《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：

K-menas 的优化函数如下：

$$\min_{\Omega}\sum_{i=1}^{k}\sum_{t\in\sigma_i}(x(t)-e_{\sigma_i}(x))^T(x(t)-e_{\sigma_i}(x))$$

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：

$$\begin{aligned}\hat{\boldsymbol{w}}_i=\Big\{t\in J(N)|&(x(t)-c_i)^T(x(t)-c_i)\\ &\le(x(t)-c_j)^T(x(t)-c_j),\forall j\}\end{aligned}$$

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。

(2) 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。

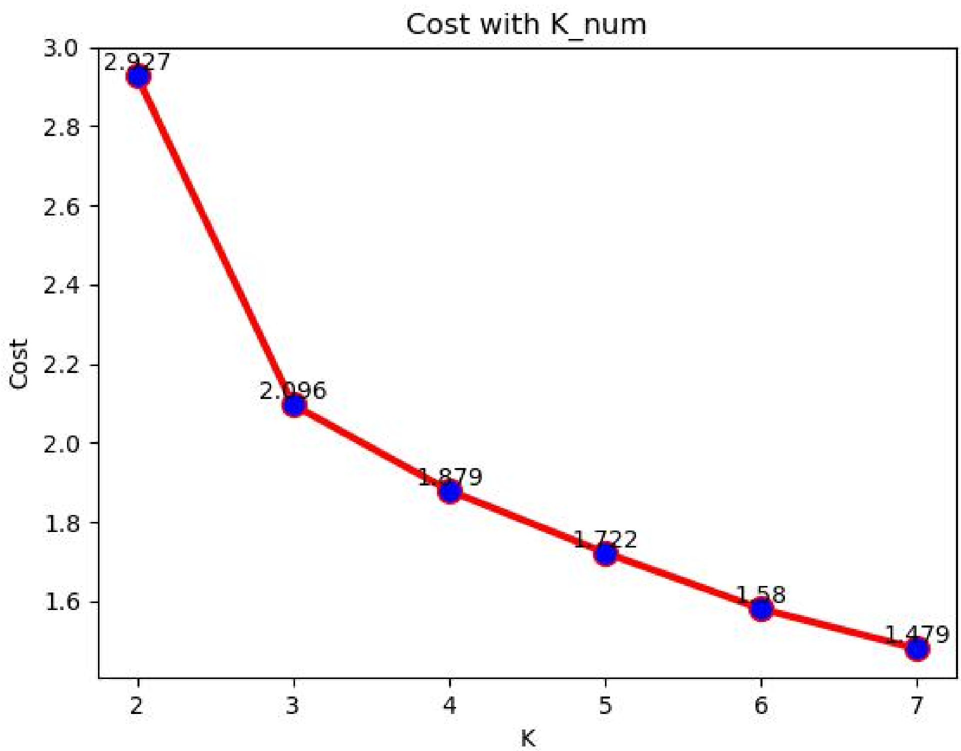
2. K-menas 聚类实验

2.1 确定聚类数目

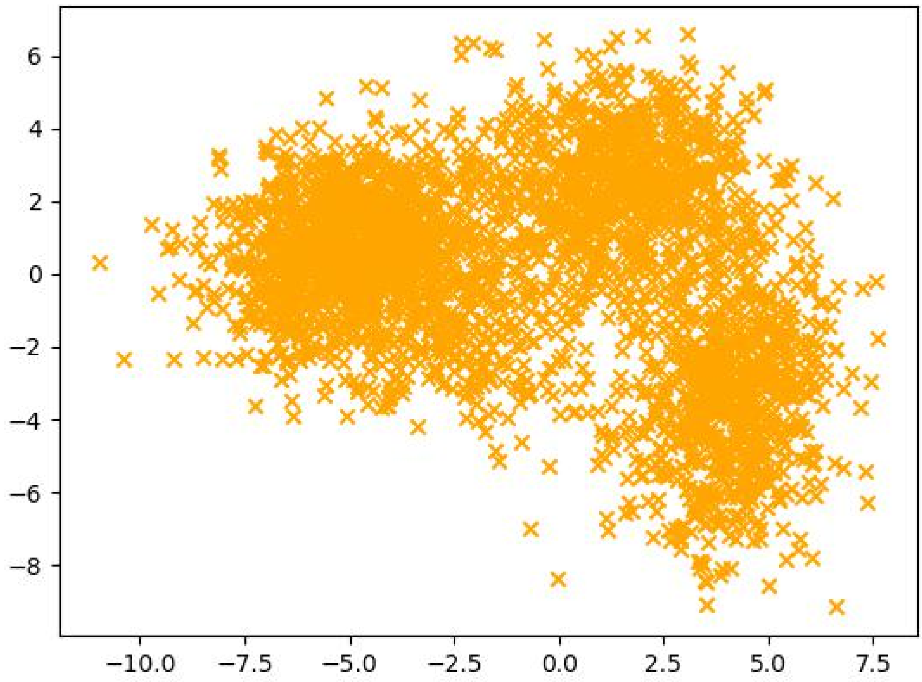
此处使用了吴恩达机器学习当中 K-menas 种类的确定——**“肘部法则”**。即选择 cost function 点的 cost 明显变缓慢的 K。其中，cost function 如下：

