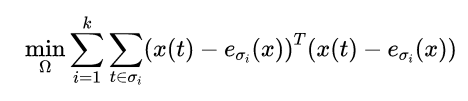
**《系统工程导论》K-means聚类分析**

1. **请简要证明k-means为何会收敛。k-means一定会收敛到最优值吗？为什么？**

答：

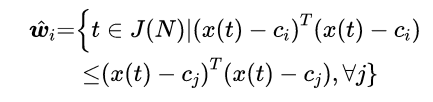
(1)收敛性证明：

1. menas的优化函数如下：



易知该目标函数有下界（极限情况，0便是一个下界）

再考虑K-means算法的迭代过程：



该迭代过程保证了每次迭代，如果没有达到收敛值，原目标函数是严格下降的。

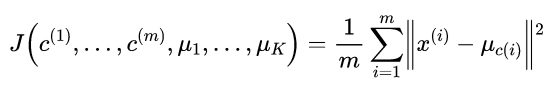
严格下降且有下界，则原目标函数必收敛。

1. K-menas不一定收敛到全局最优解，且在大多情况下都是局部最优解。

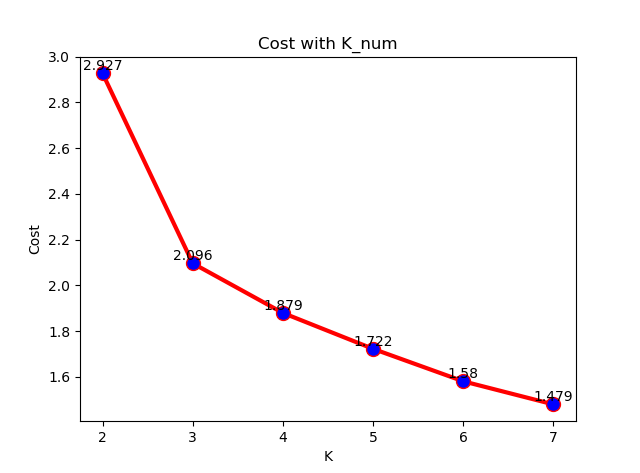
原因：首先，原目标函数为非凸函数，同时其有多个局部最优解。根据上面的证明，算法是严格单调下降的，因此，在初始选定之后，其只能收敛到其对应的局部最优解。且不能产生任何跳出该局部的解。因此K-means不能收敛到全局最优解，且最终解的局部最优解和初始值的选取有关。

1. **K-menas聚类实验**
   1. **确定聚类数目**

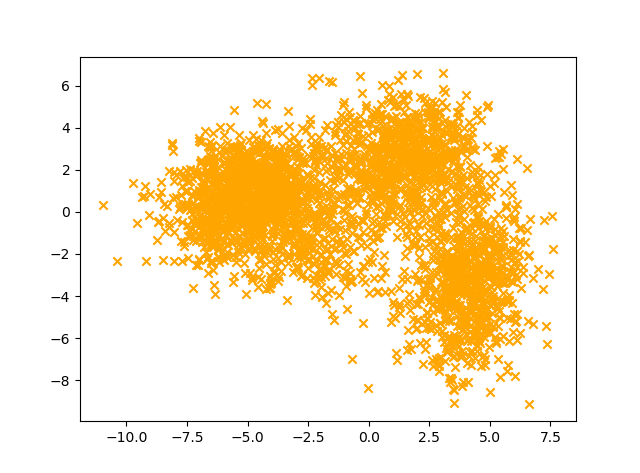
此处使用了吴恩达机器学习当中K-menas种类的确定——**“肘部法则”**。即选择cost function点的cost明显变缓慢的K。其中，cost function如下：



本次实验，K从2开始取，直到7，其结果如下：



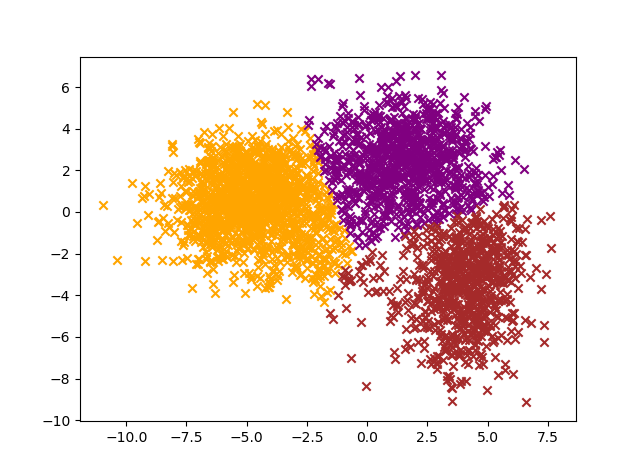
可见，当**K=3**时，cost显著下降，之后平缓下降，可知，该问题的K选择3比较合适。同时也可以从原始数据的分布看出：



