西南科技大学

计算机实验报告

课程名称：数据挖掘应用技术与实践

实验名称： K-means算法

学 号：5120180269

学生姓名：李若昊

班 级： 计科1803

指导教师：吴珏

评 分：

实验日期： 2020年10月17日

实验目的：

通过分析K-means聚类算法的聚类原理，利用编程工具编程实现K-means聚类算法，并通过对样本数据的聚类过程，加深对该聚类算法的理解与应用过程。

实验内容：

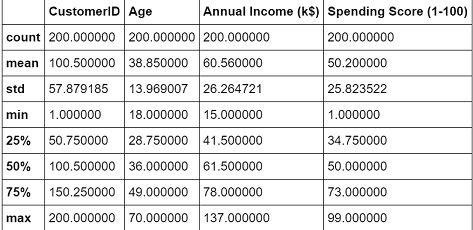
1. 分析K-means聚类算法；
2. 分析距离计算方法；
3. 分析聚类的评价标准；
4. 编程完成K-means聚类算法，并 基于相关实验数据实现聚类过程；

实验要求：

从自己所选的数据集（见下）中，综合考虑“年收入”和“消费评分”两个指标，将消费者放于这两个维度中分类。

实验步骤：

[原始数据集](https://github.com/pavankalyan1997/Machine-learning-without-any-libraries/tree/master/2.Clustering/1.K_Means_Clustering)（注：来自SuperDataScience网站）：



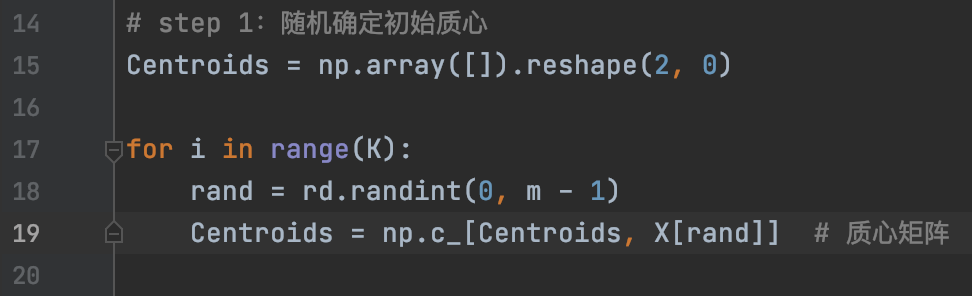
\* 为了最后结果展示方便理解和描述，仅取“年收入”和“消费评分”两列使用。

0. 先设置必要的两个初始值：迭代步数和簇数

a) 先选取一个直觉认为应该可以让算法收敛的迭代步数，当然较优的步数值需要多次尝试。这里先假设为100。

b) 再设置合理的簇数（有多种方法可以辅助判断多少簇合适，如肘部算法），这里假设为5簇。

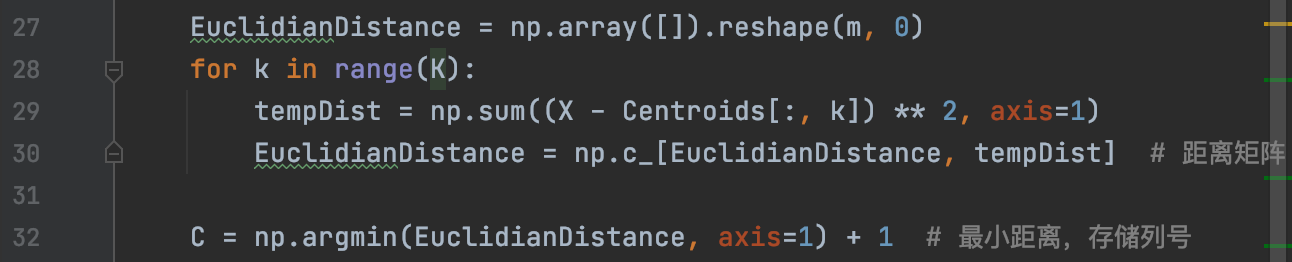
1. 现在可以开始算法了。算法初始从所有样本点中随机选簇数个点作为初始质心们，存放于一个n\*K维的矩阵中，其中每一列是一簇的质心。