$$J\left(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K\right)=\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-\mu_{c(i)}\right\|^2$$

本次实验，K 从 2 开始取，直到 7，其结果如下：



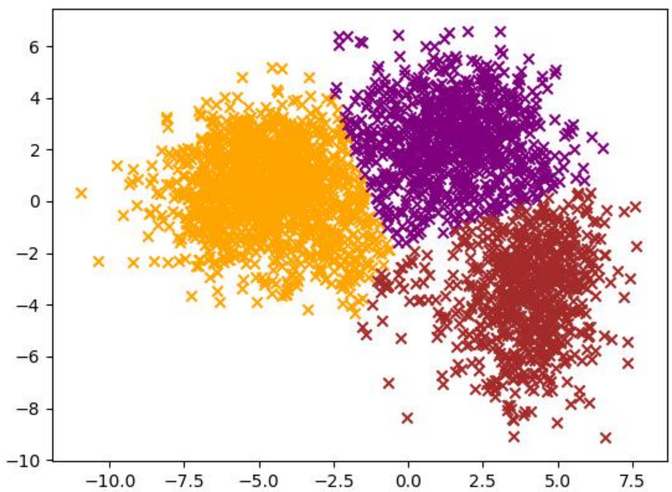
可见，当 K=3 时，cost 显著下降，之后平缓下降，可知，该问题的 K 选择 3 比较合适。同时也可以从原始数据的分布看出：



