**K-means算法——基于对HCV数据的分析**

1. **问题描述**

在日常生活中，有许多人奉献爱心进行献血。但是并不是所有人的血都可以作为医用，有些人的因为先天或后天的原因，导致血液中携带一些病毒或不符合要求。其中肝问题是血液不合格无法献血的一个重要因素。因此此次聚类实验是为了将合格的献血者与不同类型肝问题人的血液进行分类（即献血者与丙型肝炎，包括其进展情况（“仅”丙型肝炎、纤维化、肝硬化））。

在HCV相关的数据中，共有13个特征，其分别为 ‘Age (in years)‘，’Sex (f,m)‘，’ALB‘，’ALP’， ‘ALT’， ‘AST’， ‘BIL’，‘CHE’，‘CHOL’，‘CREA’，‘GGT’，‘PROT’，其中，最后的聚类结果应为：

Category (diagnosis)

'0=Blood Donor',

'1=suspect Blood Donor',

'2=Hepatitis',

'3=Fibrosis',

'4=Cirrhosis'

由于在数据集中已经给出了最终的分类结果，因此我们在训练模型中需要先将这些分类结果隐藏，从而看最终的分类结果的准确率。

数据集链接为：https://archive.ics.uci.edu/ml/datasets/HCV+data

1. **模型理论**

**1、传统K-means算法**

Kmeans算法解决的问题是，在事先不知道如何分类的情况下（即无监督），让程序根据距离的远近，把N个对象（局部）最优的划分为k个类。它是无监督算法中比较常见的一种算法，原理比较简单易懂。本质是通过循环，不断迭代类中心点，计算各个对象到新的类中心点的距离并根据距离最近的原则重新归类，当类内距离最小、类间距离最大时，即可停止迭代（使用中，常常会限定迭代次数，防止陷入死循环。当达到预先设定得循环次数或类中心点不再发生变化时，最后一次迭代得到的结果，即为最终聚类结果）。

如果用数据表达式表示，假设簇划分为(C1,C2,...Ck)(C1,C2,...Ck)，则我们的目标是最小化平方误差E：

其中是簇的均值向量，有时也称为质心，表达式为：

如果我们想直接求上式的最小值并不容易，这是一个NP难的问题，因此只能采用启发式的迭代方法。

以下是K-means算法的具体步骤：

1. 对于K-Means算法，首先要注意的是k值的选择，一般来说，我们会根据对数据的先验经验选择一个合适的k值，在本次实验中，因为由数据集的结果我们知道可以分为5类，因此将k设成5.
2. 在确定了k的个数后，我们需要选择k个初始化的质心。由于我们是启发式方法，k个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的k个质心，最好这些质心不能太近。
3. 输入是样本集聚类的簇树k,最大迭代次数N；输出是簇划分
4. 从数据集D中随机选择k个样本作为初始的k个质心向量：
5. 将簇划分C初始化为
6. 对于计算样本和各个质心向量的距离：，将xi标记最小的为所对应的类别。此时更新
7. 对于,对中所有的样本点重新计算新的质心
8. 如果所有的k个质心向量都没有发生变化，则转到步骤4）

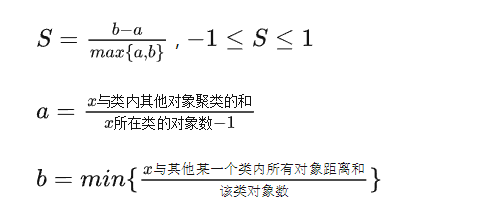
4）输出簇划分

**2、K-means++**

k个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的k个质心。如果仅仅是完全随机的选择，有可能导致算法收敛很慢。K-Means++算法就是对K-Means随机初始化质心的方法的优化。在本次实验中，我也采取了K-means++的方法，找到了间距最大的五个质心。

**3、评价指标**

聚类质量的评估主要是轮廓系数：其对应的公式如下：

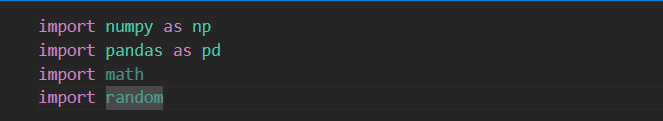


其中，a为x与其所属类内所有其他对象的平均距离，反映所属类的紧凑性，a越小越紧凑；b为x与其距离最近类内所有点的平均距离，反映x与其他类的分类程度，b越大越分离。

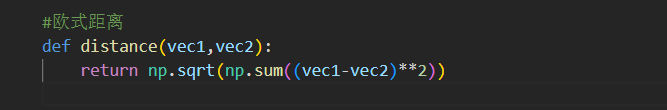
S取值范围在-1到1之间，S越接近1，说明包含x的类是紧凑的且x远离其他类，故聚类效果越好；

