# 《模式识别》课程

# 实 验 报 告



**姓 名： 金家耀**

**专 业：**  人工智能

**学 号： 1193210320**

**江南大学人工智能与计算机学院**

# K均值聚类算法

**1实验目的**

C均值聚类（更常用的叫法是K均值算法，K-means clustering）是经典的非监督数据处理方法。实验目的在于加深学生对C均值聚类原理的理解，掌握算法的实现过程，体会其在模式识别中的作用。

**2实验原理**

设定 C个类别并选取 C个初始聚类中心，按最小距离原则将各样本分配到 C类中的某一类；之后不断地计算各类中心并调整各样本的类别，最终使各样本到其所属类别中心的距离平方之和最小。

**3实验内容**

处理男生和女生的身高、体重数据，分别保存在文件FEMALE.TXT、MALE.TXT中，利用C均值方法进行聚类分析。

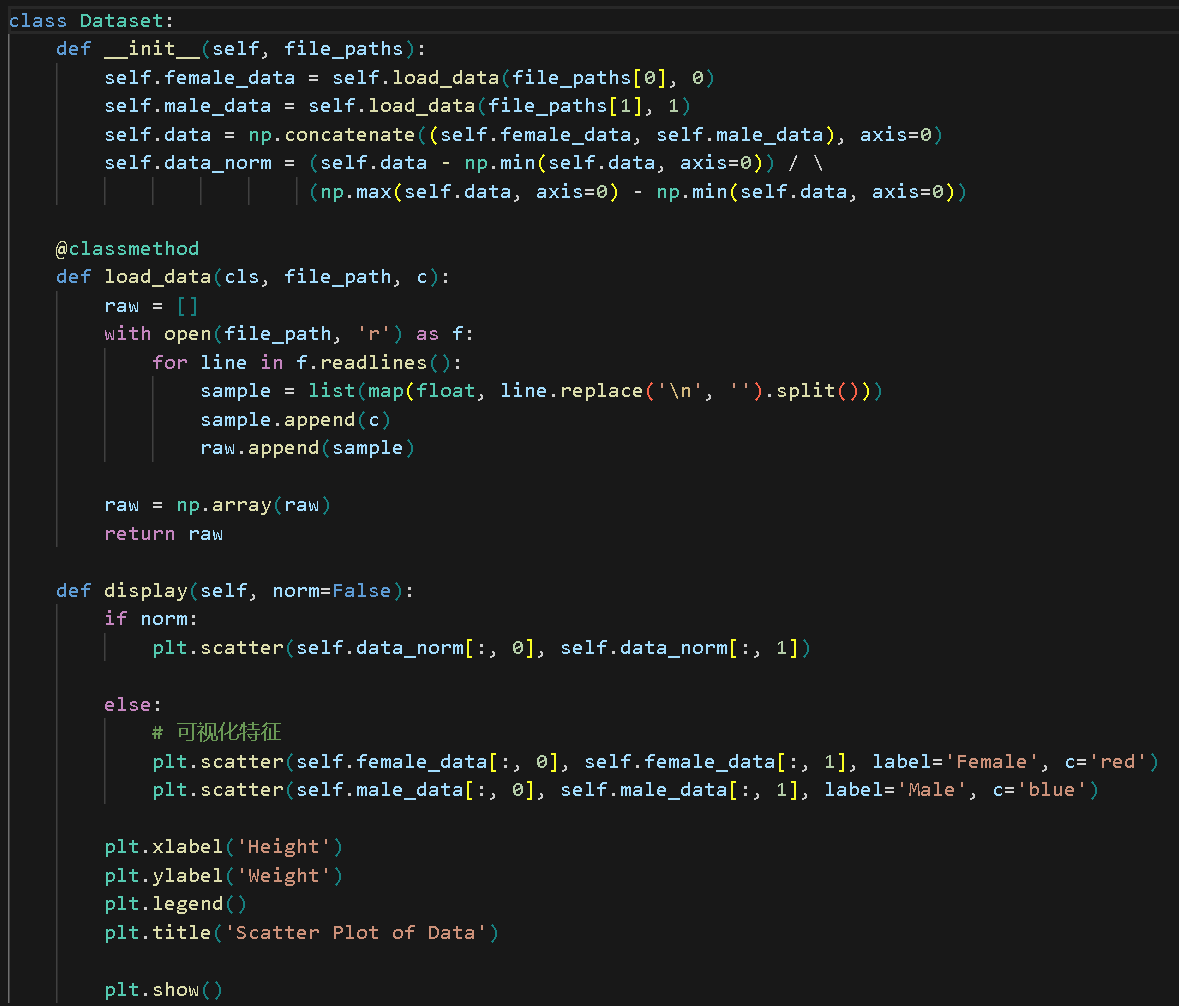
