

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： 校交1902**

**学 号： U201912633**

**姓 名： 张睿**

**指导教师： 王蔚**

**报告日期： 2022.1.3**

**计算机科学与技术学院**

**目录**

[实验四 kmeans算法及其实现 1](#_Toc58793873)

[**4.1实验目的** 1](#_Toc58793874)

[**4.2 实验内容** 1](#_Toc58793875)

[**4.3 实验过程** 2](#_Toc58793876)

[4.3.1 编程思路 2](#_Toc58793877)

[4.3.2 遇到的问题及解决方式 2](#_Toc58793878)

[4.3.3 实验测试与结果分析 2](#_Toc58793879)

[**4.4 实验总结** 2](#_Toc58793880)

# 实验四 kmeans算法及其实现

## **4.1实验目的**

1、加深对聚类算法的理解,进一步认识聚类算法的实现；

2、分析kmeans流程,探究聚类算法院里；

3、掌握kmeans算法核心要点；

4、将kmeans算法运用于实际，并掌握其度量好坏方式。

## **4.2 实验内容**

提供葡萄酒识别数据集，数据集已经被归一化。同学可以思考数据集为什么被归一化，如果没有被归一化，实验结果是怎么样的，以及为什么这样。

同时葡萄酒数据集中已经按照类别给出了1、2、3种葡萄酒数据，在cvs文件中的第一列标注了出来，大家可以将聚类好的数据与标的数据做对比。

编写kmeans算法，算法的输入是葡萄酒数据集，葡萄酒数据集一共13维数据，代表着葡萄酒的13维特征，请在欧式距离下对葡萄酒的所有数据进行聚类，聚类的数量K值为3。

在本次实验中，最终评价kmean算法的精准度有两种，第一是葡萄酒数据集已经给出的三个聚类，和自己运行的三个聚类做准确度判断。第二个是计算所有数据点到各自质心距离的平方和。请各位同学在实验中计算出这两个值。

实验进阶部分：在聚类之后，任选两个维度，以三种不同的颜色对自己聚类的结果进行标注，最终以二维平面中点图的形式来展示三个质心和所有的样本点。效果展示图可如图1.1所示。

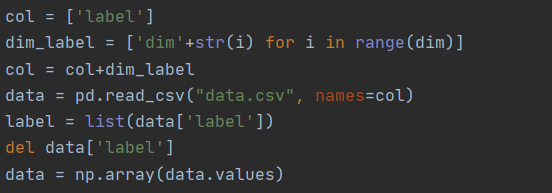


图4.1 葡萄酒数据集在黄酮和总酚维度下聚类图像（SSE为距离平方和，Acc为准确率）

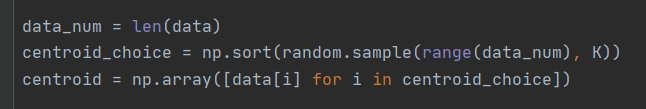
## **4.3 实验过程**

### 4.3.1 编程思路

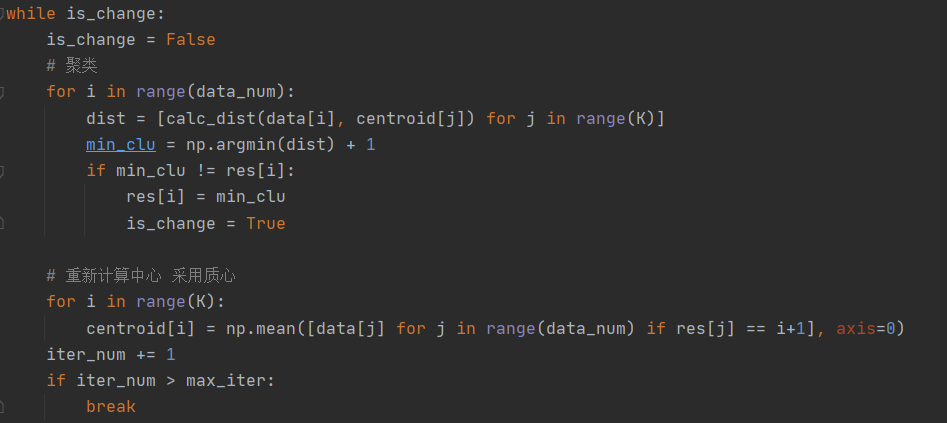
K-means思路很简单，首先读入原始数据并将其转换为矩阵方便操作。



接着利用random模块随机选择三个初始点。

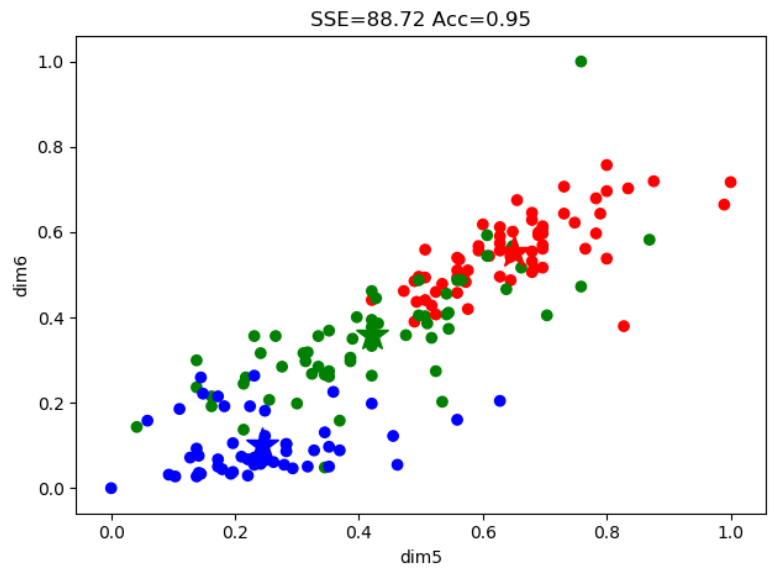


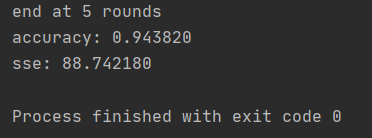
随后开始迭代，每一次迭代都重新计算每个点到中心点的距离并重新归到距其距离最小的中心点的类，并重新计算中心点，计算中心点的方式采用质心，及当前类中所有的点对应维度的平均值作为中心点当前维度的值，迭代结束的条件一方面是每个点不再变化类别以及不超过最大迭代次数。



最后利用matplotlib进行可视化分析

### 4.3.2 实验测试与结果分析





这里利用第五个维度和第六个维度进行平面的分析，星型的点代表中心点，可以看到，基本上分的非常好SSE为88.72，准确度有95%，需要注意的是，聚类算法有随机性，有一定可能结果非常差，这是因为最初的三个点是随机选的，并且是随机分配类的，这有两个影响因素，一方面初始点是随机的，会导致结果偏差，因为有可能选到三个特别近的点，另一方面，初始分配的类也是随机的，所以有可能点虽然聚对了，但是这个类的label不对，这也会使准确率很低，但是sse并不会受此影响。除此之外，由于数据集处理地非常好，可以看到5轮左右迭代就会停止。

## **4.4 实验总结**

K-means的实验较为简单，主要就是理解算法的过程，由于数据集处理地非常好，效果也很喜人，通过这次实验，学到的更多的是可视化的使用，这是之前接触比较少的知识。