## 大数据分析实验报告——聚类

### 苏致成 201250104

### 注明：

本文档采用markdown书写，因此，转为word部分公式可能出现显示异常等问题。

### 使用方法

#### SimpleKMeans

聚类属于无监督学习，KMeans聚类是最基础常用的聚类算法，其基本思想是，通过迭代寻找k个聚簇的一种划分方案，使得聚类结果对应的损失函数最小。其中，损失函数可以定义为各个样本距离所属聚簇中心点的误差平方和。即：

其中， 代表第 个样本， 是 所属的聚簇， 代表聚簇对应的中心点， 代表样本总数。

核心部分即为先固定中心点，调整样本所属类别来减少 ;再固定每个样本类别，调整中心点继续减小 。两个过程交替循环， 单调递减直到最小值，中心点和样本划分的类别同时收敛。

##### 算法流程

1. 数据预处理。包括标准化、异常点过滤等。
2. 随机选取k个中心，记为 。
3. 定义损失函数： 。
4. 令 为迭代次数，重复如下过程直至 收敛：
   * 对于每一个样本 ，将其分配到距离最近的中心，即：

* 对于每一个类中心 ,重新计算该类的中心。即：

#### EM

EM算法是一种迭代优化策略，它的每一轮迭代过程都分为两步，一个是期望步（E步），另一个是极大步（M步）。基本思想是根据观测值，估计出模型参数的值，然后再根据上一步估计出的参数值估计缺失数据的值，然后再根据估计出的缺失数据加上观测值重新再对参数值进行估计，反复迭代，直至收敛。

缺点：对初始值敏感，聚类结果随不同的初始值波动较大。

##### 算法流程

1. 输入观察的数据 ，联合分布 ,条件分布 ,最大迭代次数 。
2. 随机初始化模型参数 的初值 。
3. 令 开始 算法迭代步骤：
   * E步骤：计算联合分布的条件概率期望：

$$Q\_{i}(z\_{i})=p(z\_{i}|x\_{i},\theta\_{j})\\
l(\theta,\theta\_{j})=\sum\_{i=1}^{n}\sum\_{z\_{i}}Q\_{i}(z\_{i})log\frac{p(x\_{i},z\_{i};\theta)}{Q\_{i}(z\_{i})}$$

* + M步骤：极大化 ,得到 ：
  + 若 已经收敛，则算法结束。否则继续进行E步骤和M步骤进行迭代。

### 数据集处理思路

#### SimpleKMeans

1. 缺失值处理。将含有缺失值的条目均去除。
2. 读入 ，调用 ,并且将其聚类的数量设置为3（注：在KMeans中聚类的数目k的值一般比较难以确定，此处确定为3的部分原因是已知其结果的可靠性）。
3. 输出结果。

#### EM

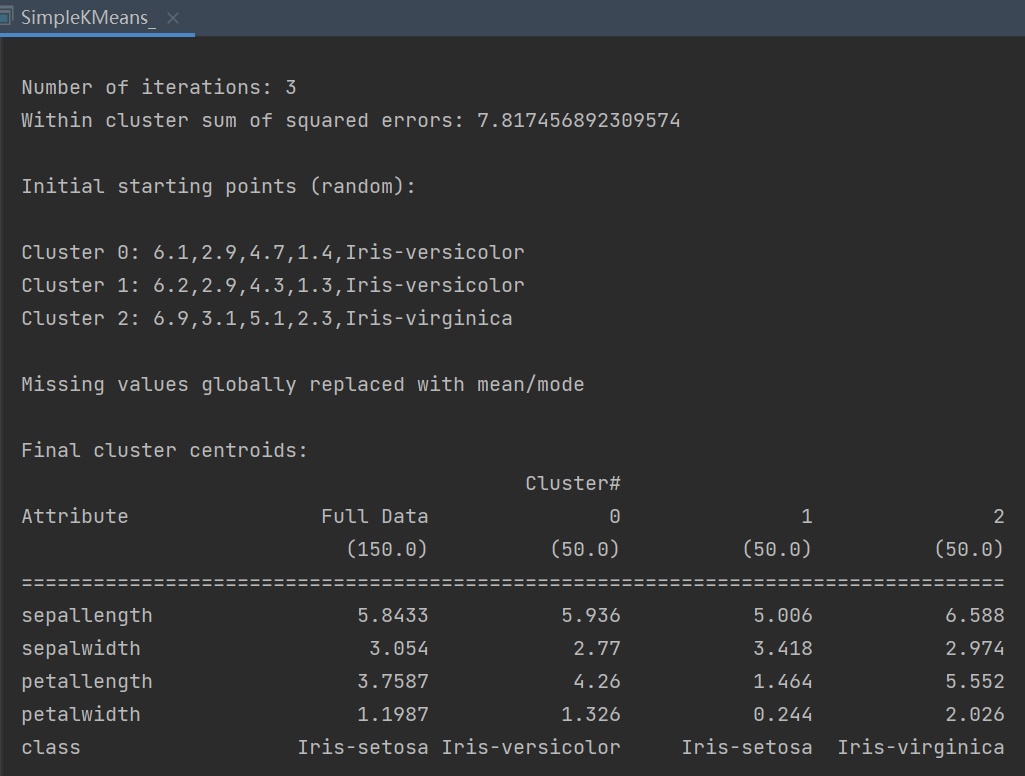
1. 缺失值处理：观察数据集可知，并无缺失数据值。
2. 读入 ，调用 将最后一列属性不予纳入聚类考虑。
3. 设置最大迭代次数为100，并且聚簇的数目是3。
4. 打印出对数似然度量等值。

