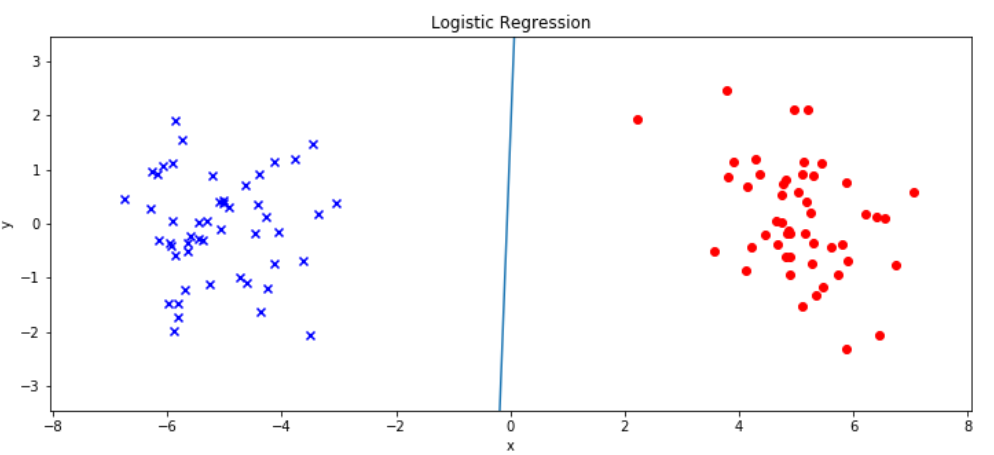
Lecture4编程作业

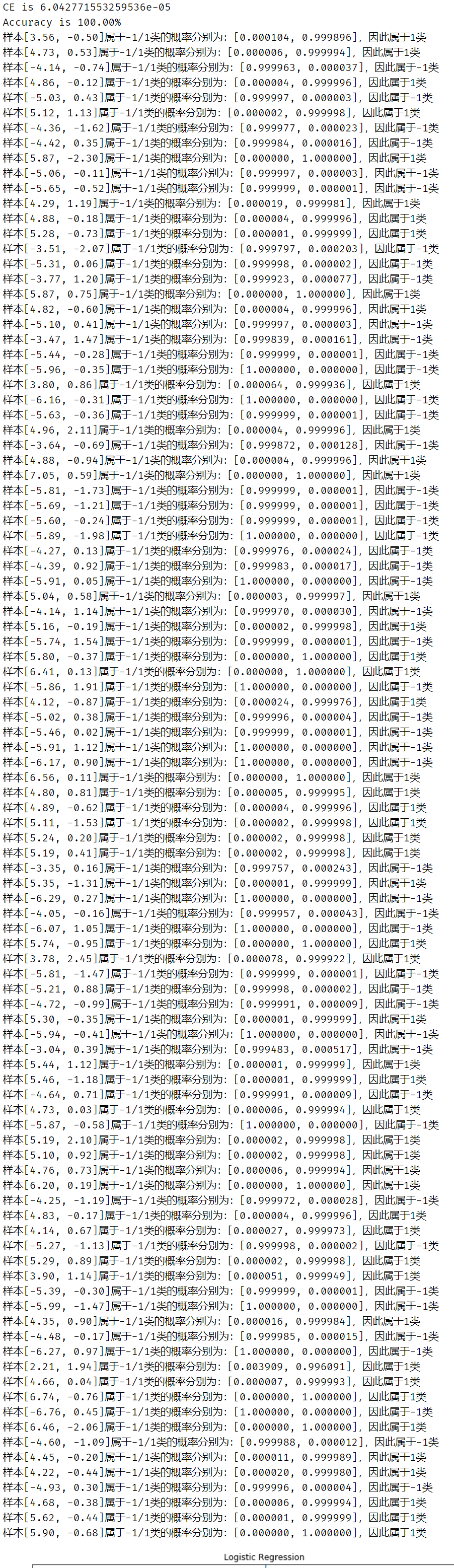
李星毅 U201712072 自实1701

1，编程实现Logistic regression算法。

（a）产生两个都具有200个二维向量的数据集和（注意：在生成数据集之前最好使用命令randn(‘seed’,0)初始化高斯随机生成器为0（或任意给定数值），这对结果的可重复性很重要）。向量的前半部分来自均值向量的正态分布，并且协方差矩阵。向量的后半部分来自均值向量的正态分布，并且协方差矩阵。其中是一个2\*2的单位矩阵。

