第6次作业

简答题

1. 请简述adaboost算法的设计思想和主要计算步骤。

   答：adaboost是一种集成学习方法，它使用若干个弱学习器的线性组合构造一个精度更高的模型，其中弱学习器是指性能稍高于随机猜测的模型。adaboost的主要思想是逐个训练若分类器，并提高前面分类器分类错误的权重，最后将所有若分类器线性组合构成强分类器，线性组合的权重取决于弱分类器的错误率，错误率越低，权重越高。

   下面以二分类为例阐述adaboost的计算步骤。依次训练M个弱分类器将N个样本分成两类。

   1. **初始化权重**:$w_n = \frac{1}{N}, n = 1, 2, \cdots, N$。
   2. 从$m = 1, 2, \cdots, M$开始循环
      1. **训练第m个若分类器**，目标是极小化误差函数J_m,其中

      $$J_m = \sum_{n=1}^{N} w_n^{(m)} \delta(y_m(x_n) \neq t_n)$$

      上式中的$w_n^{(m)}$表示第m个弱分类器对第n个数据的权重，$\delta(y_m(n) \neq t_n)$在预测错误时取1，否则取0。因此J_m表示第m个弱分类器对n个数据错误率的加权平均。

      2. **更新权重**。按照下面的计算公式更新权重。易知$\epsilon_m \in [0, 1]$，而a_m是ϵ_m的单调递减函数。我们要求弱分类器的错误率比随机分类好，即$\epsilon_m < 0.5$，则$a_m > 0$。当第m个弱分类器将第n个样本正确时，$\exp\left(a_m \delta(y_m(n) \neq t_n)\right) = 1$，因此下一个分类器对此样本的权重不变；反之当错误分类时，$\exp\left(a_m \delta(y_m(n) \neq t_n)\right) > 1$，即错分样本的权重变大。

      $$\epsilon_m = \frac{J_m}{\sum_{n=1}^{m} w_n^{(m)}}$$
      $$a_m = \log\left(\frac{1-\epsilon}{\epsilon}\right)$$
      $$w_n^{(m+1)} = w_n^{(m)} \exp\left(a_m \delta(y_m(n) \neq t_n)\right)$$

      3. 将M个弱分类器线性组合，得到强分类器，计算公式如下

      $$Y_M(x) = sign\left(\sum_{m=1}^{M} a_m y_m(x)\right)$$

      权重a_m是错误率的单调递减函数，即错误率越高时，弱分类器的权重越小。

2. 请从混合高斯密度函数估计的角度，简述K-means聚类算法的原理（请主要用文字描述，条理清晰）；请给出K-Means聚类算法的主要步骤；请说明哪些因素会影响K-Means算法的聚类性能。

   答：高斯混合模型（GMM）实质多个单高斯模型的线性组合，理论上高斯混合模型可以拟合出任何类型的分布。GMM常用于聚类，如果要GMM的分布中随机选取一个点，可以分成两步：首先随机从选择一个单高斯模型，每个单高斯模型被选中的概率是π_k。选好单高斯模型之后，再考虑从这个模型中选择一个点。将GMM用于聚类时，本质是根据已有数据推断出GMM的概率分布。我们先假定GMM由K个单高斯模型组成，因此我们需要推断K个成分各自的均值向量和协方差矩阵以及他们的权重π_k。而K-Means是GMM的特殊情况，当GMM中每个成分的协方差矩阵退化成对角阵且对角线上的元素很小的时候，样本之间的马氏距离退化成欧氏

距离，此时我们只需要估计π_k和各个成分的均值向量。此时软指派退化成硬指派。

K-means聚类算法的主要步骤如下

1. 确定超参数k，并随机初始化k个聚类中心。
2. 遍历每个样本，计算该样本和k个聚类中心的距离，选择距离最小的一个作为自己的类别。
3. 根据第2步更新的标签，重新计算聚类中心。计算聚类中心的方法是求本类所有样本点的均值。
4. 循环上述2-3步，直到收敛或者达到最大迭代次数。

影响K-Means算法聚类性能的主要因素有

1. 初始聚类中心的选取。聚类中心选取不当可能会导致算法收敛到局部最优，因此要多次随机初始化聚类中心，从中选择性能不错的结果。
2. 聚类数k的选择。k越大，误差越小，但是很可能不符合数据的分布。可以采用“肘部法则”选择k值。