本实验使用MATLAB平台，对不同特征的样本进行分类

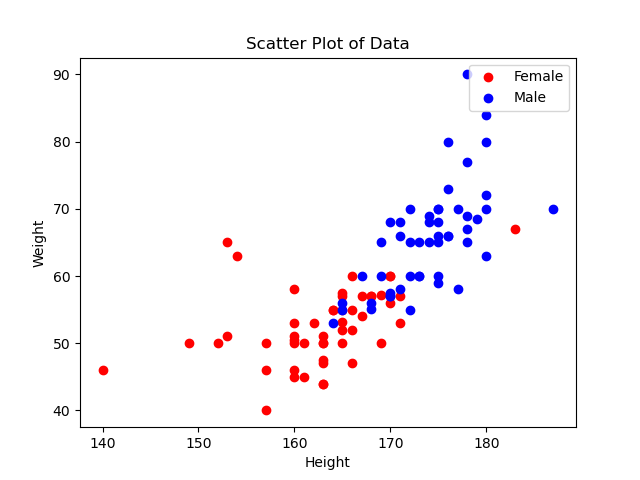
**4实验要求**

1. 同时采用身高和体重数据作为特征，类别数设为2，利用C均值聚类方法对数据进行聚类，并将聚类结果显示在二维平面上；
2. 尝试不同的初始值，观察聚类结果是否发生变化；
3. 改变类别数，分别进行两类、三类、四类、五类聚类，画出聚类指标J与类别数之间的关系曲线，探讨是否可以确定出合理的类别数目。

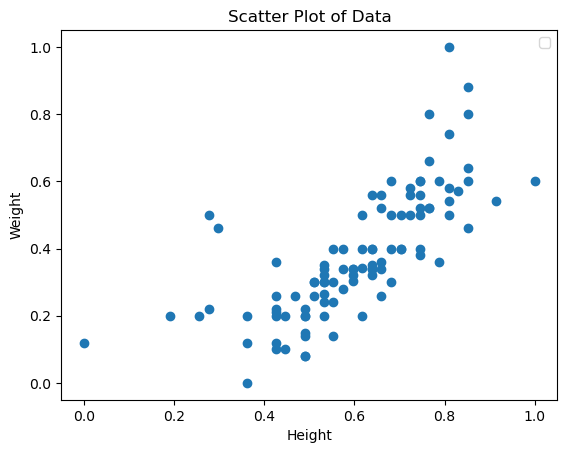
**5实验代码和结果**

1. **读入原始数据以及归一化后并可视化**



****

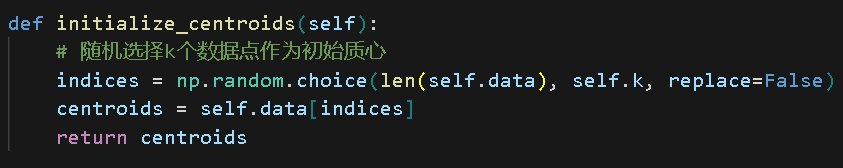
将包含标签的样本数据使用matplotlib作可视化，从散点图可以大致看出男性和女性大致分布在两个区域。