（b）在上述数据集上和分别属于+1类和-1类，请在上述数据集的两类中各随机抽取150个样本作为训练集，运用Logistic regression算法得到的分类面，然后对余下的各50个样本进行分类，画出测试样本及其分类面，统计错误率，给出每个样本属于该类别的概率值。

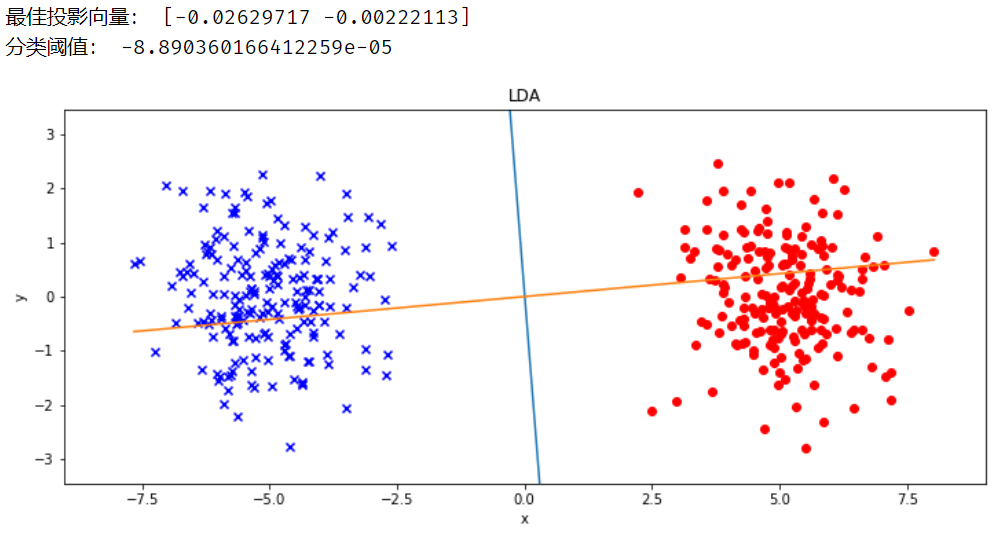




2，编程实现Fisher线性判别算法。

（a）产生两个都具有200个二维向量的数据集和（注意：在生成数据集之前最好使用命令randn(‘seed’,0)初始化高斯随机生成器为0（或任意给定数值），这对结果的可重复性很重要）。向量的前半部分来自均值向量的正态分布，并且协方差矩阵。向量的后半部分来自均值向量的正态分布，并且协方差矩阵。其中是一个2\*2的单位矩阵。

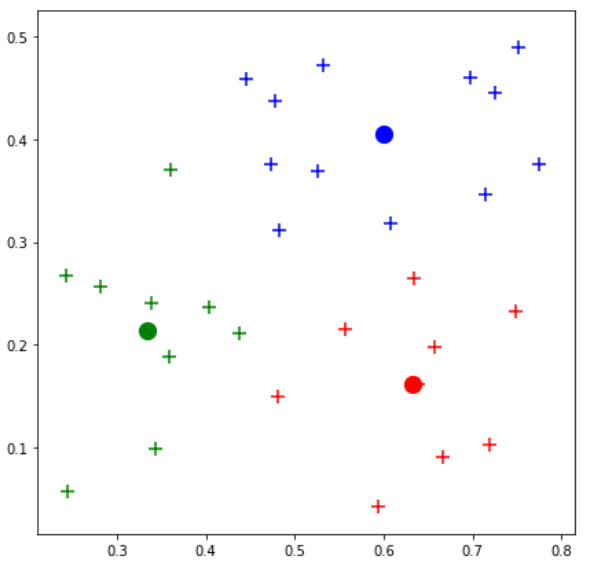
（b）在上述数据集上和分别属于+1类和-1类，请运用Fisher线性判别算法，进行分类，画出最佳投影向量，并给出分类阈值。