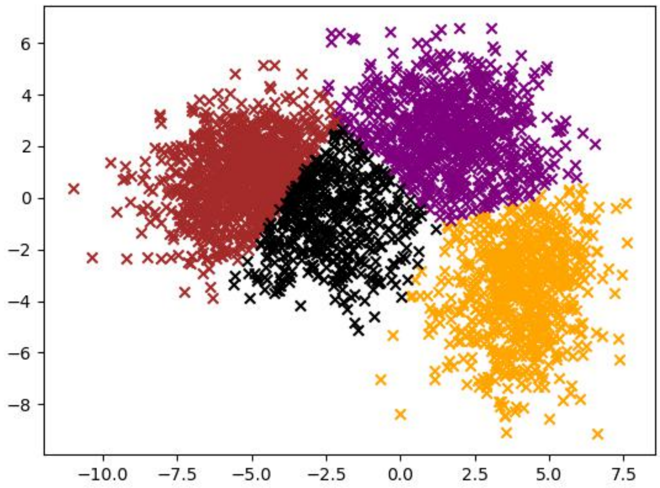
数据分为明显的 3 块。

2.2 聚类实验结果

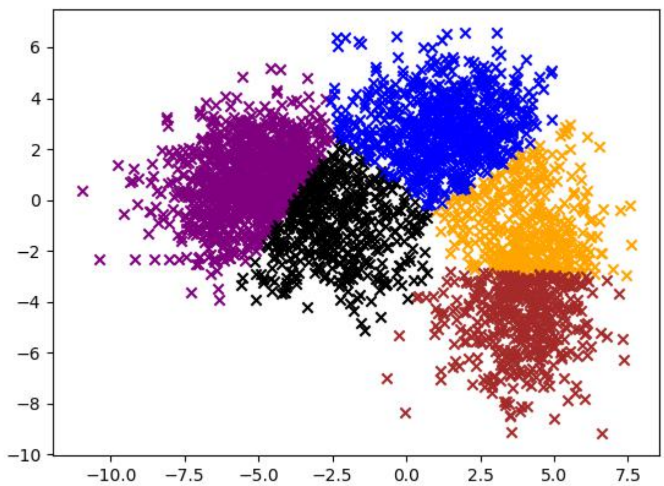
【K=3】



【K=4】



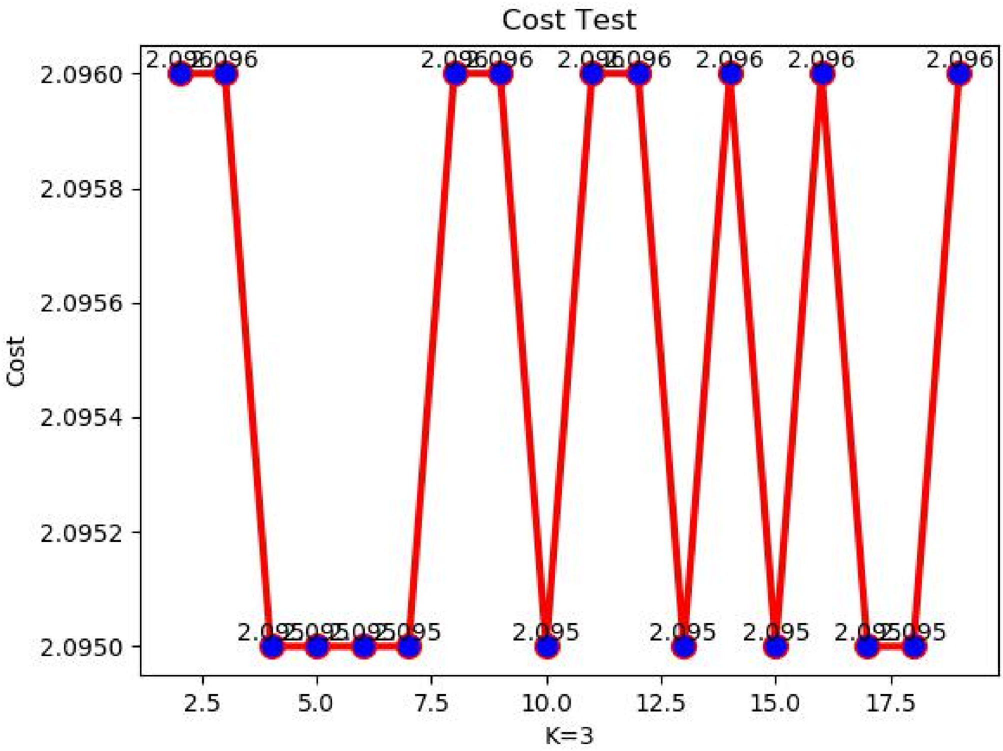
【K=5】



2.4 选择不同初始点

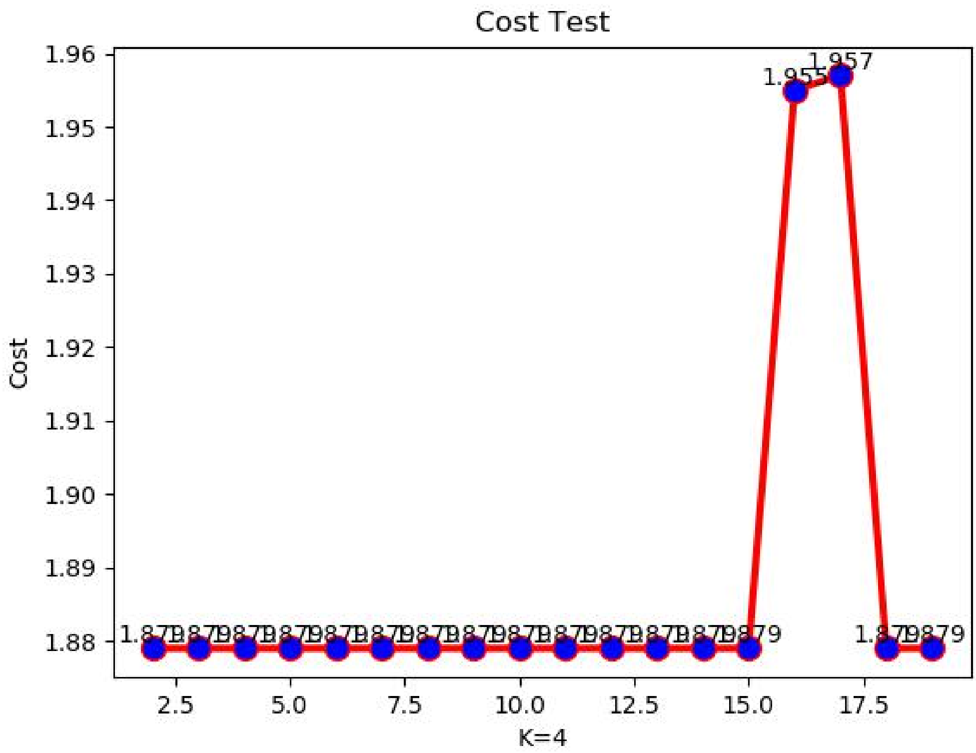
选择不同的初始点，进行试验，观察 cost 变化。

（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）

