教程 2 : K-means

提交日期：2020.7.18

提交人：詹紫琦

目录

[1. 题目 3](#_Toc2061552594)

[1.1无标签数据集上的聚类方法K-means 3](#_Toc1320375436)

[1.2 运行环境 3](#_Toc850114103)

[2. 算法阐述或实验步骤说明 3](#_Toc486833554)

[2.1变量说明 3](#_Toc44406358)

[2.2算法阐述 3](#_Toc2021832392)

[2.3 实验步骤说明 4](#_Toc1433194579)

[3. 实验结果与截图 5](#_Toc722759633)

[3.1 随机初始化簇类中心点的值的K-means实验结果 5](#_Toc323276968)

[4. 总结 11](#_Toc866465257)

[4.1 总结 11](#_Toc787472945)

[5. 参考文献 11](#_Toc1148868495)

# 题目

## 1.1无标签数据集上的聚类方法K-means

1. Means[1]是一种聚类算法，从之前的监督学习中的分类方法中转换到无监督学习中的聚类算法，分类与聚类之间既有联系又存在着差别，分类其实是从特定的数据中挖掘出有用信息，最后根据该有用信息对新输入的信息作出判断的过程。聚类的目的也是将数据进行分类，但是在收集到原始数据之时，收集人员不会得到任何关于数据类别的有用信息，完全是通过算法来判断数据中各条数据之间的相似性，在聚类产生结果之前，无法根据数据表面信息对数据进行有效分类。在无监督学习中，存在着多类聚类算法，其中K-means方法是一种比较简单，易于理解，适用于连续型数据，而且运算速度较快，并且需要事先确定类别数，而类别数的确定一般是根据人工经验或者时现实项目需求等方面来确定类别数。

## 1.2 运行环境

系统：Ubuntu20.04，python3.7，Anaconda集成工具Jupyter编写。

# 算法阐述或实验步骤说明

## 2.1变量说明

：类别数，事先由人为定义，代表将该数据集所分的类别种数

：该算法中所应用的数据集，总共有m项

：代表K类中心点的K个不同值

：对应数据集中每一项对应的类别种数

：对应该算法的损失函数

## 2.2算法阐述

首先根据已知类别数K，进行随机初始化簇中心点的值，被随机初始化得到一个初始值。然后进行以下步骤：

Repeat{

For i =1 to m

Ci更新为对应与第i项数据欧式距离最近的类别k的值

得到了Ci值后，然后更新，更新公式如公式2-1。

（2-1）

其中是对应第K类中的数据散点。

}（minimize（J））

当迭代至损失函数J收敛至较小值时，K-means算法至此执行完毕。损失函数的表达式如公式2-2

（2-2）

## 2.3 实验步骤说明

首先读取数据集，初步观察无监督数据的散点分布，如图3-1，然后编写出一次迭代的K-means主要功能函数，接下来编写计算损失函数的功能函数。最后设置K-means的LOOP次数，即进行多少次K-means，进行递归，得到聚类结果与每进行一次K-means操作后的损失函数值结果，最终进行实验结果分析与结论得出。

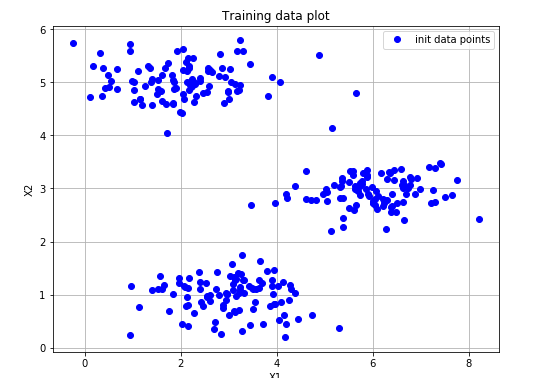


图3-1初步分析数据，每一个数据项的维数是2

初步观察数据，发现聚类的类别数大致上可以观察出为3，即K为3。

# 实验结果与截图

## 3.1 随机初始化簇类中心点的值的K-means实验结果

进行第一次随机初始化簇类中心点的值，得到的初始簇类中心点的分布如图3-1初始簇类中心点的值分别为：第一类簇中心点值（3.04367119,1.01541041），第二类簇中心点的值为（1.95399466,5.02557006），第三类簇中心点的值为：（6.03366736,3.00052511）。