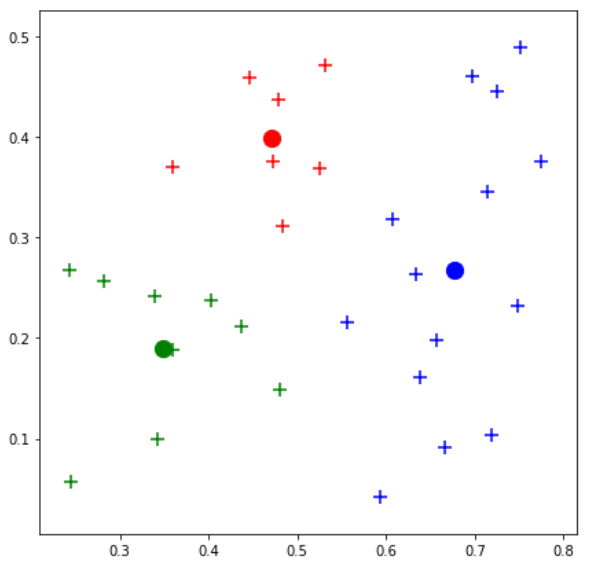


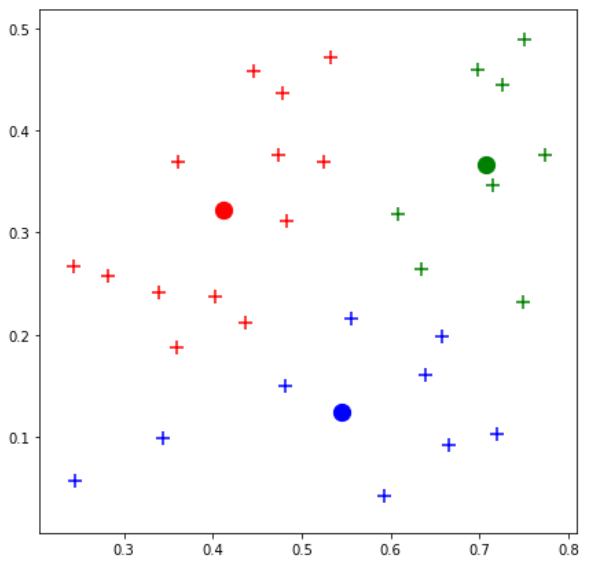
3，编程实现k均值算法，实验数据集为：(0.697 0.460)、(0.774 0.376)、( 0.634 0.264)、(0.608 0.318)、(0.556 0.215)、(0.403 0.237)、(0.481 0.149)、(0.437 0.211)、(0.666 0.091)、(0.243 0.267)、(0.245 0.057)、(0.343 0.099)、(0.639 0.161)、(0.657 0.198)、(0.360 0.370)、(0.593 0.042)、(0.719 0.103)、(0.359 0.188)、(0.339 0.241)、(0.282 0.257)、(0.748 0.232)、(0.714 0.346)、(0.483 0.312)、(0.478 0.437)、(0.525 0.369)、(0.751 0.489)、(0.532 0.472)、(0.473 0.376)、(0.725 0.445)、(0.446 0.459)。

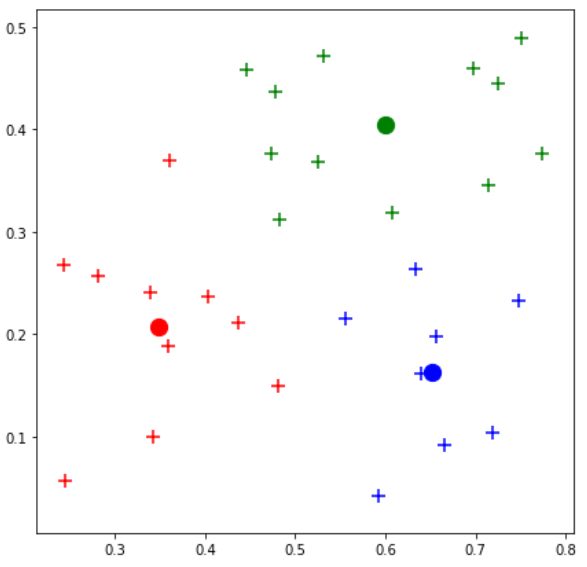
（a），k=3时聚类结果；（b），k=5时聚类结果。请图示聚类结果，用“+”代表样本，不同颜色代表不同类别，用“。”代表不同类别的类心。

