****

****

**QG中期考核详细报告书**

**题 目**  **­**

**学 院 物理与光电工程学院 ­**

**专 业 电子科学与技术**

**年级班别 19级1班 ­**

**学 号 3119007189**

**学生姓名 王天成**

**2020年04月**

# 1.Apriori算法

**数据集的处理：**

依次读取数据集里的每行元素，并去掉每行头尾空白，将数据集中的内容保存到list1，遍历list1去掉其中的’,’,然后将数据存到dataSet

**算法步骤和思想：**

思想：如果一个项集是非频繁的，那么它的所有超集也是非频繁的。如果一个项集是非频繁的，那么它的所有超集也是非频繁的。

步骤：

1. 首先会生成所有单个元素的项集列表（对输入数据集进行去重，排序，放入list中，并转换所有元素为frozenset格式）
2. 对输入的数据集进行去重

3、扫描候选数据集（单个元素的数据集）来查看哪些项集满足最小支持度要求，那些不满足最小支持度的集合会被去掉，并返回支持度大于最小支持度的数据

4、对剩下的集合进行组合以生成包含两个元素的项集

5、接下来重新扫描候选数据集（两个元素的数据集），去掉不满足最小支持度的项集，重复进行直到所有项集都被去掉

**算法实现结果评估：**

利用mushroom数据集进行测试，e代表可食用，p代表不可食用。调用apriori函数，筛选支持度大于最小支持度的数据集。然后打印出对应长度的数据集，查看结果。打印了两种数据集，分别是长度为2，3的数据集。

**不足：**数据集中，除了可食用用e表示以外，还有其他的特征（如红色）也用e表示，另外还有一些数据缺失，用“？”表示，频繁项集中不能将这两类数据很好地分别开，造成结果不准确。

# 2.K-means算法

**数据集的处理：**

依次读取数据集里的每行元素，并去掉每行头尾空白，将数据集中的内容保存到list1，遍历list1去掉其中的’,’,然后将切片好的数据存到dataSet

**算法步骤和思想：**

思想：对于给定的样本集，按照样本之间的距离大小，将样本集划分为K个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。

步骤：

1. 创建一个与输入数据集的行数相同但有两列的数据集clusterAssment用来保存簇的信息
2. 在数据集范围内随机产生k个质心
3. 遍历每一行的数据，计算与k个质心的距离，算出最小距离，将最小距离和最小距离的质点索引保存到clusterAssment中
4. 更新质点为簇中所有点的平均值
5. 重复上述步骤，直至质点位置不再发生改变

**算法实现结果评估：**

利用dataSet数据集，多次调用kmeans函数，打印出质心的位置进行测试。

**不足：**测试结果出现错误，可能是由于数据集最后一行没有数据造成的

**优化之处：**通过调用numpy中seterr函数解决

# 3.Knn算法

**数据集的处理：**

依次读取数据集里的每行元素，并去掉每行头尾空白，将数据集中的内容保存到list中

**算法步骤和思想：**

思想：当预测一个新的值x的时候，根据离x最近的k的值的类别来判断x属于什么类别。

步骤：

1. 计算实验集中每个数据与输入数据data的距离，并将距离和对应的类别保存到字典c中
2. 将字典c保存到列表中
3. 将列表中数据按照字典中的距离进行由小到大排序
4. 提取最近的前k个值
5. 利于距离计算最近的k个值中，每种类别出现的加权平均值（因为离输入data越近的数值距离越小，所以最后要用1减去求得的平均值来表示加权平均值）
6. 最后比较得出加权平均值最大的类别，即为data最可能的类别

**算法实现结果评估：**

将iris.data数据集作为训练集，bezdekIris.data数据集作为测试集进行测试。测试失败。

**不足：**每次测试总距离都逐渐减小，最后减为零，导致运算出错。

**优化之处：**可以采用高斯函数来避免最后减到0的情况。