%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

第1题

试析k均值算法能否找到最小化(9.24)的最优解

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

解：

不能

因为 最小化式9.24，找到它的最优解需考察样本集中所有可能的簇分组，这是 NP难问题，K均值算法利用贪心，迭代优化找到的只是近似解而非最优解

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

第2题 试析AGNES算法使用最小距离和最大距离的区别。

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

最大距离：可以认为是所有类别先生成一个能包围所有类内样本的最小圆，然后所有圆同时慢慢扩大相同的半径，哪个类圆能完全包围另一个类则停止，并合并这两个类。

最小距离则是扩大时遇到第一个非自己类的点就停止，并合并这两个类。

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

第3题

聚类结果中若每个簇都有一个凸包，且凸包不相交，则称为凸聚类。试析本章介绍的哪些聚类方法只能产生，哪些能产生非凸聚类。

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

在一个簇的凸包之内，是否有其他簇的样本。

K均值，LVQ，以线性边界作为最终分类解，必然是凸聚类。

高斯混合聚类其分类边界为弧线，为非凸聚类。

DBSCAN，当目标分类为多个同心弧时，其内簇会被包含，因此是非凸聚类。

AGENS，采用自底向上的方式，其衡量方式非常近似于线性的方式，应是凸聚类。

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

第4题

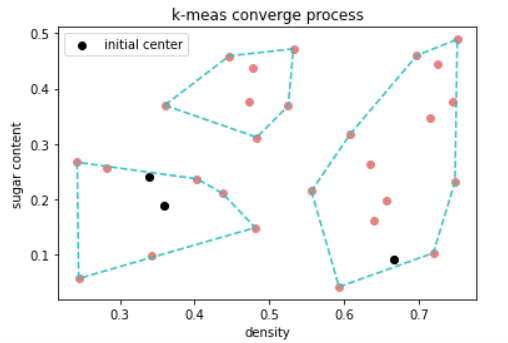
编程实现k均值算法，设置三组不同的k值，三组不同的初始中心点，在西瓜数据集4.0上进行实验，并讨论什么样的初始中心有利于取得好结果。。

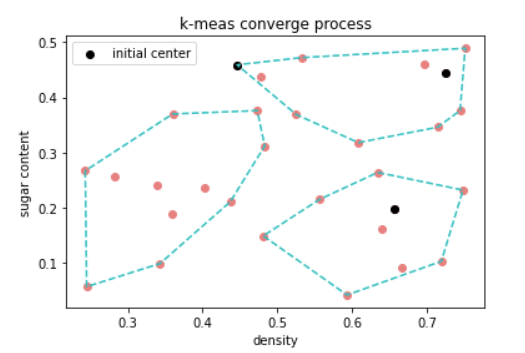
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

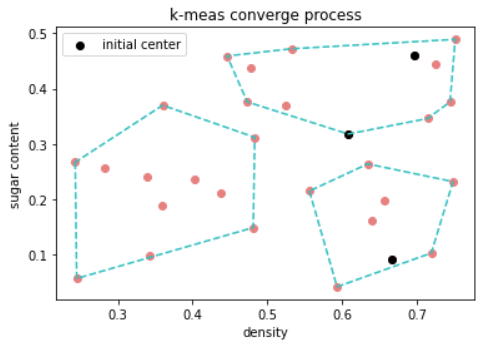
代码：



三次结果：







初始中心的选取：

各个初始点相聚的越远，结果越好。

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

第5题

用任意方法对手写数字进行聚类，数据集如下。

# -------------数据加载

digits\_train = pd.read\_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/optdigits.tra',

header=None)

digits\_test = pd.read\_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/optdigits.tes',

header=None)。

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