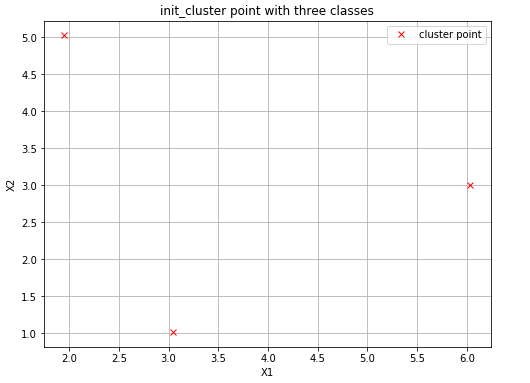


图3-1 第一次随机初始化簇类中心点的分布

根据该初始簇类中心点的值进行第一次K-means聚类，得到的结果如图3-2：

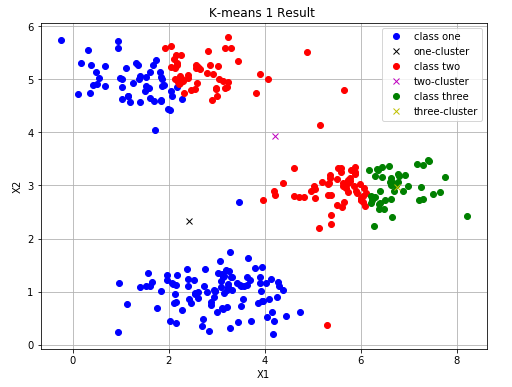


图3-2 第一次K-means结果

然后进行第二次K-means聚类，得到的结果如图3-3：

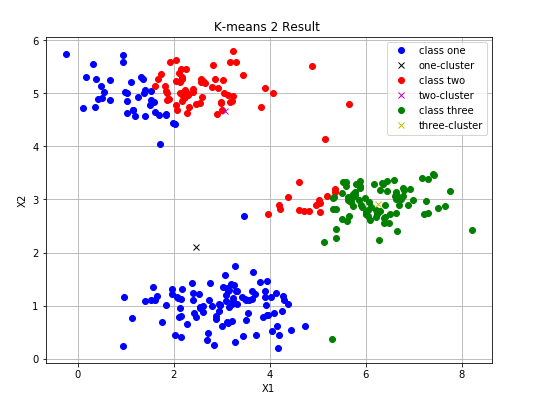


图3-3 第二次K-means结果

发现聚类效果还不理想，继续进行迭代，第三次K-means结果如图3-4：

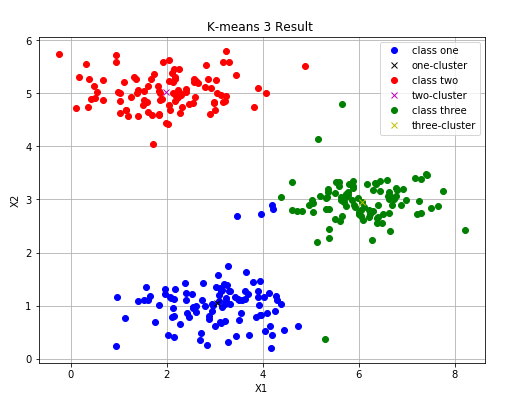


图3-4 第三次K-means结果

观察第三次K-means结果发现聚类效果还比较好，继续进行第四次K-means，试验结果如图3-5：进行完第四次聚类后，发现只有少数点会变化类别，观察第五次，第六次聚类结果，发现基本无变化，得到的该随机初始化下的簇类中心点计算出的损失函数如图3-6。