1. **代码实现**

**1、引用库**

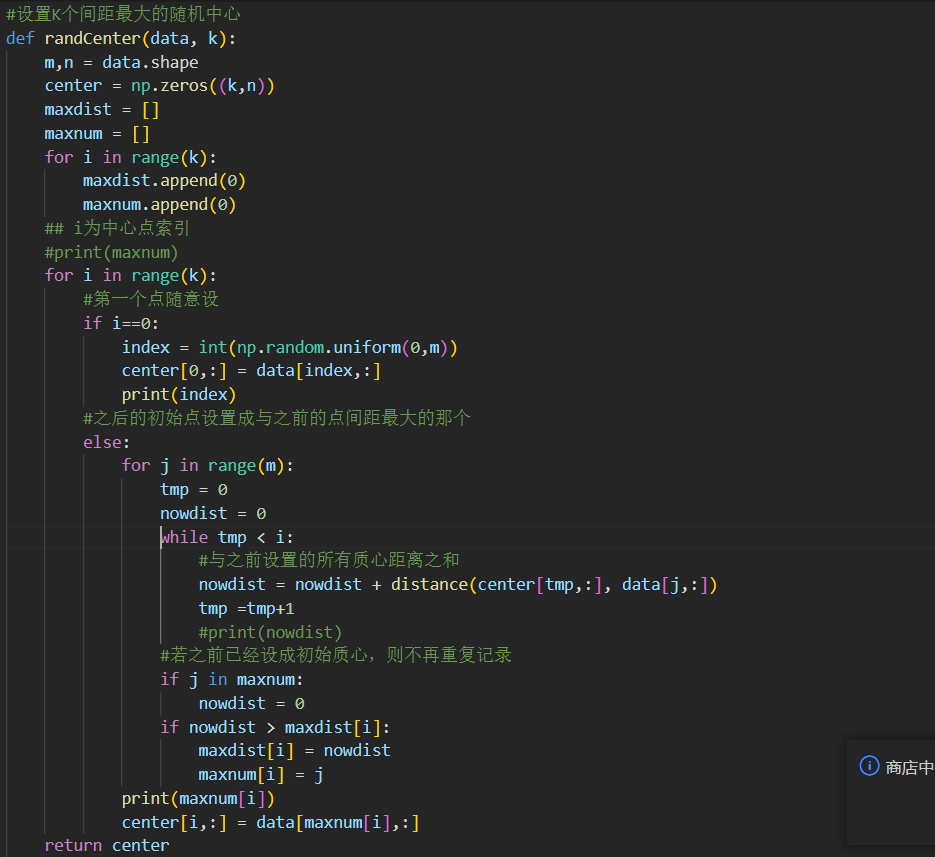
****

**2、计算欧式距离**

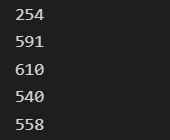
****

**3、初始化质心**

其中，这里采用了K-means++中的算法，设置间距最大的k个质心，使最终的结果更加准确。

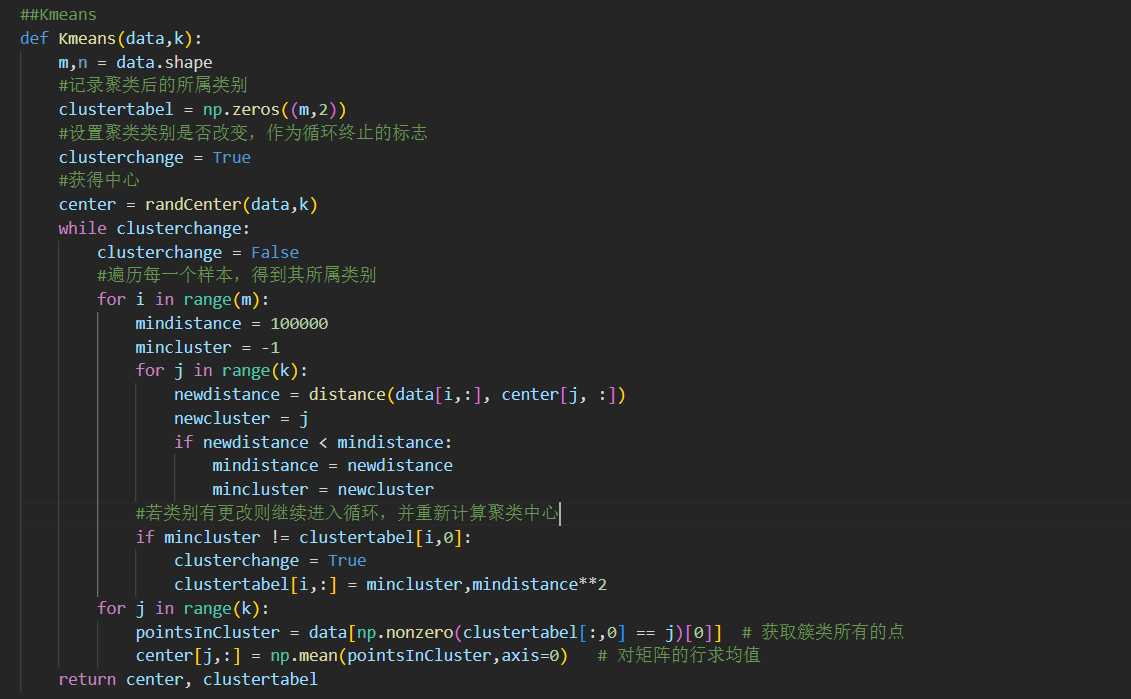
****

由于最终选择距离最大的质心中，未包含数据集中的有一类点，因此在这里通过手动增加一个包含该类的质心，并剔除其中的一个程序跑出来的质心。其结果如下：



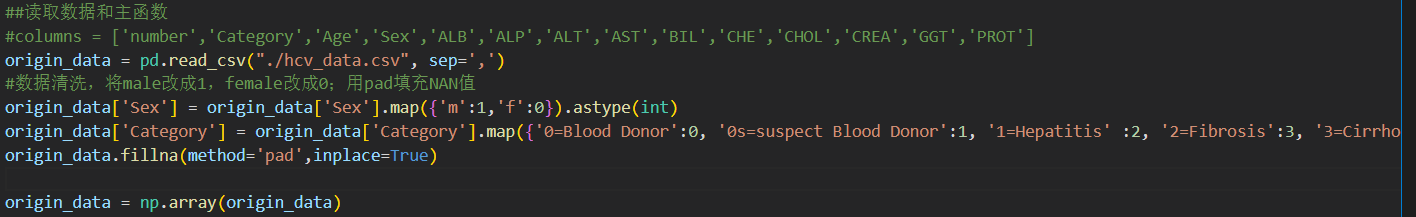
根据数据集比对，我保留了序号为254、540、558、610的质心，分别代表第0，1，2，4类，剔除了序号为591的质心，改成属于第3类中与其他三类最大的序号为574的数据。

**4、K-means算法主函数**

****

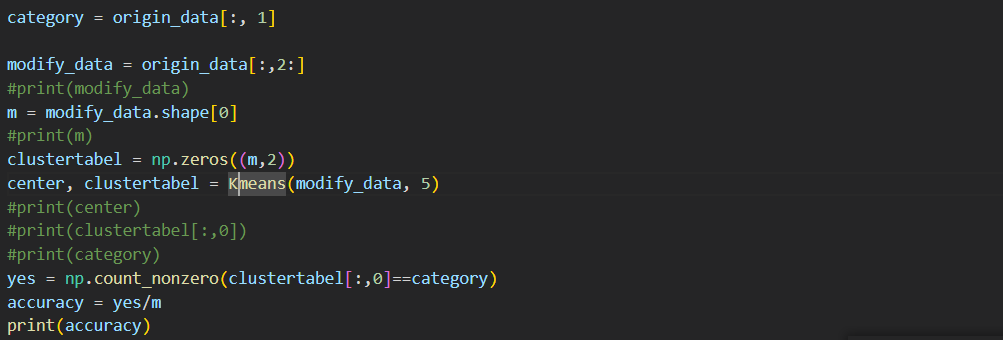
**5、数据预处理、清洗**

其中，用padding的方式填补数据集缺失值，且将二值类别（性别）进行赋值，其中男性赋值为1，女性赋值为0。同时，剔除标签列，实现无监督学习。在最开始的实验中，我加入了年龄和性别这两个变量（探究是否对聚类结果有影响），之后我又将这两个属性的数据删去，再运行结果，进行一定的比较分析。

****

**6、主函数并计算结果**

读入经过处理后的数据，并计算出最终的准确率

****

1. **结果分析**

