Kmeans 聚类作业报告

2017011010 杜澍浛 自 71

题目

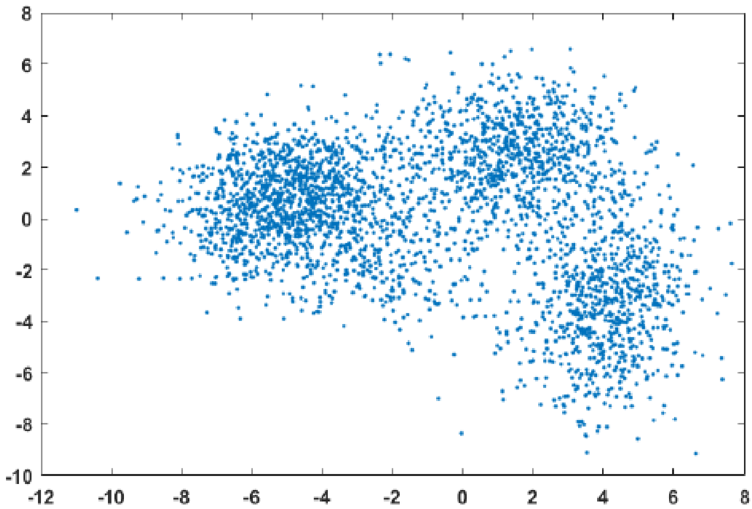
给定样本集合$\Omega = \{x_1, x_2 ... x_n\}$，其中每个样本都是$d$维的向量，k-means聚类的目标是将集合中的样本划分为k个类别，使得下述目标函数最小化：

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

其中$e_{\widetilde{\omega}_i}(x)$为第$\widetilde{\omega}_i$类的中心，即：

$$e_{\widetilde{\omega}_i}(x) = \frac{1}{|\widetilde{\omega}_i|} \sum_{t \in \widetilde{\omega}_i} x_t$$

附件 data.mat 中包含3000个二维平面上的点，请根据课堂所学知识，编写 k-means 聚类方法对这些点进行聚类。这些点的分布情况如下：



具体要求

(1) k-means 聚类一定会收敛吗？为什么？

(2) 完成函数 $\boldsymbol{function}\ label = \boldsymbol{kmeans\_clustering}(data, num)$，其中输入变量 $data$为 N 行 m 列，每一行为一个数据点， num 表示聚类数目；输出变量 $label$ 为 N 行 1 列，表示对应的数据点属于哪一类（比如属于第一类的点 $label$ 就为 1）

(3) 聚类数目从2类开始逐渐增加，分别进行计算并分析聚类效果，决定最合适的聚类数目并说明理由

(4) 选择不同的初始点多次实验，观察初始点的选择对最终结果的影响，并分析为什么会有这种影响

(5) 选择不同的数据规模进行实验，计算你的程序耗时，观察耗时与数据规模之间的关系，从中你能得到什么结论？提示：MATLAB 中可以使用 tic 和 toc 语句组合来计算某一段代码的耗时，具体可以查看帮助

【解答】

（1）k-means 聚类一定会收敛，这是因为：

假设有一组数据，变量为 n 维，在$t = 1,2,\dots,N$时刻，记为：

$$x(t) = [x_1(t) \quad x_2(t) \quad \dots \quad x_n(t)]^T, 1 \le t \le N$$

聚类的目标为：

$$\bigcup_{1 \le i \le k} \varpi_i = J(M), \varpi_i \cap \varpi_j = \phi, \forall i \ne j$$

$$\Omega = \{\varpi_i \subseteq J(M), 1 \le i \le k\}$$

并且使得下述目标函数最小：

$$\sum_{i=1}^{k} \rho(\varpi_i)$$

$$\rho(\varpi) = \sum_{t \in \varpi} (x(t) - e_\varpi(x))^T (x(t) - e_\varpi(x))$$

$$e_\varpi(x) = \frac{1}{|\varpi|} \sum_{t \in \varpi} x(t)$$

k-means 方法在一开始就固定了聚类的个数，假设初始聚类划分如下：

$$\Omega = \{\varpi_i \subseteq J(M), 1 \le i \le k\}$$

此时各聚类中心点的值为：

$$c_i = e_{\varpi_i}(x) \in R^n, 1 \le i \le k$$

对于每个点，根据其与各个中心点的距离将其划入与之最近的聚类：

$$\bar{\varpi}_i = \left\{ t \in J(M) \mid (x(t)-c_i)^T (x(t)-c_i) \le (x(t)-c_j)^T (x(t)-c_j), \forall j \right\}$$