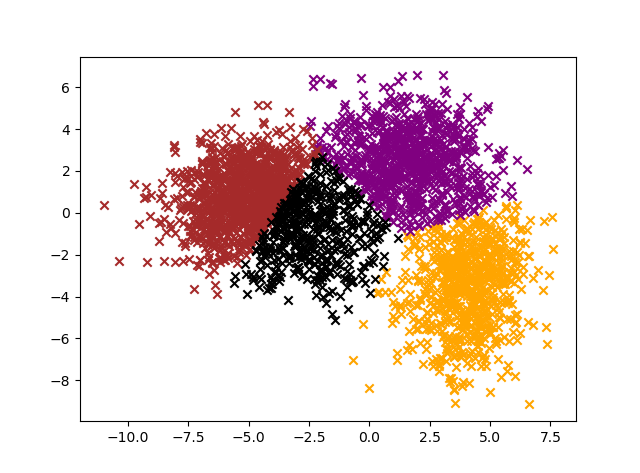
数据分为明显的3块。

**2.2 聚类实验结果**

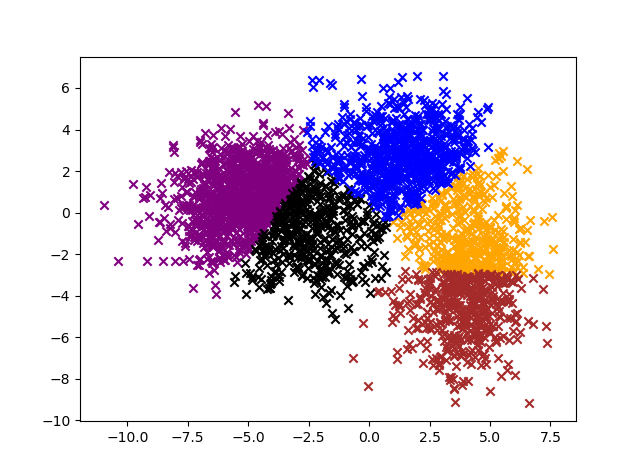
【K=3】



【K=4】



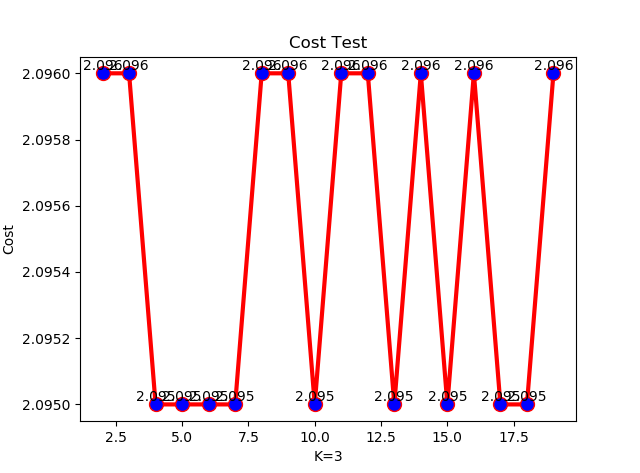
【K=5】



**2.4 选择不同初始点**

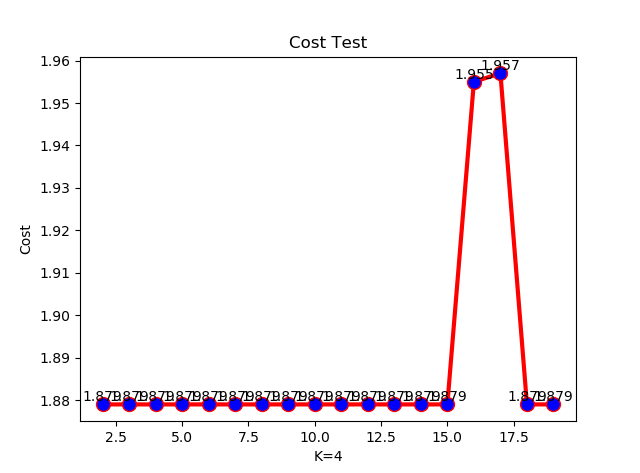
选择不同的初始点，进行试验，观察cost变化。

（1）K=3，进行18次实验，其cost变化如下：（**注**：此处横轴表示实验次数）

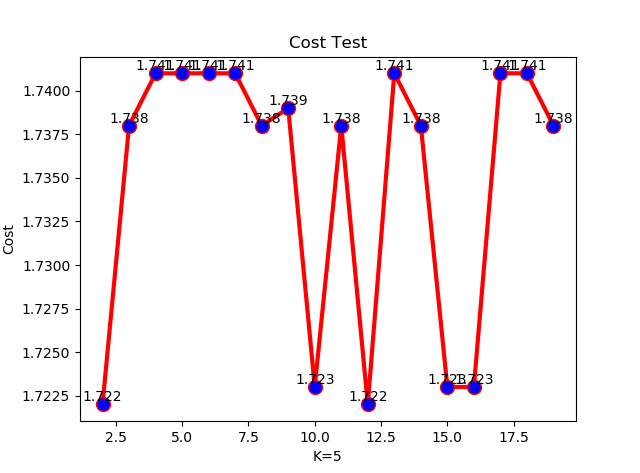


说明该问题不同的初始化会到达不同的聚类结果，即有不同的局部极小。同时，不同的初始点，其迭代次数差别也很大。