****

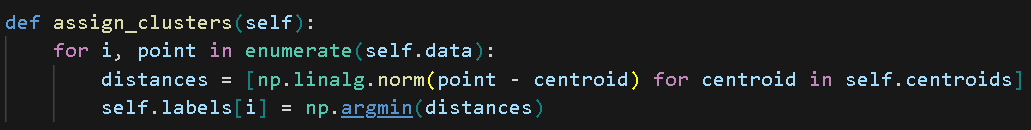
为保证聚类效果不受数据点特征分布影响，将数据点进行0-1归一化后进行数据可视化。

1. **k-means当类别为2时并可视化**
2. **means算法流程代码：**

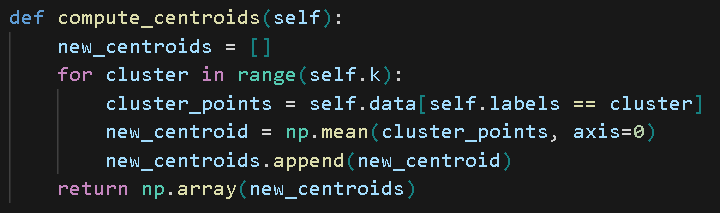
* **初始化聚类中心点**



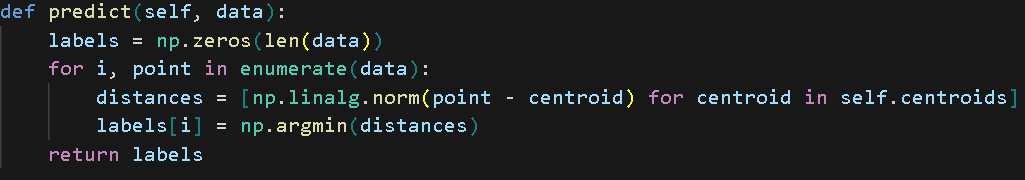
* **根据中心处理聚类簇**



