姓名： 姜现日

学号：18201291086

实验日期： 2024-04-11

实验四：

**（1）实验内容**：K-means算法是经典的聚类算法，其基本思想是：以空间中k个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。假设要把样本集分为K个类别，算法描述如下：

（1）适当选择k个类的初始中心

（2）在第I次迭代中，对任意一个样本，求其到K个中心的距离，将该样本归到距离最短的中心所在的类

（3）利用均值方式更新该类的中心值

（4）对于所有的K个聚类中心，如果利用（2）（3）的迭代法更新后，值保持基本不变，则迭代结束，否则继续迭代。

要求用java编写K-means算法（k值可以自己设定），根据花的属性对数据集Iris Data Set进行聚类，并将聚类结果（sepal length，sepal width，petal length，petal width ，cluster label）打印至cluster.txt文件。

iris数据包括四个属性：sepal length花萼长度，sepal width花萼宽度，petal length花瓣长度，petal width花瓣宽度。其中第五个值表示该样本属于哪一个类。Iris.data 可以用写字板打开。

注意：样本点间的距离直接用向量的欧氏距离。

1. **实验思路**

用RANDOM类实现随机抽签，选K个初始中心值。用Hash Map管理K个聚类属于的点。利用TryCatch局管理输入的格式，用Split函数分离每个点以后在二维Array List保存。通过迭代计算样本归到距离最短的中心所在的类。李彤均值港式更新个类的中心值，如果中心值不变，迭代结束。计算完以后用Print Writer类创建文件，写每个点的属性。

1. **实验源码**

**（4）实验心得**

知道K-means 算法的定义，怎么聚类每个点。用random函数选择K个类的初始中心值的话每次的结果可能不同。理解选初始值是很重要的部分。学会了怎么管理和用Hash Map，利用Get和Put函数管理对每个Key的value。