3. 请简述谱聚类算法的原理，给出一种谱聚类算法（经典算法、Shi算法和Ng算法之一）的计算步骤，请指出哪些因素会影响聚类的性能。

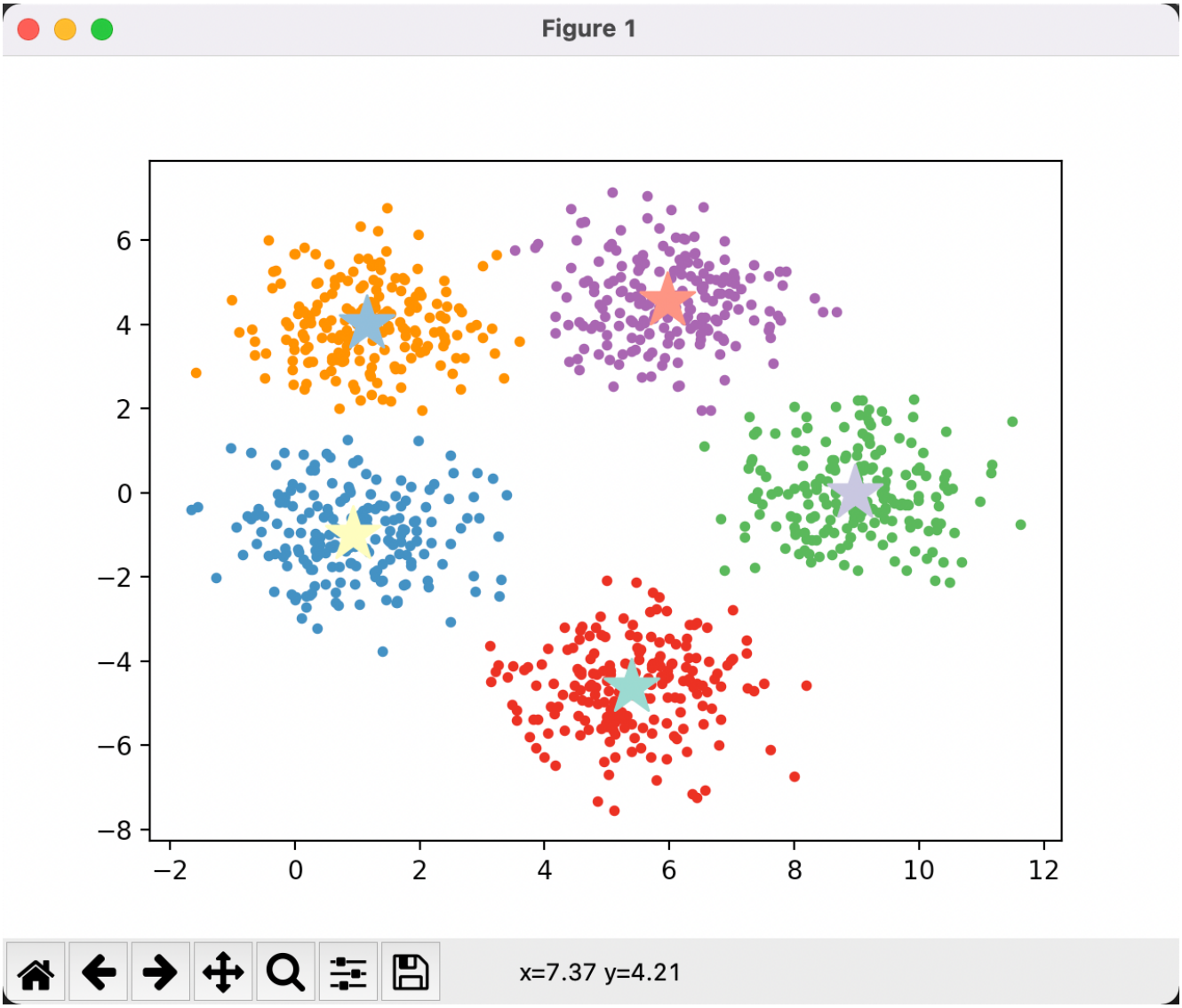
   答：谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。经典的聚类算法步骤如下：

   1. 根据输入数据构造相似矩阵。根据相似矩阵构造邻接矩阵W和度矩阵D。
   2. 计算拉普拉斯矩阵L。
   3. 构建标准化的拉普拉斯矩阵$D^{-1/2}LD^{-1/2}$。
   4. 计算$D^{-1/2}LD^{-1/2}$的最小的k_1个特征值及其对应的特征向量f。
   5. 标准化特征向量f并组成$n \times k_1$维的特征矩阵F。
   6. 对F中的每一行作为一个k_1维样本，使用K-means等方法聚类。

