Homework of Pattern classification IV

*Name: Xue Yuan — Student number: 202228015926034

***Abstract*—This document is about the first homework for Pattern classification by LaTeX.**

I. Short Answers and Descriptions

Q1: 请描述使用高斯混合模型进行数据聚类的过程。

A: 采用gauss混合模型进行聚类,其主要包括三个过程:

1) 样本的生成:通过类先验概率密度$p(\omega_j)$随机选择一个样本类别, 然后通过类条件概率密度函数$p(x|\omega,\theta_j)$随机生成相应的样本。其中,高斯混合模型GMM(Gaussian Mixture Model)的概率函数为:

$$P(y|\theta)=\sum_{k=1}^{k}\alpha_k\varphi(y|\theta_k) \tag{1}$$

$$\sum_{k=1}^{k}\alpha_k=1;\alpha_k\geq 0 \tag{2}$$

$$\varphi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \tag{3}$$

2) 我们认为:样本来自c个不同类别,c的值是已知的。每一类的先验概率$p(\omega_j)$也是已知的, 类条件概率密度函数$p(x|\omega,\theta_j)$形式上是已知的,存在c个参数向量$\theta_j, j=1,2,\ldots,c$。
3) 根据极大似然估计,求解类条件概率密度函数中的参数θ_j,再根据最大后验概率进行聚类划分:

$$C(\mathbf{x})=\arg\max_j p\left(\omega_j\mid\mathbf{x},\boldsymbol{\theta}\right)$$

后验概率:

$$p\left(\omega_j\mid\mathbf{x},\boldsymbol{\theta}\right)=\frac{p\left(\mathbf{x}\mid\omega_j,\boldsymbol{\theta}\right)P\left(\omega_j\right)}{\sum\limits_{i=1}^{c}p\left(\mathbf{x}\mid\omega_i,\boldsymbol{\theta}\right)P\left(\omega_i\right)}$$

Q2:对于:$x_1=(4,5)^T, x_2=(1,4)^T, x_3=(0,1)^T, x_4=(5,0)^T, x_5=(4,1)^T, x_6=(0,6)^T$现有以下三种聚类划分。

（1）$\{x_1,x_2,x_6\}$, $\{x_3,x_4,x_5\}$
（2）$\{x_1,x_4,x_5\}$, $\{x_2,x_3,x_6\}$
（3）$\{x_1,x_2,x_3,x_6\}$, $\{x_4,x_5\}$

Fig. 1: Three types of division

假定我们聚类的准则是最小平方和误差，请判断上述三个划分中哪个更好?

A:由最小平方和误差定义,可以计算各个划分相应的误差大小为$tr\left[S_w\right]$,其中,S_w为:

$$\begin{aligned}\mathbf{S}_W&=\sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}\left(\mathbf{x}-\mathbf{m}_i\right)\left(\mathbf{x}-\mathbf{m}_i\right)^t\\&=\sum_{k=1}^{6}\mathbf{x}_k\mathbf{x}_k^t-\sum_{i=1}^{c}n_i\mathbf{m}_i\mathbf{m}_i^t\end{aligned}$$

容易计算:

$$\sum_{k=1}^{6}\mathbf{x}_k\mathbf{x}_k^t=\begin{pmatrix}58&28\\28&94\end{pmatrix}$$

计算三种聚类划分

$$\mathbf{m}_i=\frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}$$

1) 第一种划分的聚类中心为:

$$m_1=\frac{1}{3}\left(\begin{pmatrix}4\\5\end{pmatrix}+\begin{pmatrix}1\\4\end{pmatrix}+\begin{pmatrix}0\\6\end{pmatrix}\right)=\begin{pmatrix}5/3\\5\end{pmatrix}$$

$$m_2=\frac{1}{3}\left(\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}5\\0\end{pmatrix}+\begin{pmatrix}4\\1\end{pmatrix}\right)=\begin{pmatrix}3\\2/3\end{pmatrix}$$

可以计算:

$$\sum_{i=1}^{c}n_i\mathbf{m}_i\mathbf{m}_i^t=3\left(m_1m_1^T+m_2m_2^T\right)=\begin{pmatrix}106/3&31\\31&229/3\end{pmatrix}$$

$$S_{w1}=\begin{pmatrix}58&28\\28&94\end{pmatrix}-\begin{pmatrix}106/3&31\\31&229/3\end{pmatrix}=\begin{pmatrix}68/3&-3\\-3&53/3\end{pmatrix}$$