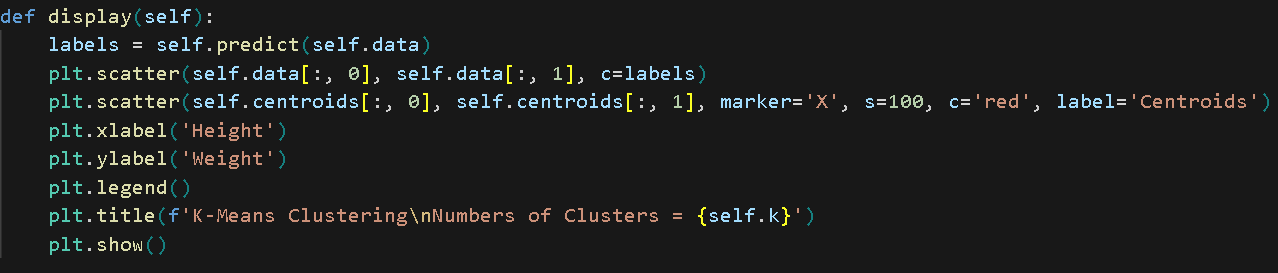
* **重新计算聚类中心**

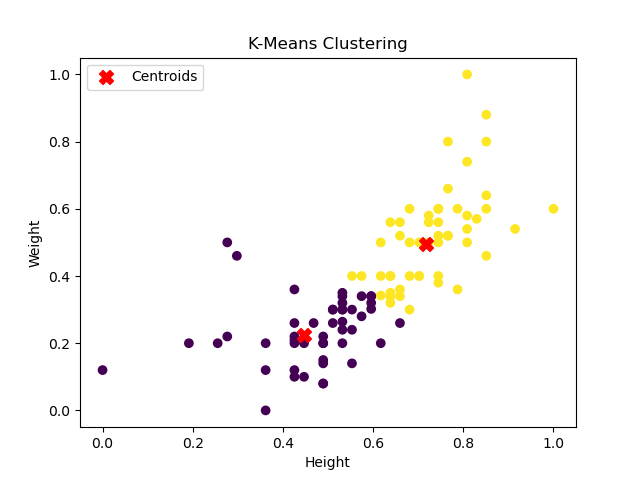


* **预测标签**



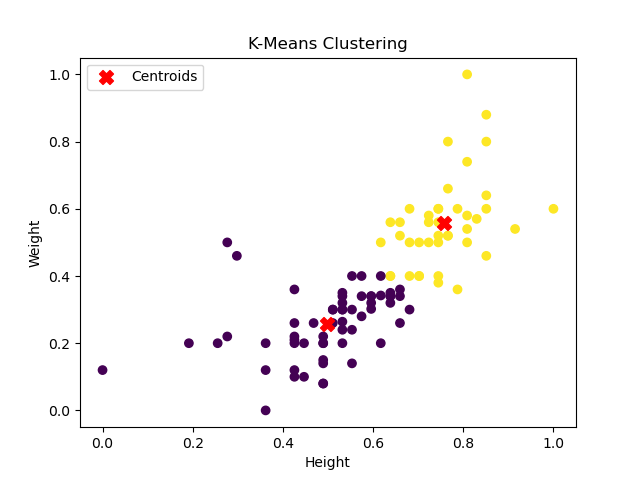
* **可视化预测结果**

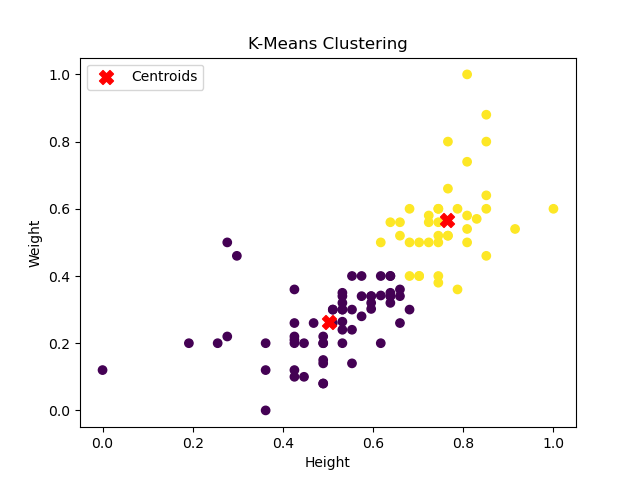


****

通过K-means（k=2）对于归一化之后的数据聚类，并将结果可视化在二维平面中，不同颜色表示一类。