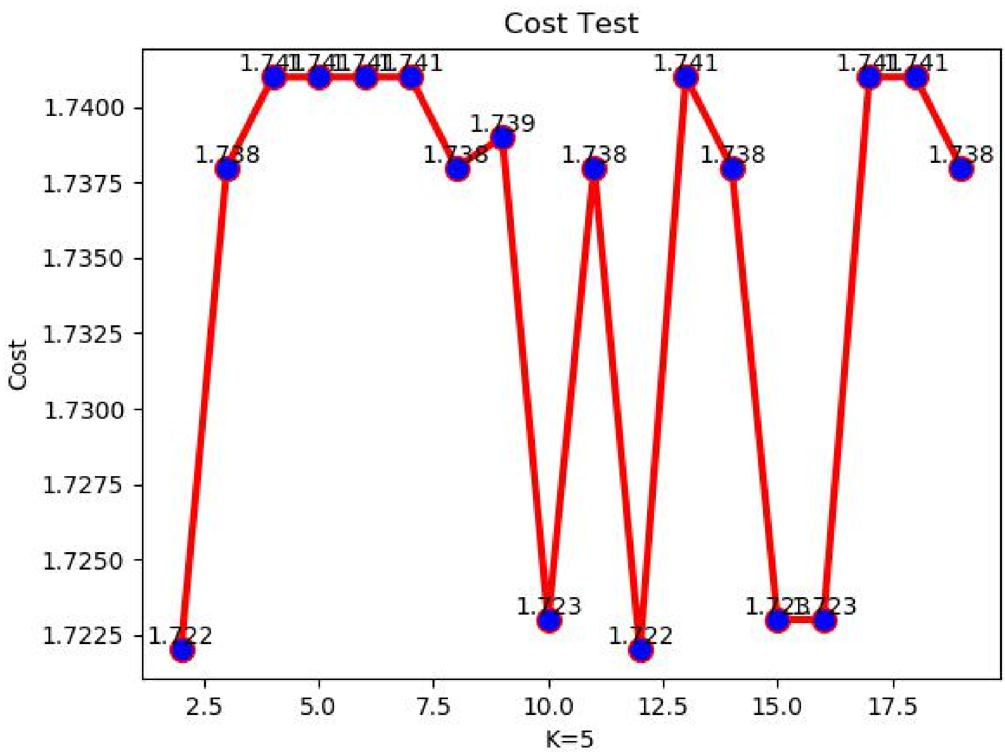


说明该问题不同的初始化会到达不同的聚类结果，即有不同的局部极小。同时，不同的初始点，其迭代次数差别也很大。

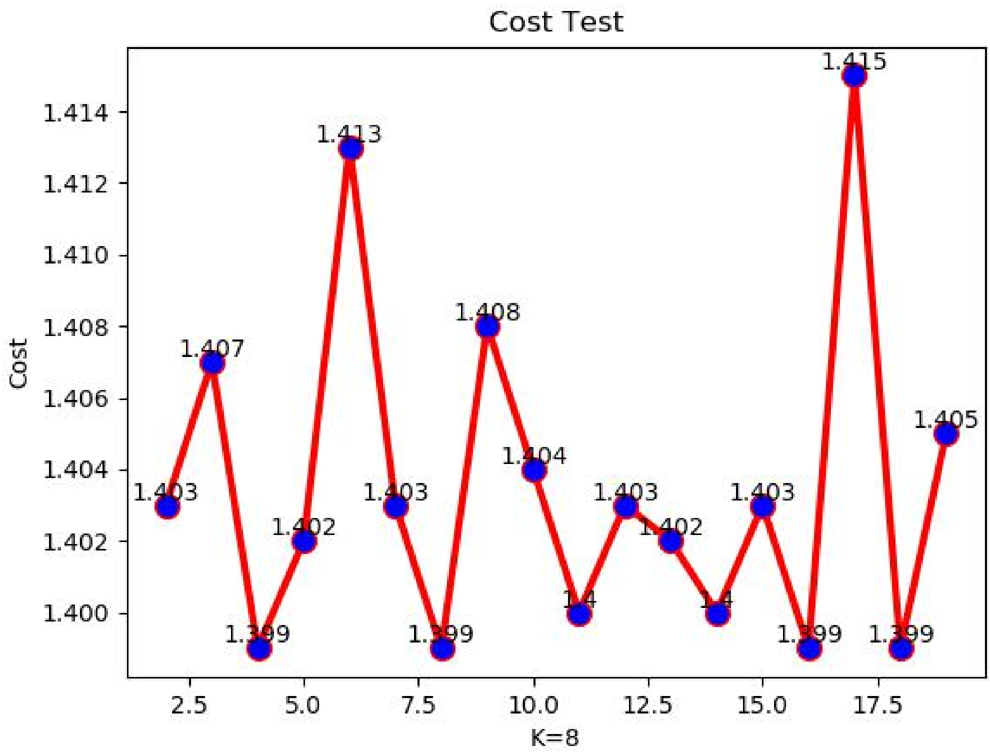
（2）K=4


Cost Test
Cost
K=4
1.955
1.957
1.879
1.96
1.95
1.94
1.93
1.92
1.91
1.90
1.89
1.88
2.5
5.0
7.5
10.0
12.5
15.0
17.5