更新后的聚类如下：

$$\bar{\Omega} = \{ \bar{\varpi}_i \subseteq J(M), 1 \le i \le k \}$$

进一步得到：

$$\sum_{i=1}^{k} \sum_{t \in \bar{\varpi}_i} (x(t)-c_i)^T (x(t)-c_i) \le \sum_{i=1}^{k} \sum_{t \in \varpi_i} (x(t)-c_i)^T (x(t)-c_i)$$

$$\sum_{i=1}^{k} \sum_{t \in \varpi_i} (x(t)-c_i)^T (x(t)-c_i) \le \sum_{i=1}^{k} \rho(\varpi_i)$$

可见每一次迭代后目标函数值不会变大，最终一定收敛到局部最优解。

（2）完整代码在 assignment6.py 文件中，kmeans_clustering(data,num)函数的步骤如下：

①根据聚类数随机挑选出各个聚类中心

②对所有数据点进行：

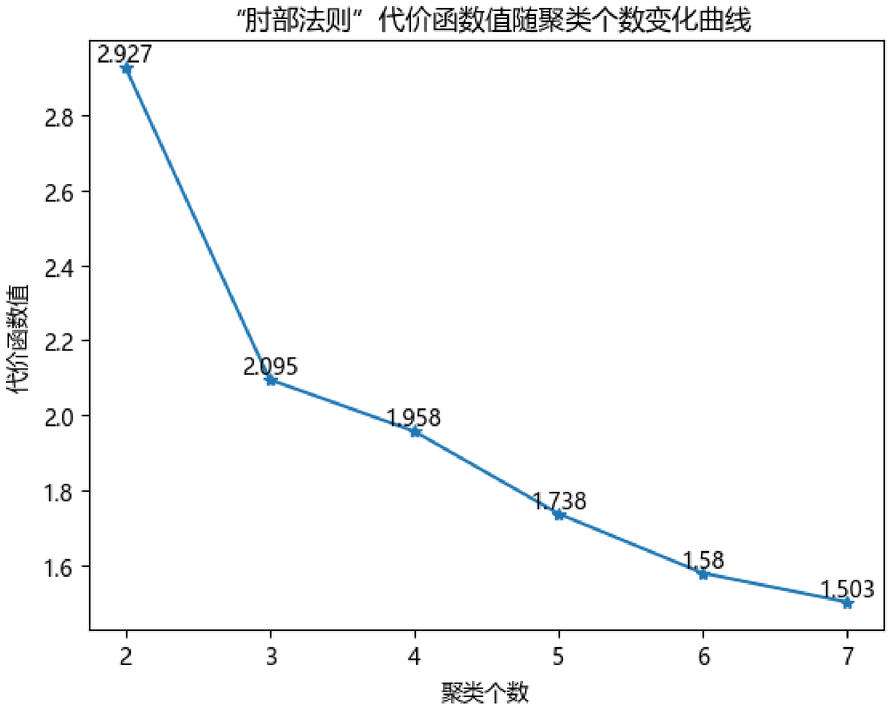
a、将其归入与其最近的中心点所在聚类

b、重新计算该聚类的重心，并用重心替换聚类中心