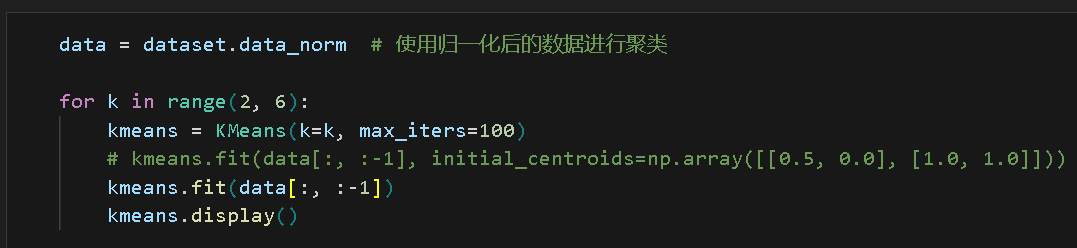
1. **尝试不同的初始值并可视化**

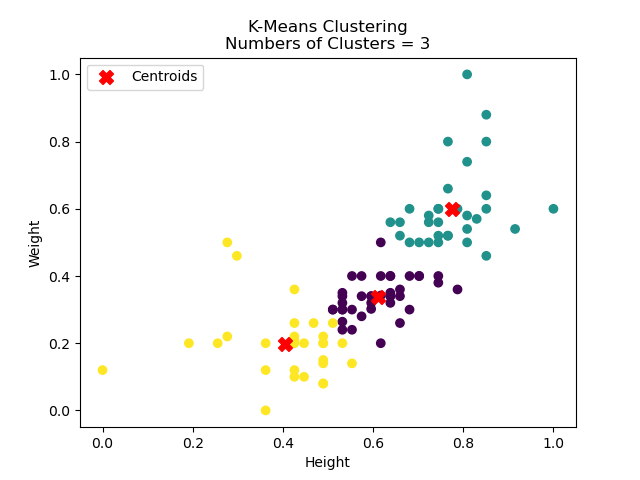
****

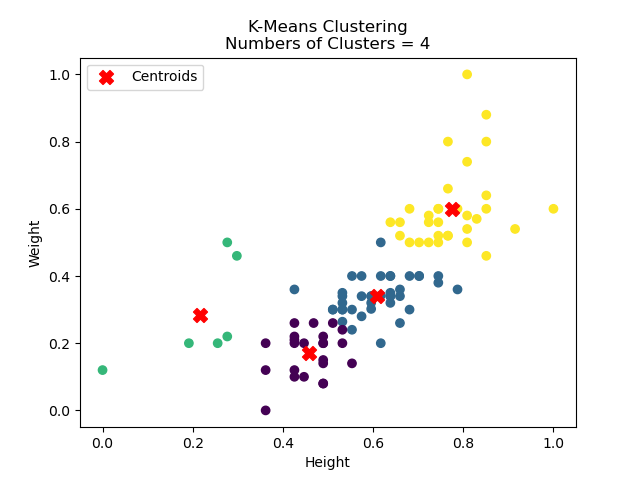
****

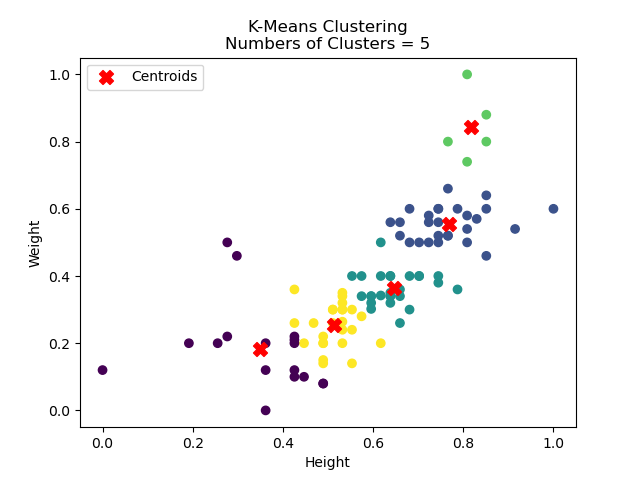
我尝试使用不同的初始值进行聚类算法并且把最大迭代次数改小，通过多次尝试以及可视化，由图中可以大致看出最终聚类结果不同。