计算编程题

1. 编写一个程序，实现经典的K-均值聚类算法。
2. 令聚类个数等于5，采用不同初始值，报告聚类精度，以及最后获得的聚类中心，并计算所获得的聚类中心与对应的真是分布的均值之间的误差。

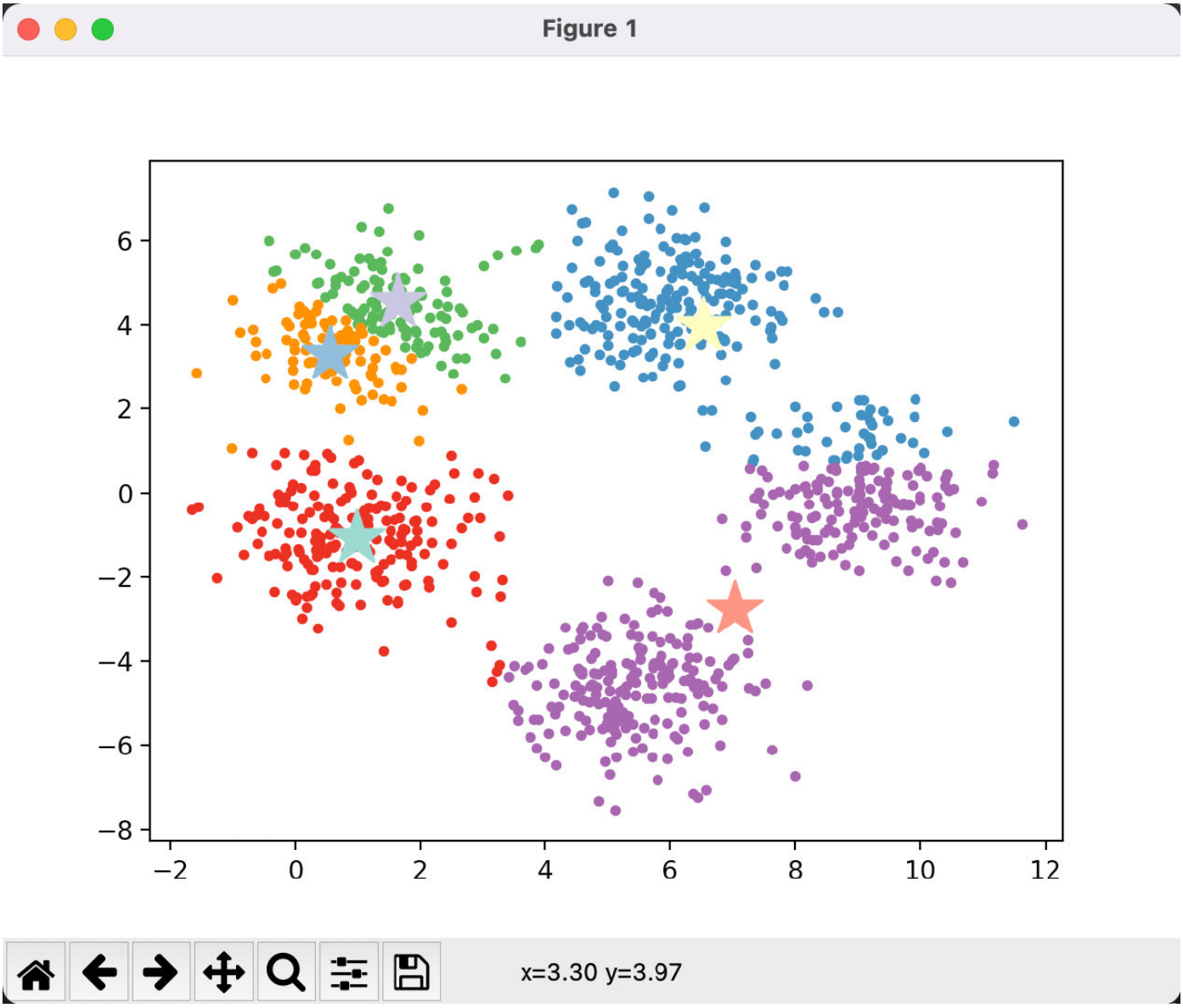
   答：某次实验的结果如下



准确率以及预测聚类中心和实际聚类中心的误差如下

```
聚类精度： 0.9960460460460461
预测中心与真实中心的距离： 0.49840606869908505
```

随机初始化聚类中心可能导致Kmeans陷入局部最优，实验结果如下：



top/Learning_materials/模式识别/作业/HW6/co
聚类精度： 0.8796576576576577
预测中心与真实中心的距离： 9.448498525486