

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1806**

**学 号： U201814670**

**姓 名： 李田田**

**指导教师： 崔金华**

**报告日期： 2020.12.17**

**计算机科学与技术学院**

**目录**

[实验四 kmeans算法及其实现 1](#_Toc58252477)

**[4.1实验目的](#_Toc58252478)** [1](#_Toc58252478)

**[4.2 实验内容](#_Toc58252479)** [1](#_Toc58252479)

**[4.3 实验过程](#_Toc58252480)** [2](#_Toc58252480)

[4.3.1 编程思路 2](#_Toc58252481)

[4.3.2 遇到的问题及解决方式 2](#_Toc58252482)

[4.3.3 实验测试与结果分析 3](#_Toc58252483)

**[4.4 实验总结](#_Toc58252484)** [3](#_Toc58252484)

# 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所示。



图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

因为数据为13维的数据，实验中通过将归一化数据的每一行作为一个样本，得到二维数组data；

先在随即选取k个点，坐标范围为0-1（数据已经归一），此处k为3；

对于数据中的每一个样本点，选取与其最近的聚类中心，并计算距离，记录每一个样本点所选取的聚类中心序号和距离。

现在有k个聚类，计算每一个聚类的中心，替换原来的中心。

循环前两个步骤，直到记录的每一个样本点的聚类中心都不再变化，且每一个点到质心的和不再改变。

利用 python 第三方库中的可视化工具 matplotlib.pyplot 对聚类后的元素显示（散点图），查看结果。

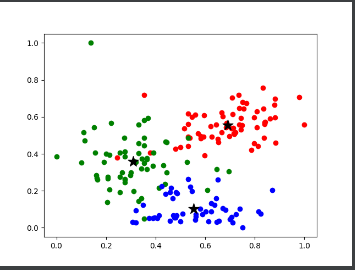
### 4.3.2 遇到的问题及解决方式

在计算每个聚类的中心的时候，刚开始想到直接判断，然后遍历，总觉得有点麻烦。后来一想，将二维数组转化为矩阵，由矩阵可以直接用函数mean计算每一列的均值。得到的列均值，而且是十三列可以直接赋给该聚类中心centers[j, :]。

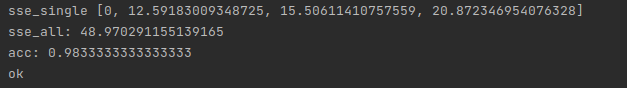
在使用可视化工具的时候，一开始使用的时候一直没有图像，后来看了教程例子才知道plt的具体用法，要用show函数展示出来。

### 4.3.3 实验测试与结果分析

测试结果：横坐标是酒精，纵坐标是黄酮。



SSE与ACC如下图所示：



结果分析：从散点图中可以清楚地看出该算法将数据分为三大聚类，得到的聚类相对比较明显。

## **4.4 实验总结**

如果没有被归一化，实验结果将比较分散，而且循环次数也会变多。因为未归一化之前，数据集的书记分布的区间较大，各特征之间数值差异过大；同时循环次数也会变多。

在本次实验的过程中，主要学习了kmeans聚类算法的过程，使用pyplot画图，并将自己的测试结果与已分结果进行判断，收获良多。同时也学会了kmeas算法度量好坏的方式，掌握了该算法的核心。