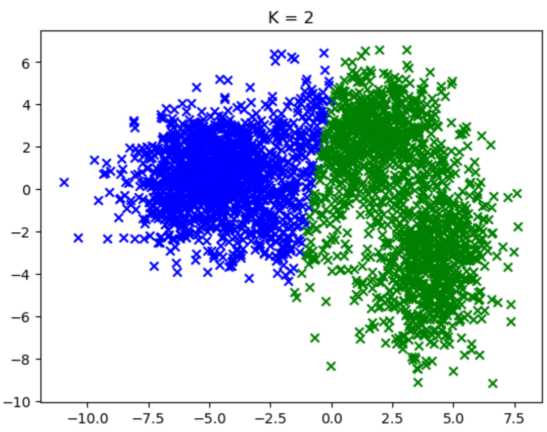
③如果新的中心和原中心足够接近则停止分类，否则继续

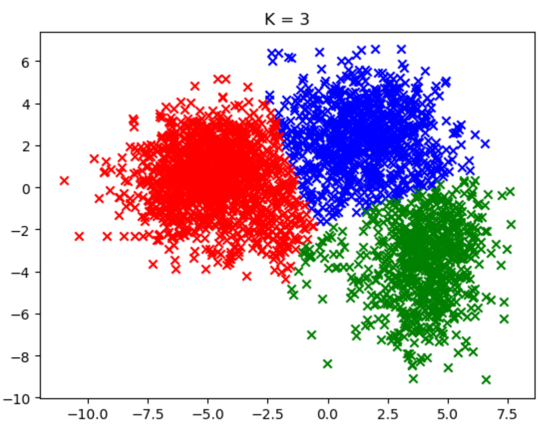
（3）最合适的聚类数目为 3，理由如下：

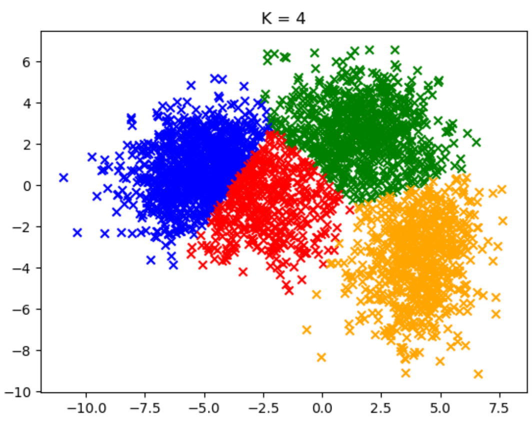
首先，根据原数据的分布，直观感受是分为 3 个聚类比较合适。由“肘部法则”（k-means 是以最小化样本与质点的平方误差作为目标函数，将每个簇的质点与簇内样本点的平方距离误差和称为畸变程度(distortions)，那么，对于一个簇，它的畸变程度越低，代表簇内成员越紧密，畸变程度越高，代表簇内结构越松散。 畸变程度会随着类别的增加而降低，但对于有一定区分度的数据，在达到某个临界点时畸变程度会得到极大改善，之后缓慢下降，这个临界点就可以考虑为聚类性能较好的点），定义代价函数为每个数据点与其对应中心的距离误差的平均值，得到如下曲线：

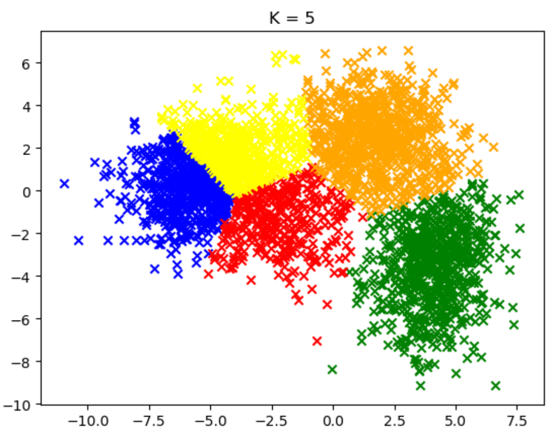


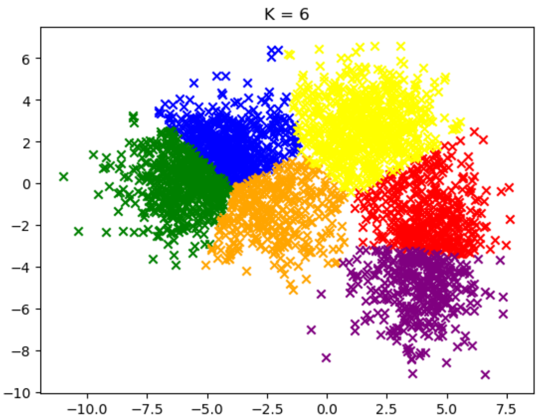
可见聚类数目为 3 时下降最明显，下面依次给出聚类数目取 2 到 7 时的效果：

