**《大数据分析》第五次实验报告**

201250070郁博文

1. **K-means聚类**
2. **K-means聚类简介**

K-means是一种基于欧氏空间距离进行聚类的方法。它是一种无监督学习。

1. **个人理解**

K-means的目标就是把给定的数据集划分为K个簇，K为预先设置的超参数。这个算法也十分好理解。

在聚类开始前，我们先随机选取K个初始点作为中心点，然后，在剩下的点中，每次选取一个点，计算该点到K个中心点的欧式距离，选取欧氏距离最近的那个初始点，那么该点就与初始点为同一聚类。在这之后，我们需要更新中心点。具体做法就是在坐标系中找到一个点，使得该点到旧中心点和新加入的点的距离最小。重复以上过程，最终每个点到它对应的中心点距离最小，就可以得到K个簇。实际上，如果我们定义如下的损失函数：



那么上述过程也是该损失函数收敛的过程。

K-means尽管是一个十分强大的无监督学习算法，但是任然存在许多问题。我们可以发现，超参数K的取值完全是随心所欲地，而K-means聚类结果的好坏很大程度上是超参数K决定的。这就给算法带来了很大的不确定性：我们应该如何确定K，使得算法结果更好呢？在查阅相关资料后，笔者了解到了一个名为Gap Statistic的方法。记损失函数为 Dk ，当分为K类时，Gap Statistic定义为： Gap(k)=E(logDk)−logDk 。 E(logDk) 是 logDk 的期望，一般由蒙特卡洛模拟产生。我们在样本所在的区域内按照均匀分布随机地产生和原始样本数一样多的随机样本，并对这个随机样本做KMeans，得到一个 Dk ，重复多次就可以计算出 E(logDk) 的近似值。

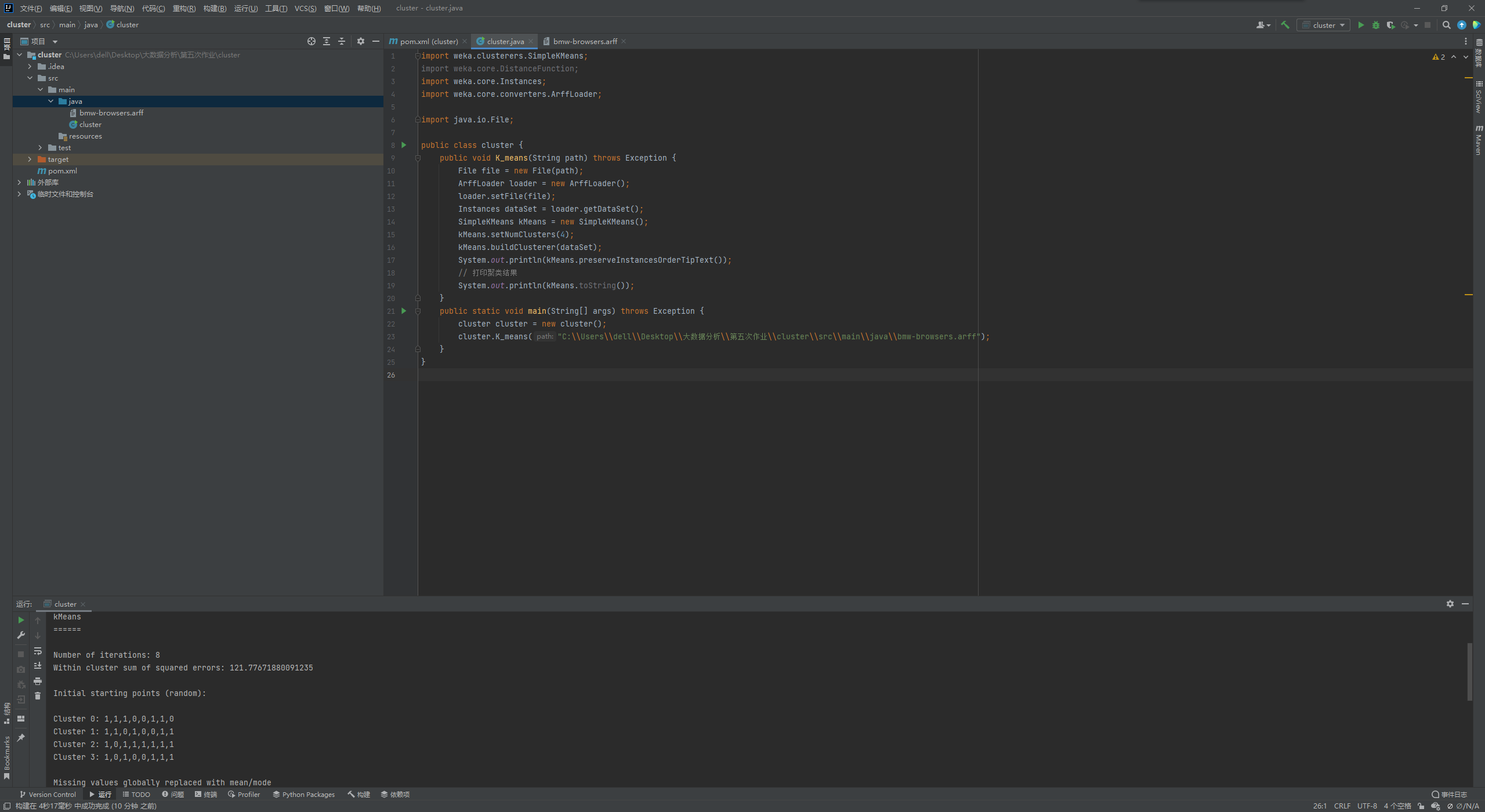
Gap(K) 的物理含义是随机样本的损失与实际样本的损失之差。Gap越大说明聚类的效果越好。一种极端情况是，随着K的变化 Gap(K) 几乎维持一条直线保持不变。说明这些样本间没有明显的类别关系，数据分布几乎和均匀分布一致，近似随机。此时做聚类没有意义。

同时，初始值的选择也带来很大的不确定性，如果初始值本身距离就相近，那么无疑会花费更多迭代次数算法才能收敛，因此在选择初始值时，可以使这些点尽可能远离，这也是K-means++算法的思路。

1. **实验结果**

编写的实验代码如下：

这个类是K-means的实现。



设置聚类结果为4类，运行结果如下：