图3-5 第四次K-means结果

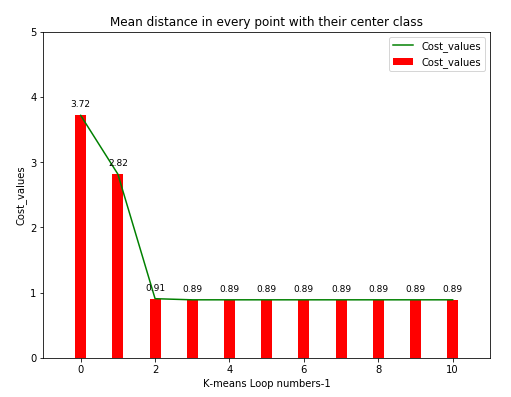


图3-6 第一次随机初始化K-means的损失函数大小随K-means迭代次数变化图

观察图3-6，我们可以发现损失函数的值最终是固定为0.89。由于一次随机初始化的簇类中心点的值不具有普适性，而且很容易得不到全局最小值，而是很容易得到局部最小值。所以进行多次随机初始化，第二次，第三次，第四次随机初始化后的损失函数值的示意图如图3-7，3-8，3-9。

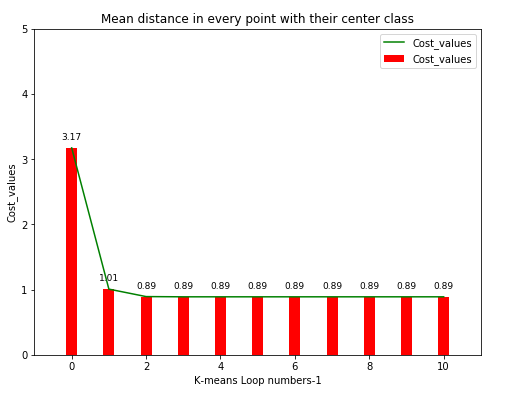


图3-7第二次随机初始化簇类中心值损失函数的变化

观察四次随机初始化簇类中心值损失函数的变化图可以看出，通过K-means聚类方法在该数据集上进行聚类的结果已经得出，能使损失函数收敛到一个全局最小值，即0.888861732，约为0.89，该值便为k-means聚类方法在该数据集上损失函数的全局最小值。

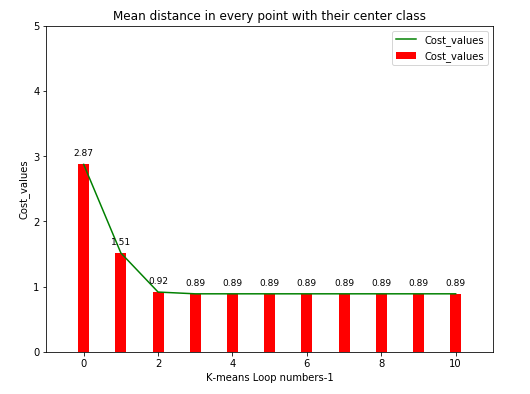


图3-8第三次随机初始化簇类中心值损失函数的变化

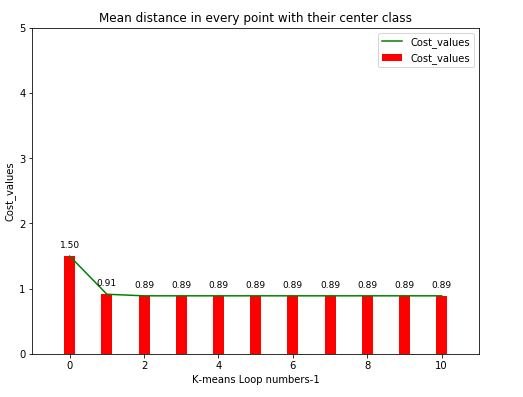


图3-9第四次随机初始化簇类中心值损失函数的变化

# 总结

## 4.1 总结

通过多次随机初始化簇类中心点的值的K-means方法，同时经过多次K-means迭代，能得到出拥有最小损失函数值的聚类结果。通过该次实验能发现出，K-means方法适用于数据集连续，而且比较简单的数型数据，即K-means方法能拥有较快的处理速度，以及较少的迭代次数，通过很少次数的迭代就能使损失函数收敛到全局最小值。但K-means方法存在着显著的局限性，需要提前设置期望所分类别数，在本次实验中所设置的所分类别数便为3，缺少一种自主确定类别的机制，需要人工甚至是初始观察数据的经验来设置类别数的大小，这便是K-means方法的局限所在。

# 参考文献

[1] 周志华等.机器学习（西瓜书）