(3) K=5


Cost Test
Cost
K=5
1.741
1.739
1.738
1.723
1.722
1.7400
1.7375
1.7350
1.7325
1.7300
1.7275
1.7250
1.7225
2.5
5.0
7.5
10.0
12.5
15.0
17.5

(4) K=8



通过上面的实验可以得出如下结论：

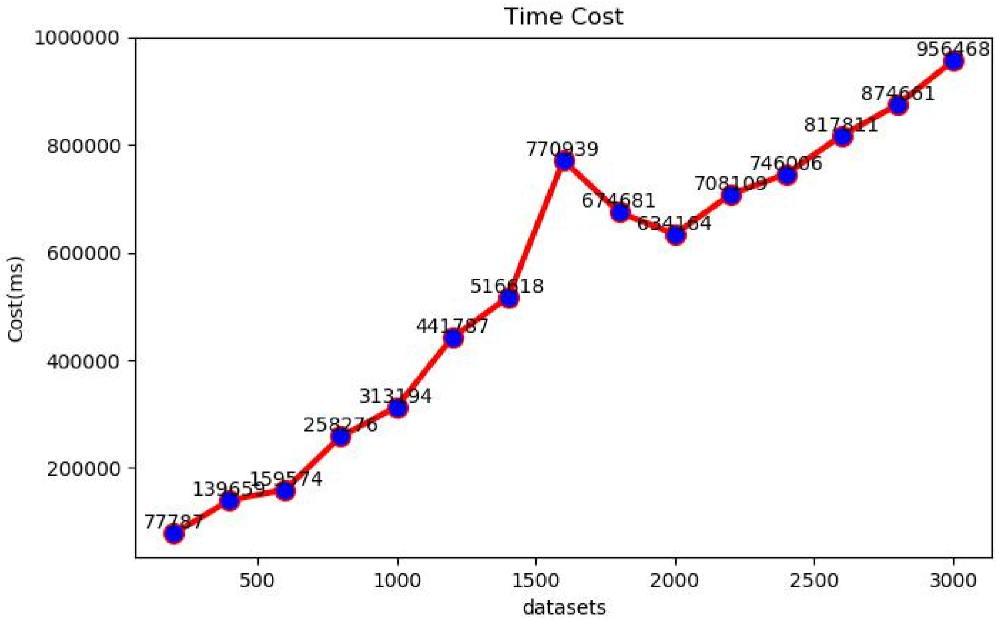
①不同的初始化会到达不同的极小值点；

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。