《系统工程导论》k-means聚类分析

题目

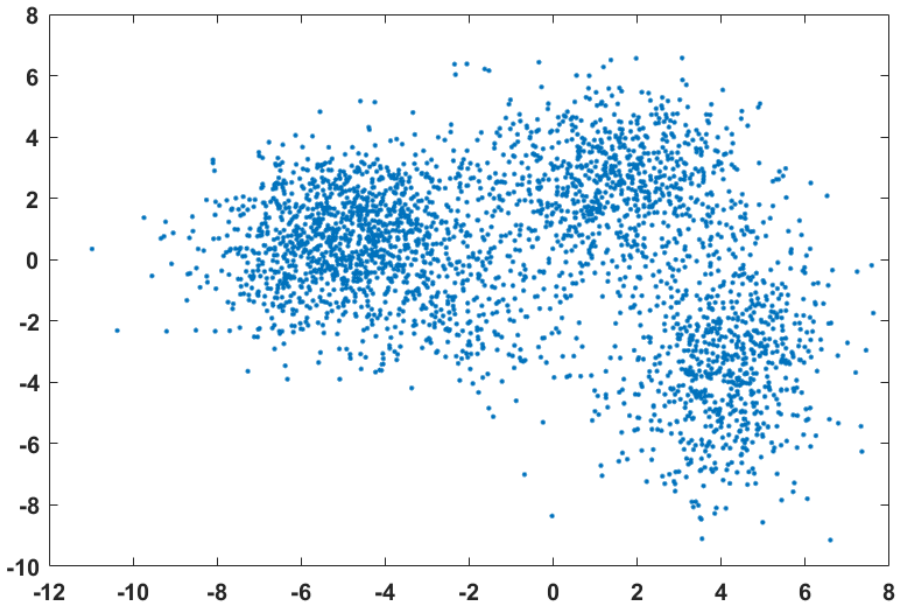
给定样本集合$\Omega = \{x_1, x_2 ... x_n\}$，其中每个样本都是$d$维的向量，k-means聚类的目标是将集合中的样本划分为k个类别，使得下述目标函数最小化：

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

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) 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？

(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

(6)请提交报告和代码文件，并且在报告中对上述问题展开分析。如果有问题，请及时联系助教