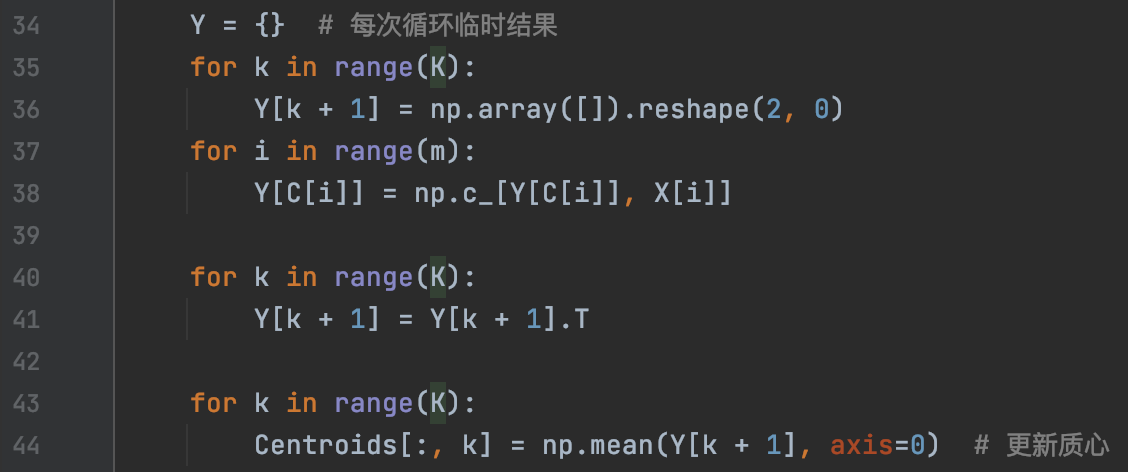
2. 分两步：

a) 对于每一个样本点，我们都算它距离各个初始质心的欧式距离，将结果存放于一个m\*K维矩阵中。其中每一行对应一个样本点，这一行的各列（共5列）分别对应距离各个质心的远近。

然后，对每一个样本点，选取距离它最近的质心所代表的簇作为它当下所归属的簇。将簇标号存于一个临时向量C中，每次迭代将会更新它。



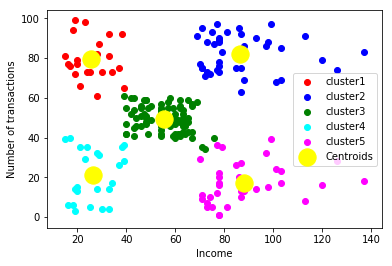
b) 根据刚才得到的向量C重新给样本点们分组，分组结果记入临时结果字典Y。同时还要对新的各个簇算簇内所有样本点的均值，并更新质心为此均值（对应的点）：



3. 重复上述步骤2共迭代数次：



4. 利用现成的库将聚类结果简易可视化：



其中，

* 红色的一簇表示低收入但消费频率高的人群
* 青色的一簇表示低收入且消费频率低的人群
* 绿色的一簇表示中等收入且消费频率相对适中的人群
* 品红色的一簇表示高收入但消费频率低的人群
* 蓝色的一簇表示高收入且消费频率高的人群

实验思考：

1. 考虑选择不同的距离函数，对结果的影响。

不同的距离函数适合不同类型的数据，故根据数据特性的不同，应该采用不同的度量方法。

1. 评价函数的计算。

实验体会：

在上学期的机器学习课程中已经利用一些工具直接使用过该算法，现在我知道了算法背后的原理，可以帮助我在应用时更好地理解结果和参数的意义。

感谢吴老师的悉心指导。