选取数据集中c个点作为聚类中心，而K-means++按照如下的思想选取K个聚类中心：假设已经选取了n个初始聚类中心(0<n<K)，则在选取第n+1个聚类中心时：距离当前n个 聚类中心越远的点会有更高的概率被选为第n+1个聚类中心。在选取第一个聚类中心(n=1)时同样通过随 机的方法。

具体过程：

1. 数据点之间随机选择一个中心u1

2. 对于尚未选择的每个数据点x，计算x到与其最近的中心点之间距离

3. 使用加权概率分布随机选择一个新的数据点作为新中心，其中选择点x的概率与步骤2中计算的距离成比例

4. 重复步骤2、3，直到选择了k个中心（j=k）

5. 初始中心选择完毕之后继续使用标准c均值聚类

（可以采用ppt25页的初始类心选择步骤方法或其他初始类心选择方法）

实验要求：

1. 编码实现改进的初始聚类中心选择方法

2. 采用一定的聚类效果衡量方法，比较改进初始聚类中心选择方法和原始随机选择聚类中心的方法的优劣

参考资料：

https://blog.csdn.net/qq_42364307/article/details/111451367

kmeans++算法

子实验3：用类核代替类心

实验原理：

模式分析的一般任务是在一般类型的数据（例如序列，文本文档，点集，向量，图像等）中 找到并研究一般类型的关系（例如聚类，排名，主成分，相关性，分类）图表等）。内核方法将数据映 射到更高维的空间，希望在这个更高维的空间中，数据可以变得更容易分离或更好的结构化。

具体过程：

1.通过核函数将数据集映射到新的向量空间

2.计算新空间的欧氏距离

3.执行c-means方法

实验要求：

1.采用马氏距离作为类核进行聚类，对比马氏距离和传统欧氏距离对聚类效果的影响

2.（选做）采用核函数将原数据映射到高维向量空间，执行聚类

3.针对数据集3、4，使用不同颜色表示聚类画出聚类结果，并对比原始c-menas聚类结果和类核方法的聚类结果

参考资料：

https://blog.csdn.net/North_City_/article/details/113194193 kernel kmeans

https://blog.csdn.net/bluesliuf/article/details/88862918 欧氏距离和马氏距离

提交时间：下次实验课之前