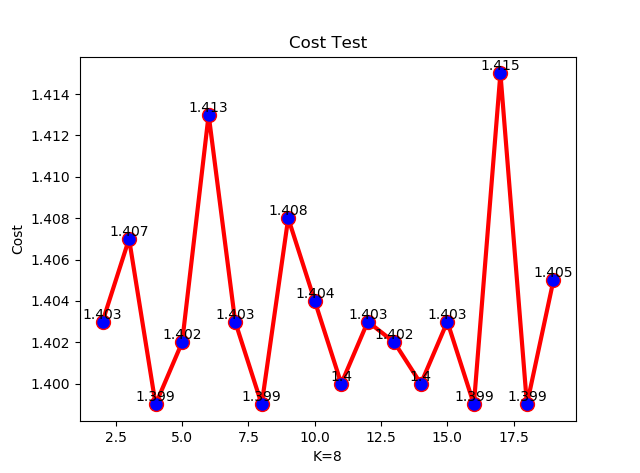
（2）K=4



1. K=5



1. K=8



通过上面的实验可以得出如下结论：

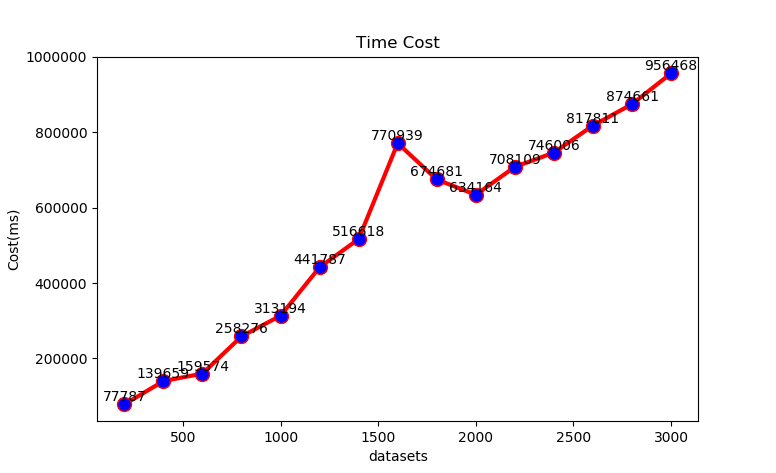
①不同的初始化会到达不同的极小值点；

②不同的初始化，即使到达了相同的局部极小值点，其对应的迭代次数差别也很大；

③聚类数目越多，对应的局部极小越多，越难取得全局最优。

**2.5 不同的数据规模，观察耗时和数据规模之间的关系**

使用定量观测的方法，即确定初始化的点（前num个），对不同的数据量进行聚类，聚类数为3。统计结果如下：(微秒为单位）



（\*注：横坐标为数据量，纵坐标为用时（单位微秒））

可见，在相同的条件下（CUP，初始化，系统等等），数据量越大，其对应的运行耗时越大。同时可以看到，耗时和数据量基本上呈线性关系。