K-Means 聚类方法

一、算法应用

K-means 聚类算法也称 K 均值聚类算法，是集简单和经典于一身的基于距离的聚类算法。它采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为类簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。 K-means 聚类算法可以视作为一种无监督机器学习算法，是做聚类应用必入门算法之一。

二、算法核心思想

K-means 聚类算法是一种迭代求解的聚类分析算法，其步骤是：1）随机选取 K 个对象作为初始的聚类中心；2）然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心，聚类中心以及分配给它 们的对象就代表一个聚类；3）每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算；4）这个过程将 不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或 最小数目）聚类中心再发生变化，误差平方和局部最小。

三、详细算法流程（以2维点聚类为例）

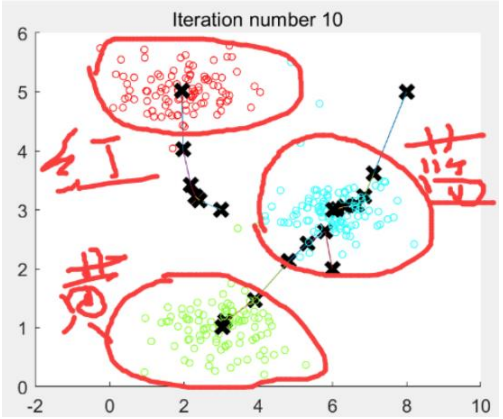
- 可参考视频或百度K-Means算法： https://www.bilibili.com/video/av63310788/

- 对象数据：

X 是300x2 的数据（即 300 个 2 维数据（可视为300个平面上的点），每一行代表一个数据（平面点），即 xi∈R1x2）

- 算法目的：

将上述300个2维数据（平面点）按空间距离聚集为K类（K为自己设置），即实现类似下图的分类结果



三、详细算法流程（算法细节）

主要完成A和B问题
代码编写！

Step1	先假设初始质心$u=\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}=\begin{pmatrix}3&3\\6&2\\8&5\end{pmatrix}\in\mathbb{R}^{3\times2}$ (假设 K=3 类)
Step2 (A: 聚类)	300 个数据x_i分别计算与u_1、u_2、u_3的距离 for $i=1:m$ % $(m=300)$ $\mathbf{dis(j)}=\|x_i-u_j\|_2^2, j=1,2,3$ %找出dis中最小距离的系数，如dis中最小距离的系数为 2，则x_i属于第 2 类 end
Step3 (B: 更新均值)	经上述计算后，更新中心$u=\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}$ for $i=1:\mathrm{K}$ %$(\mathrm{K}=3)$ % (1)将上述所有属于第 1 类的x_i数据拿出来，求平均值得更新的 $u_1=(*,*)\in\mathbb{R}^{1\times2}$ % (2)同上述 1 更新u_2，u_3 end
Step4	重复 step2、step3（循环 Step2-3）直至要求的迭代次数；如迭代设为$iter=10$

四、课题要求与思考

请使用提供的软件包：Project1_K_Means_ToStudent

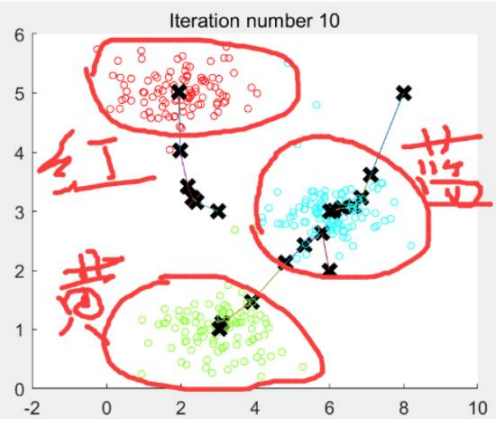
✓ 实现提供代码包里 Demo_K_Means_ToStudent.m 中的两个子函数，即

(1) idx = findClosestCentroids.m [即实现算法流程图中的(A)]

(2) centroids = computeCentroids(X, idx, K) [即实现算法流程图中的(B)]

✓ 检验代码是否正确？

若最终分类结果 idx 与文件夹中 idx_correct.mat 内的 idx_correct一致，或为右图聚类结果，则为正确。



✓ 请用： 设置断点+调试方式，将代码所有步骤过一遍！

✓ 思考

(1) 若输入的数据为高维数据，不只是本样例的2维数据，本聚类算法依然可以使用么？（不强制要求）

(2) 本样例的聚类数K是手动设置，是否可以开发一种自适应算法实现K值的自适应设置，达到更准确聚类效果？（开放问题，不强制要求）