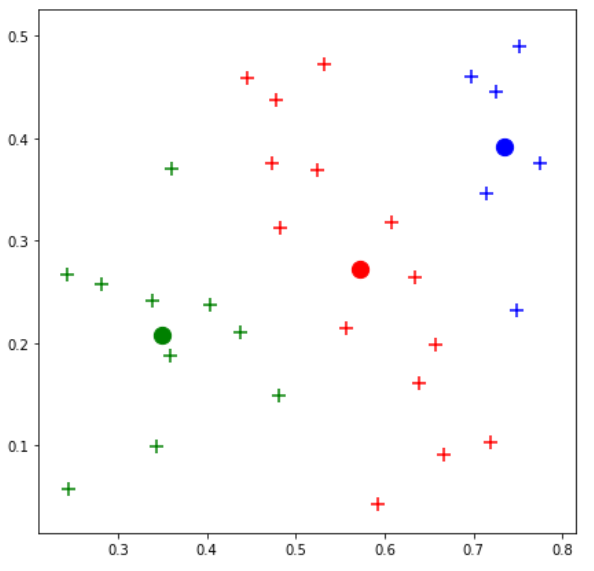
（a）k=3时



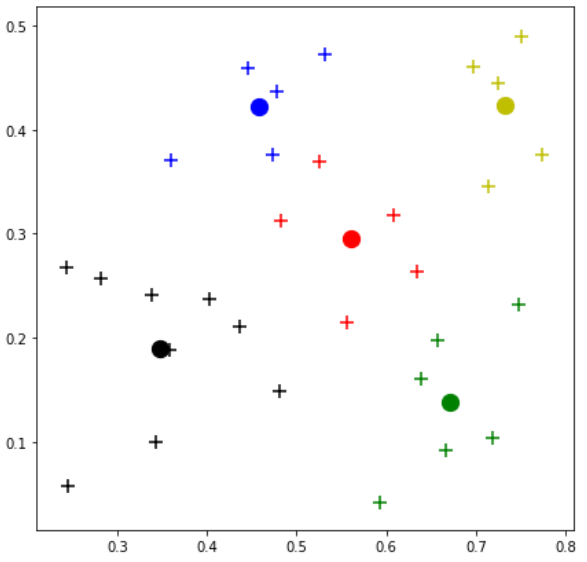


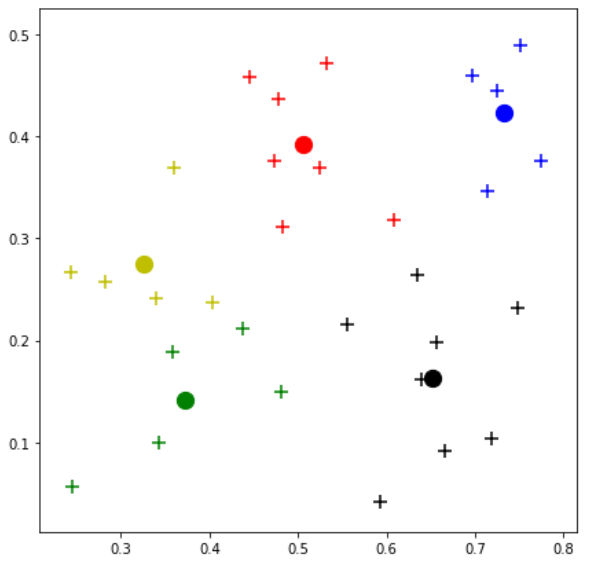


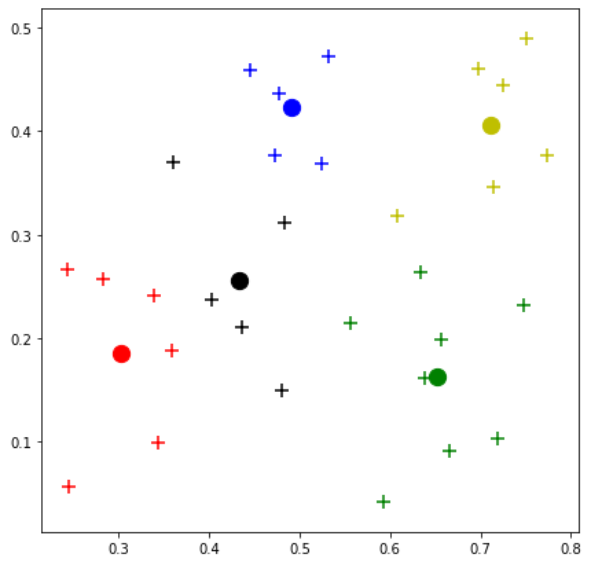


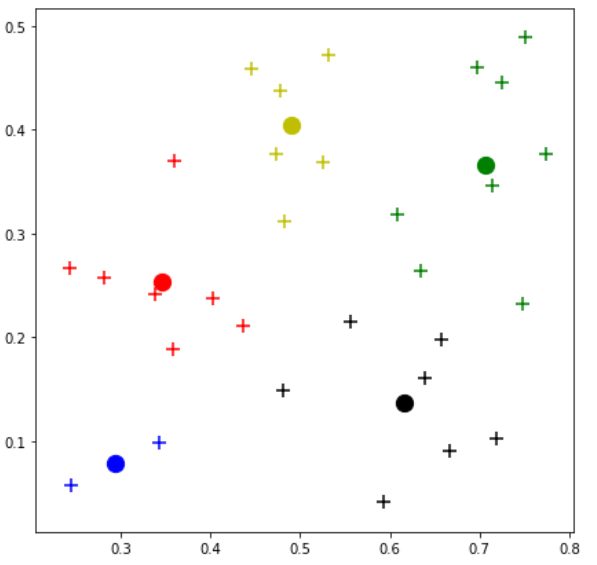


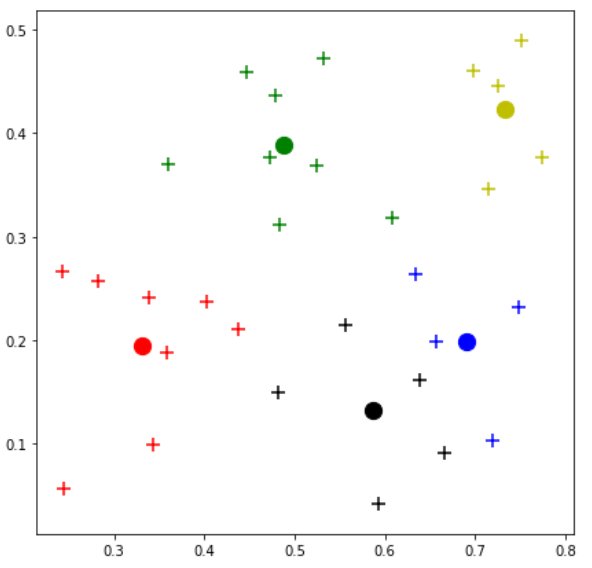
（b）k=5时











实验结果分析：

1. k均值算法有主要的两个缺陷：一个是k值需要预先给定，属于预先知识，很多情况下k值的估计是非常困难的；第二个是k均值算法对初始选取的聚类中心点是敏感的，不同的随机种子点得到的聚类结果完全不同。
2. 随机的选择初始质心可能只能得到局部的最优解，而无法得到全局的最优解。
3. 一种稍好一点的改进思路是：假设使用欧几里得距离作为度量，可以考虑使用误差平方和（Sum of the Squared Error, SSE）作为聚类的目标函数，多次运行k均值产生的不同的簇集，选择SSE最小的那个作为聚类的结果。该策略比较简单，但是效果可能也不是很好，这取决于数据集合寻找的簇的个数。