1. **类别为3、4、5时的k-means结果可视化**



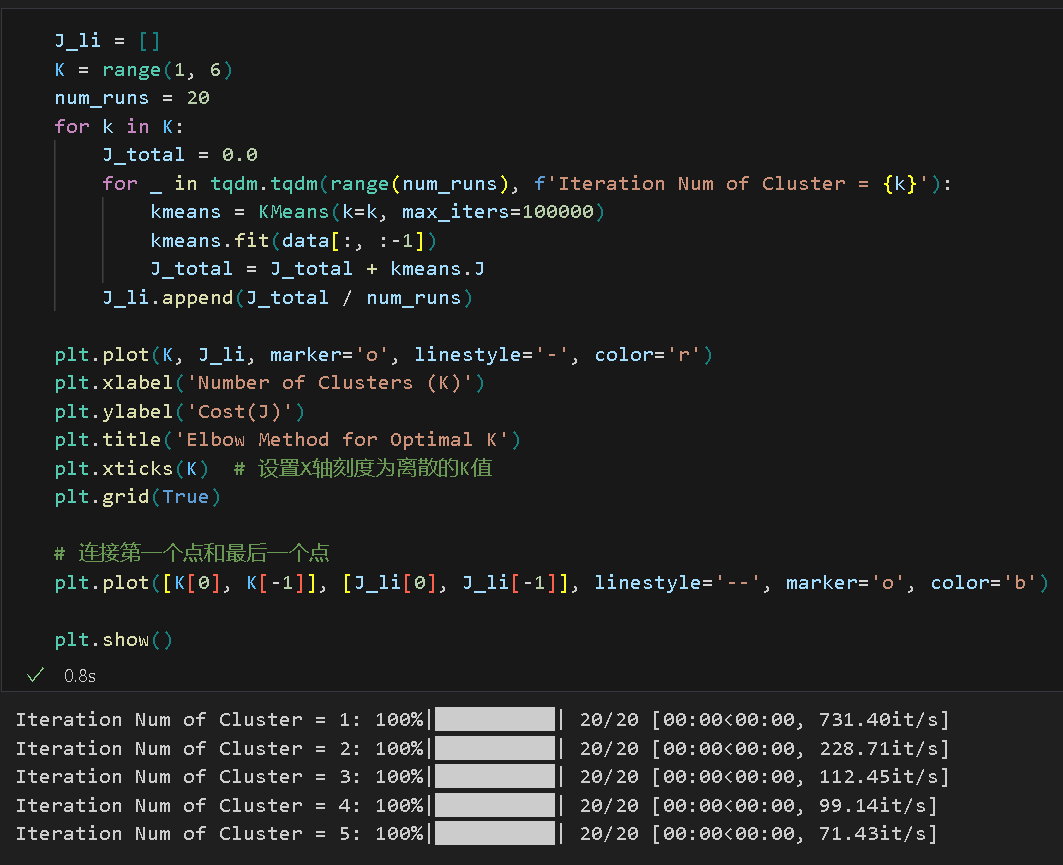
****

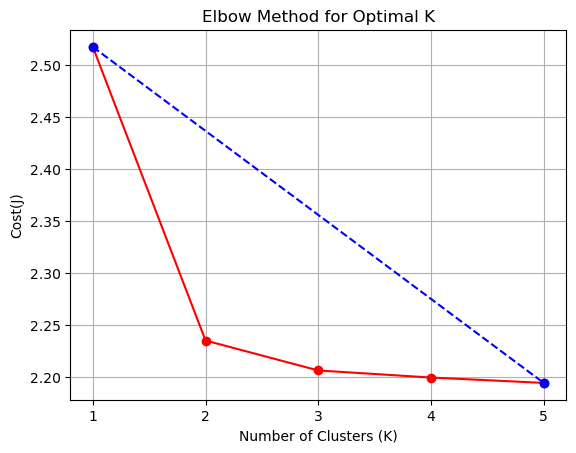
****

****

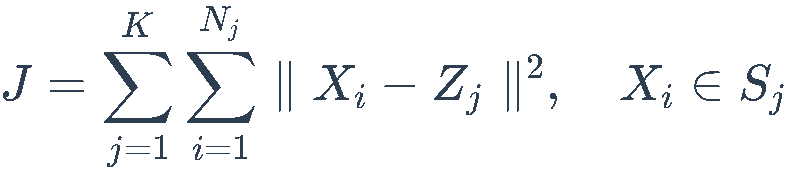
尝试将聚类个数调整成3、4、5，并可视化结果。

1. **用手肘方法判断最优K值，得K=2时效果最佳**



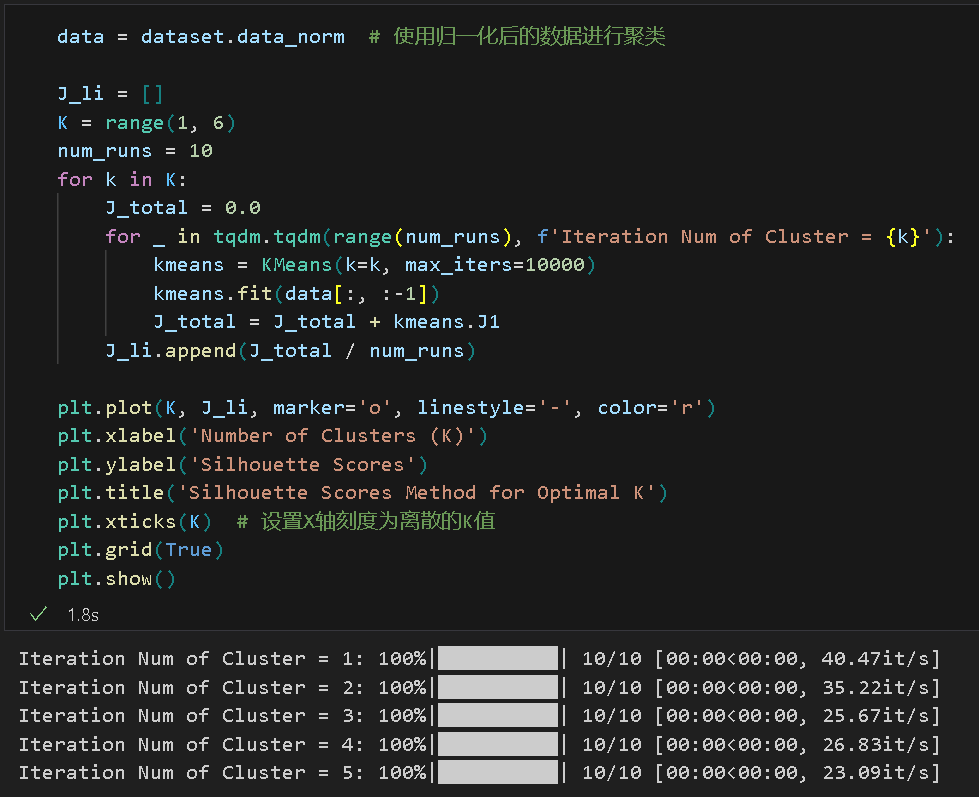
