信息基础二 作业

徐瑞 201800121045

问题：

有类二维空间点，A类和B类。

A 类点以(0, 0)为中心、(1, 0; 0, 1)为协方差矩阵的二维高斯分布;

B 类点以(1, 2) 为中心、(1, 0; 0, 2)为协方差矩阵的二维高斯分布; 随机生成300个A类点， 200个B类点， 并用K-Means和EM Algorithm进行聚类。

解答：

（1）K-Means聚类

K-Means聚类聚类算法是一种无监督学习方法大概步骤分为以下几个：

1.选取k个点，作为聚类中心；

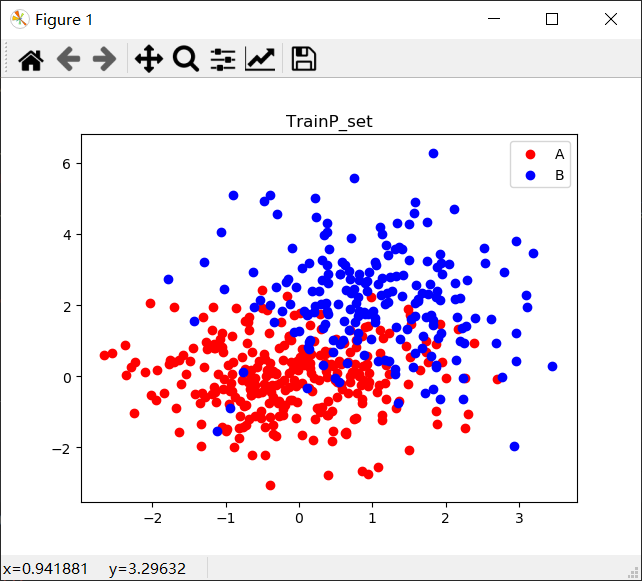
2.计算每个点分别到k个聚类中心的距离，然后把该点分到距离最小近的k那一类，这样就分成了k个簇;

3.重新计算每个簇的均值作为新的质心k；

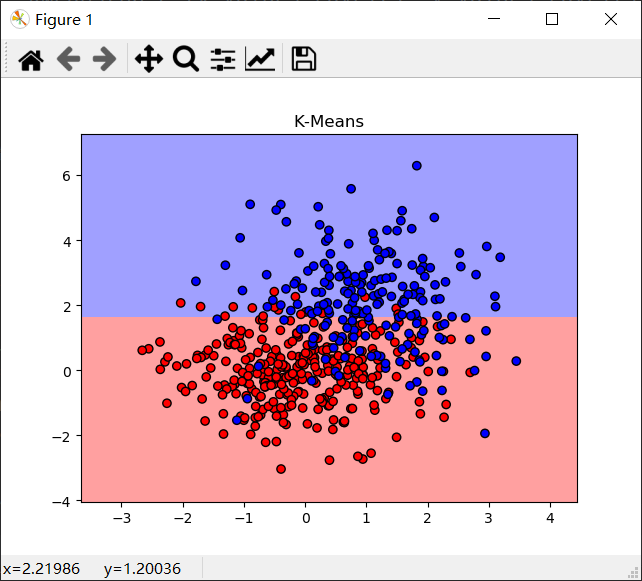
4.重复进行步骤2和3，直到质心的位置不再发生变化。

我最初选取质心的时候使用随机选取的方式，然后我又采用了最远距离法，即我先随机选取一个点作为第一个k，然后我再通过计算和比较其他所有的点到该点的距离，最后选取距离最远的那个点作为第二个k。下面是初始数据集和两种初始化方式的结果展示：

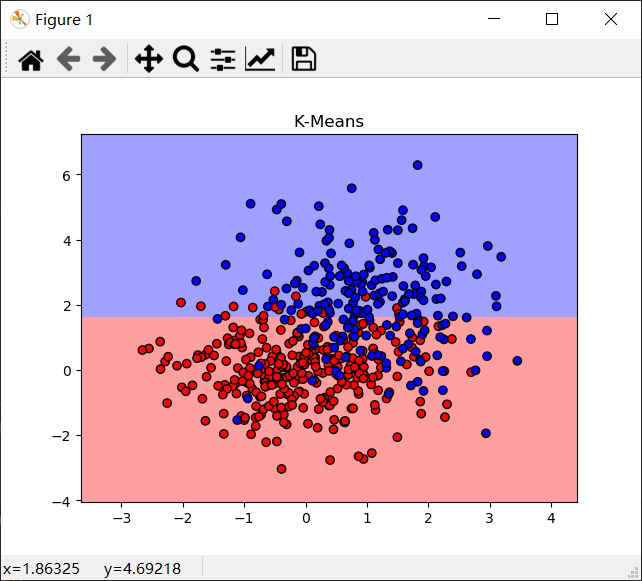
初始数据集：



随机选取k的分类决策面：



最远距离法选取k的分类决策面：



（2）EM Algorithm

SKlearn库中的GaussianMixture类是EM算法在混合高斯分布的实现，其详细参数如下：

GaussianMixture(n\_components=1, covariance\_type=’full’, tol=0.001, reg\_covar=1e-06, max\_iter=100, n\_init=1, init\_params=’kmeans’, weights\_init=None, means\_init=None,   
precisions\_init=None, random\_state=None, warm\_start=False, verbose=0, verbose\_interval=10)   
参数：   
1. n\_components:混合高斯模型个数，默认为1   
2. covariance\_type:协方差类型，包括{‘full’,‘tied’, ‘diag’, ‘spherical’}四种，分别对应完全协方差矩阵（元素都不为零），相同的完全协方差矩阵（HMM会用到），对角协方差矩阵（非对角为零，对角不为零），球面协方差矩阵（非对角为零，对角完全相同，球面特性），默认‘full’ 完全协方差矩阵   
3. tol：EM迭代停止阈值，默认为1e-3.   
4. reg\_covar:协方差对角非负正则化，保证协方差矩阵均为正，默认为0   
5. max\_iter:最大迭代次数，默认100   
6. n\_init:初始化次数，用于产生最佳初始参数，默认为1   
7. init\_params: {‘kmeans’, ‘random’}, defaults to ‘kmeans’.初始化参数实现方式，默认用kmeans实现，也可以选择随机产生   
8. weights\_init:各组成模型的先验权重，可以自己设，默认按照7产生   
9. means\_init:初始化均值，同8   
10. precisions\_init:初始化精确度（模型个数，特征个数），默认按照7实现   
11. random\_state :随机数发生器   
12. warm\_start :若为True，则fit（）调用会以上一次fit（）的结果作为初始化参数，适合相同问题多次fit的情况，能加速收敛，默认为False。   
13. verbose :使能迭代信息显示，默认为0，可以为1或者大于1（显示的信息不同）   
14. verbose\_interval :与13挂钩，若使能迭代信息显示，设置多少次迭代后显示信息，默认10次。

根据训练集，可以确定n\_components=2，混合高斯模型个数为2，然后我再设定随机数发生器为12，random\_state=12，其他参数默认就可以，训练集分布和（1）K-Means中相同，训练模型后进行预测，分类决策面如下：