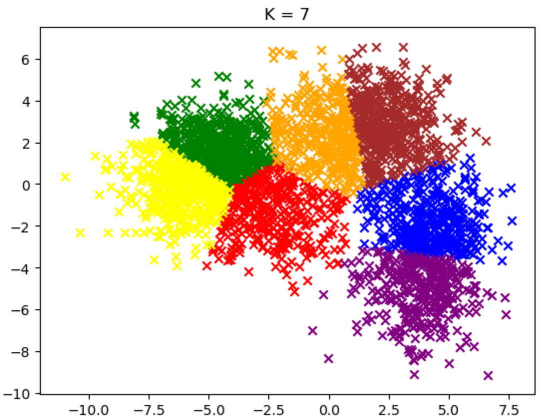








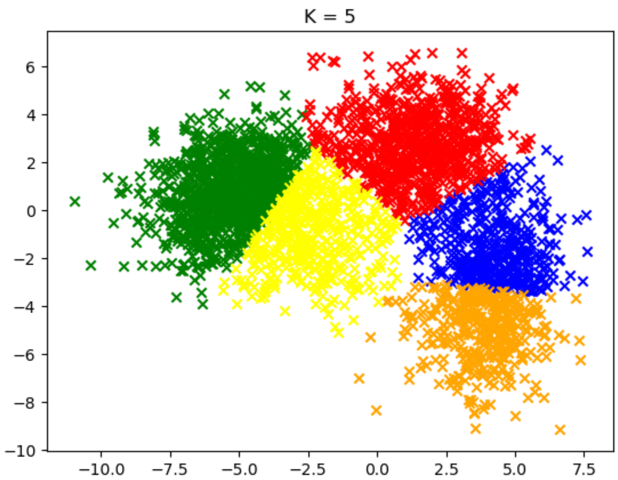




（4）以 K=5 的情况为例进行两次初始中心点不同的实验：

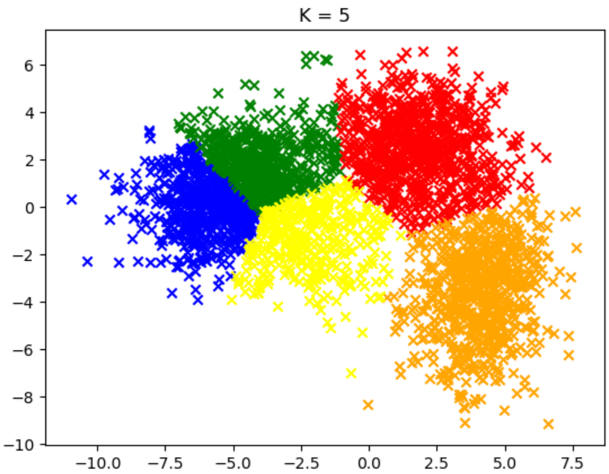
①第一次（上为初始中心点坐标，下为聚类结果）：

```
[ 5.86305787 -3.26640263]
[-3.27118813  0.22912374]
[3.2117312  1.64980798]
[ 3.65471573 -6.08698501]
[ 3.90863106 -1.17779108]
```



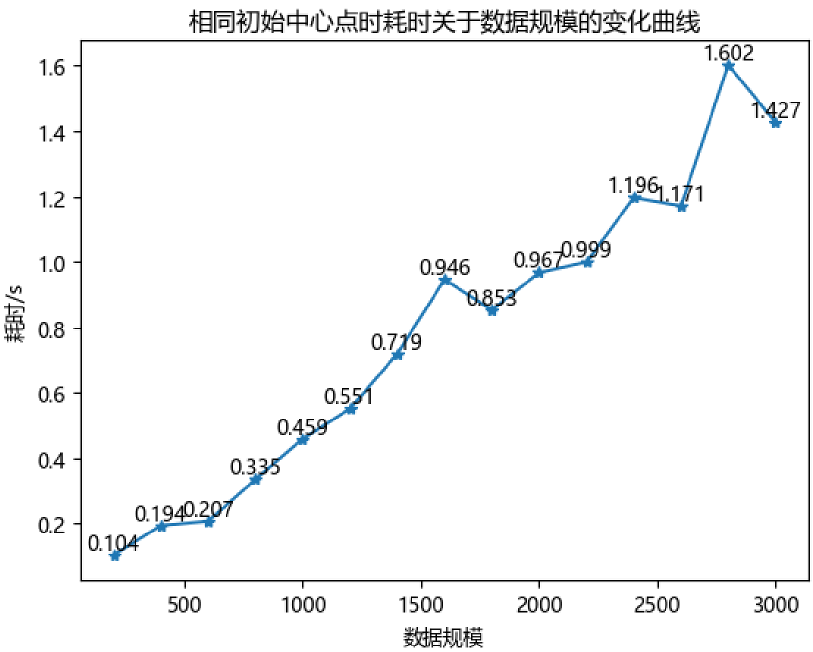
②第二次：

```
[-5.32688817 -1.14237449]
[-0.67419198  1.9358089 ]
[0.58820949 3.33623982]
[ 0.99392616 -4.94771025]
[-0.76664465 -1.65265436]
```



可以看到初始中心点的选择对最终的聚类效果有明显的影响。k-means 方法只能保证收敛到局部最优解而不能保证收敛到全局最优解，目标函数也不能确定是否是全局上的凸函数，因此不同的初始点可能会导致函数收敛到不同的局部最优解，对应地也会造成代价函数值和迭代次数的不同。

（5）为了避免随机选取初始中心点造成的耗时差异，将初始中心点确定为 data.mat 的前 K 个点(K 为聚类数目)，在 K=3 的情况下得到如下图所示的结果：



可以看到大致有耗时随数据规模增大而变大的规律，并且曲线表现出了线性特征。