****

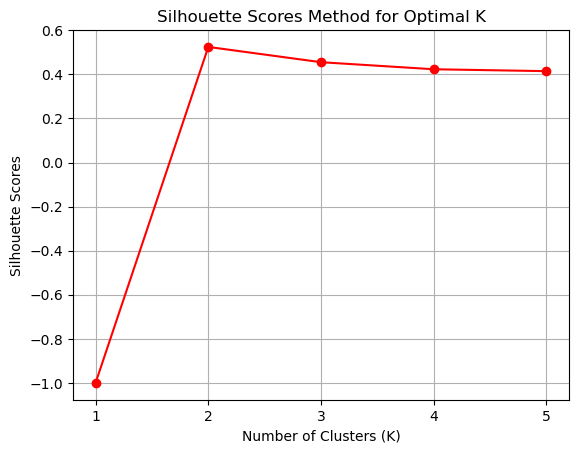
聚类算法的指标函数如下：



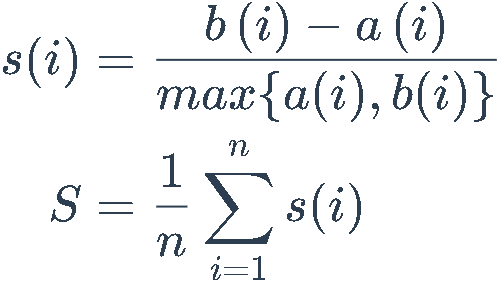
为减小每次迭代产生误差的影响，通过重复20次相同实验来计算k=1...5时聚类指标J并做可视化，从图中可知，当k=2时，距离直线段最近，故k=2时效果最佳。

1. **同时进行轮廓系数的可视化，得k=2时最优**





轮廓系数S计算公式如下：



为保证测量的准确性，我尝试使用轮廓系数来进行来一次测量，进行10次重复实验，并做可视化。由折线图可知，当k=2时，轮廓系数最高表明聚类效果最好。