**1、基础模型表现**

我将由K-means模型运行出来的结果与原来数据集中进行比较，战术部分结果如下：

运行结果：





原有的：





最终的精确度为Accuracy = = 86.7%，结果还算比较好。符合误差范围内，这说明我们的聚类模型分类效果是有一定说服力的，可以帮助我们正确地识别出该血液所属的人是否患有肝类问题。

**2、进一步讨论分析**

由于在之前提到过，我加入了年龄和性别两个属性，因此我尝试将这两个属性剔除，从而得到以下的精确度：



发现其精确度下降，说明加入年龄和性别属性还是有一定效果的。

此外，为了更好的具有实际效应，可分为正常血液和非正常血液的二分类问题，因此将原有的0，1归为一类，其他归为一类，则得到的最终结果为：



可以看出，准确度提高了将近10%，因此对于正确判断血液是否有问题有极好的效果。

由于数据集中已经有具体的标签可以验证精确度，故在此不再采用轮廓系数进行评价。

1. **总结展望**
2. **总结**

此次实验很好的将学习过的K-means算法应用到实际问题中，加深的对知识的理解，同时最终的结果还算比较符合预期，有一定的使用价值。能够将血液进行一定的区分，特别是如果只检测血液是否合格，血液主人是否存在肝问题，具有很好的效果

1. **展望**

由于K-means问题具有一定的局限性，如（算法常陷入局部最优，更换初始聚类中心后，新的聚类结果可能效果更优、对孤立点敏感，如数据集存在异常突出点，会影响聚类效果），可以对K-means算法进行进一步的改进，后续课余时间可以通过主成成分分析（PCA）对该问题进行进一步的改进，同时可以采用不同的办法，GMM算法，谱聚类算法等进行一定的尝试。