**广州大学学生实验报告**

**开课学院及实验室：**计算机科学与网络工程学院软件实验室 **2020年 11 月20日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **18级软件工程3班** | **姓名** | 李子轩 | **学号** | 1806300106 |
| **实验课程名称** | **机器学习与数据挖掘实验** | | | | | **成绩** |  |
| **实验项目名称** | k-means聚类算法 | | | | | **指导老师** | 彭伟龙 |

**实验三**

1. **组员信息**

李子轩（小组共1人）

1. **作业环境（文件说明，函数说明，调用的函数库以及涉及哪些技术）**

**文件说明：**

数据源1：iris.data

数据源2：iris.names（是对上述数据源相关数据的说明文档）

**函数说明：**

在代码中均有详细标识和说明，在此不重复描述。利用了matplotlib模块进行可视化的处理，调用了pandas、numpy下的函数和方法对数据集进行处理。

**调用的数据库以及涉及的技术：**

未调用数据库。

涉及的技术主要包括利用了matplotlib模块进行可视化的处理，在处理数据集方面使用了如pandas以及numpy来辅助，未使用基础库以外的API和库函数，均通过自定义相关的函数对k-means算法进行实现。

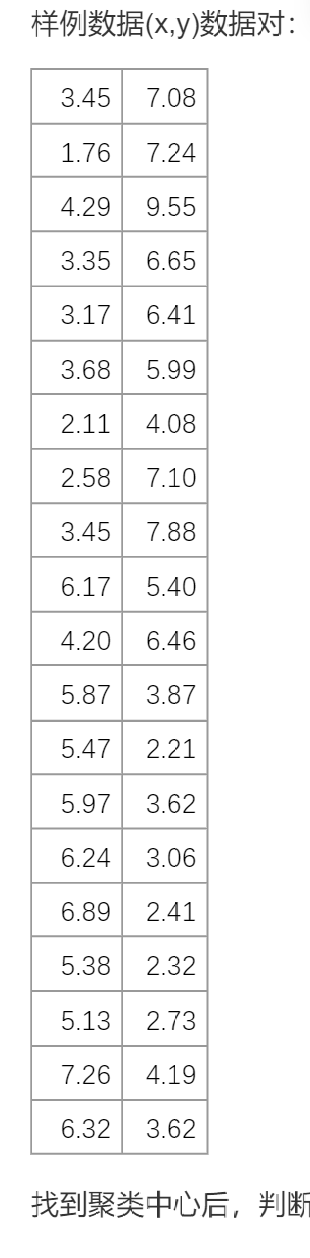
1. **作业题目和内容**

**题目：**用C++实现k-means聚类算法，

1. 对实验二中的z-score归一化的成绩数据进行测试，观察聚类为2类，3类，4类，5类的结果，观察得出什么结论？

2. 由老师给出测试数据，进行测试，并画出可视化出散点图，类中心，类半径，并分析聚为几类合适。

样例数据(x,y)数据对：



找到聚类中心后，判断(2,6)是属于哪一类？

注意

除文件读取外，不能使用C++基础库以外的API和库函数。

**（此处在实验中实验了经典的数据集iris来完成k-means聚类算法，最后将结果可视化为散点图，实验中使用的是python来完成本次实验）**

**具体的实验过程：**

**读取数据，并做预处理：**首先将iris class的数据类型转换为数值型，便于后面的计算，如图1-2所示。由于在数据中除了class类还存在四项类，所以在实验中以花萼面积为横坐标，以花瓣面积做纵坐标，

**实现算法：**kmeans算法的核心大致为先从样本集中随机选取 k个样本作为簇中心，并计算所有样本与这 k个“簇中心”的距离，对于每一个样本将其划分到与其距离最近的“簇中心”所在的簇中，对于新的簇计算各个簇的新的“簇中心”。

根据以上描述，我们大致可以猜测到实现kmeans算法的主要四点：

（1）簇个数 k 的选择；

（2）各个样本点到“簇中心”的距离；

（3）根据新划分的簇，更新“簇中心”；

（4）重复上述2、3过程，直至"簇中心"没有移动。

由于数据中的class只有三类，所以实验中的k只能取3，因为class只有3类 分别为Iris Setosa、Iris Versicolour和Iris Virginica，实验中为聚3类。

在初始时首先利用random\_select()从样本集中随机选取k个样本作为簇中心，并将随机选择到的k个簇中心的相关数据记录下来，用于后续的测试。利用distance()计算出初始时各个样本点到k个簇中心的欧式距离，其中均已花萼面积为横坐标，花瓣面积做纵坐标，然后调用kmeans()按照k值聚类，遍历所有样本点，找到每个点与簇中心之间欧式距离最短的簇中心的下标，即对于每一个样本，将其划分到与其距离最近的簇中心所在的簇中。接着调用distanceCenter()计算样本点到簇中心的距离，用于判断簇中心是否发生移动，centerPoint()则是辅助完成这一过程，找出各个簇的新的中心点，迭代一次后计算出各个样本点与新的簇中心的距离，比较前后两次与簇中心的距离，判断是否相等，不相等说明新的簇中心发生了移动，继续迭代，直至新的簇中心没有发生移动为止。

**选择适当可视化方法显示结果：**分别打印出三个簇中心的横坐标和纵坐标以及迭代的次数，以花萼面积为横坐标，花瓣面积为纵坐标进行可视化操作，最终结果如图1-3和1-1所示。

1. **难题与解决**

在本次的实验过程中主要出现的困难在于对kmeans算法的具体实现上，虽然对算法的内核已经有了基本的了解，但是在实际上手编程中由于对细节把握的不到位，导致了结果中小的瑕疵比较多，不能做到是自己满意的程度。在经过多次调试、改善数据结构后，使代码的运行的效率更高，并且结构等各方面都趋于更加的合理，基本能够达到实验所要求的目的。

1. **总结**

通过本次实验让我对经典的聚类算法有了更加深入的了解，特别是对于kmeans算法的实现有了更加清晰的认识，在编程的过程中出现的一些不应该出现的小问题也让我意识到自己的水平仍有比较大的提升空间，需要进一步的学习和锻炼，需要对一些经典的算法有更加深入的分析，了解其内核，而不仅仅停留在表面。