

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1908**

**学 号： U201915160**

**姓 名： 陈千鹤**

**指导教师： 杨驰**

**报告日期： 2021.12.24**

**计算机科学与技术学院**

**目录**

[实验四 kmeans算法及其实现 1](#_Toc91148749)

[**4.1实验目的** 1](#_Toc91148750)

[**4.2 实验内容** 1](#_Toc91148751)

[**4.3 实验过程** 2](#_Toc91148752)

[4.3.1 编程思路 2](#_Toc91148753)

[4.3.2 遇到的问题及解决方式 2](#_Toc91148754)

[4.3.3 实验测试与结果分析 2](#_Toc91148755)

[**4.4 实验总结** 3](#_Toc91148756)

# 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,探究聚类算法原理；

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

1. 实现距离计算函数

2. 实现循环计算每个点到中心的距离，并根据距离将点分到最近的簇中；随后重新计算簇的中心点；当迭代次数达到上限，或者中心点变化小于阈值后退出循环

3. 计算准确率

4. 计算所有点到中心点的距离和

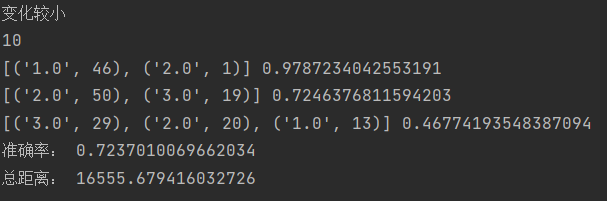
5. 选取两个维度绘制可视化坐标图

### 4.3.2 遇到的问题及解决方式

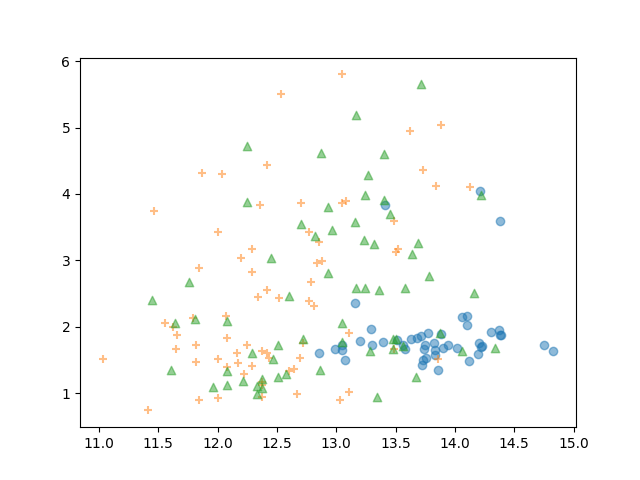
无

### 4.3.3 实验测试与结果分析

实验结果如下图所示：



可视化坐标图如下图所示：



## **4.4 实验总结**

1. 加深对聚类算法的理解,实现了常见聚类算法；

2. 学习了kmeans流程,了解了聚类算法原理；

3. 学到了kmeans算法核心要点；

4. 实现了kmeans算法，并了解了其度量好坏的方式。