### 实验结果

#### SimpleKMeans

观察到KMeans方法正好分为3种类别的鸢尾属植物。

可以观察到诸如花瓣宽度方面， 明显比 高等指标信息。



#### EM

从下图可知每个类的分类聚簇后各类的数量、各项指标等信息。

可以观察到诸如花瓣宽度方面， 明显比 高等指标信息。

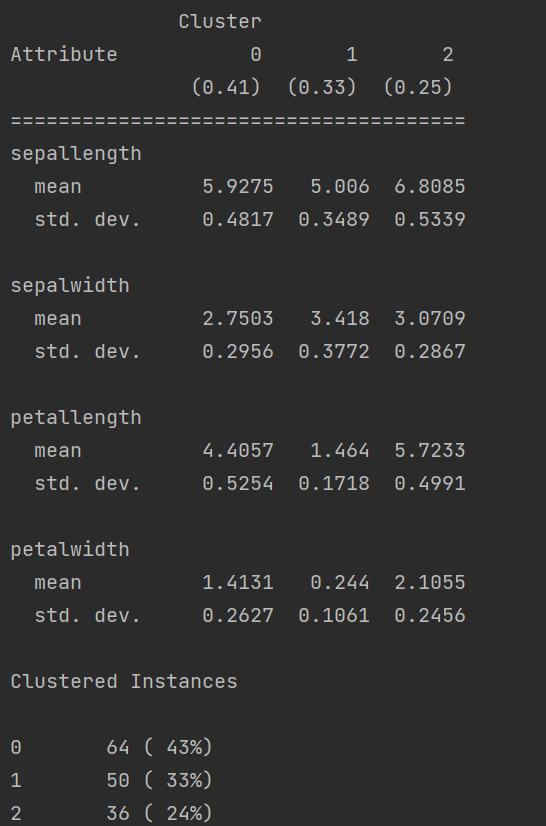


image-20221010115643829

并且根据下图可知，聚类过程很好地将三个子类进行了划分，并且错误率（9.3333%）较低。

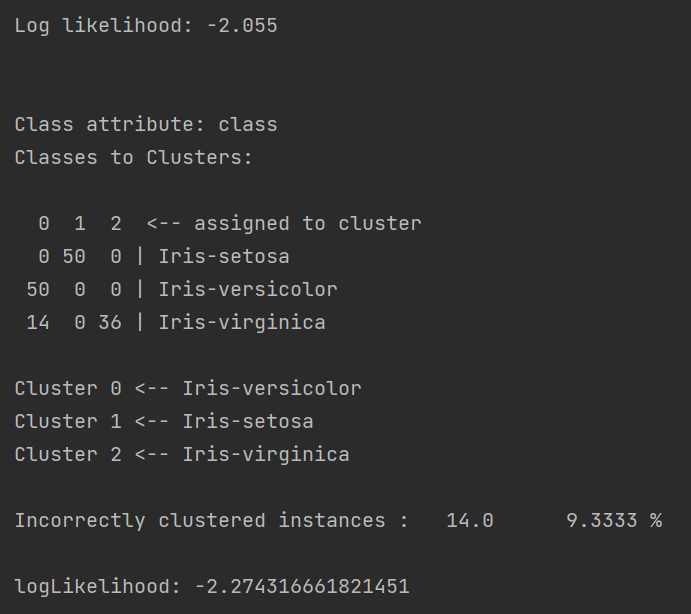


image-20221010115749967