可以得到:$tr\left[S_{w1}\right]=121/3=40.33$

2) 第二种划分的聚类中心为:

$$m_1=\frac{1}{3}\left(\begin{pmatrix}4\\5\end{pmatrix}+\begin{pmatrix}5\\0\end{pmatrix}+\begin{pmatrix}4\\1\end{pmatrix}\right)=\begin{pmatrix}13/3\\2\end{pmatrix}$$

$$m_2=\frac{1}{3}\left(\begin{pmatrix}1\\4\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}0\\6\end{pmatrix}\right)=\begin{pmatrix}1/3\\11/3\end{pmatrix}$$

同理可以计算:

$$\sum_{i=1}^{c}n_i\mathbf{m}_i\mathbf{m}_i^t=\begin{pmatrix}170/3&89/3\\89/3&157/3\end{pmatrix}$$

$$S_{w2}=\begin{pmatrix}58&28\\28&94\end{pmatrix}-\begin{pmatrix}170/3&89/3\\89/3&157/3\end{pmatrix}=\begin{pmatrix}4/3&-5/3\\-5/3&125/3\end{pmatrix}$$

可以得到:$tr\left[S_{w2}\right]=43$

3) 第三种划分的聚类中心为:

$$m_1=\frac{1}{4}\left(\begin{pmatrix}4\\5\end{pmatrix}+\begin{pmatrix}1\\4\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}0\\6\end{pmatrix}\right)=\begin{pmatrix}5/4\\4\end{pmatrix}$$

$$m_2 = \frac{1}{2}\left(\begin{pmatrix}5\\0\end{pmatrix}+\begin{pmatrix}4\\1\end{pmatrix}\right)=\begin{pmatrix}9/2\\1/2\end{pmatrix}$$

同理可以计算:

$$\sum_{i=1}^{c} n_i \mathbf{m}_i \mathbf{m}_i^t = \begin{pmatrix}187/4 & 49/2\\ 49/2 & 129/2\end{pmatrix}$$

$$S_{w3} = \begin{pmatrix}58 & 28\\ 28 & 94\end{pmatrix} - \begin{pmatrix}187/4 & 49/2\\ 49/2 & 129/2\end{pmatrix} = \begin{pmatrix}45/4 & 7/2\\ 7/2 & 59/2\end{pmatrix}$$

可以得到:$tr\left[S_{w3}\right] = 163/4 = 40.75$

综上计算可以看出:

$$tr[S_{w2}] = 43 > tr[S_{w3}] = 40.75 > tr[S_{w1}] = 40.33$$

因此,根据最小平方和误差准则,可以认定:第一种分类方法最为合理。

Q3:请阐述$k-means$聚类和模糊$k-means$聚类的异同。

A1:相同之处在于,这两种方法都需要预设类别种类数,需要预设初始化各个类别的聚类中心,并不断迭代跟新聚类中心。 并且它们都是通过误差平方和准则的方法不断迭代修正。因此,这两种方法都对初始化条件敏感,且难以处理外点。

A2:不同之处在于,模糊$k-means$聚类认为样本以一定概率隶属于不同类别(称隶属度),并由此修正了误差平方和准则的表达形式如下:

$$J_{fuz} = \sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_i(x_j)\right]^b \|x_j - m_i\|^2$$

因此,每一次的迭代不仅要跟新聚类中心,还要对隶属度进行更新,最终使聚类中心和隶属度的变化量都小于可接受极限。 相应的输出量同样包括聚类中心和样本属于它们的隶属度。

A3:最优解的推导过程如下,由优化目标函数:

$$\min_{\mu,m} \sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_i\left(\mathbf{x}_j\right)\right]^b \left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2$$

$$\text{s.t.} \ \sum_{i=1}^{c}\mu_i\left(\mathbf{x}_j\right) = 1, \quad j = 1, 2, \ldots, n$$

可以引入n个拉格朗日因子:

$$J' = \sum_{i=1}^{c}\sum_{j=1}^{n}[\mu_i(x_j)]^b \left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2 + \sum_{j=1}^{n}\lambda_i\left(\sum_{i=1}^{c}[\mu_i(x_j)]^b - 1\right)$$

欲使J最小化,可以分别使之对聚类中心m_k和隶属度$\mu_i(x_j)$求偏导。 对聚类中心,我们记,$\mu_i(x_j) = \mu_{ij}$,可得:

$$\frac{\partial J'}{\partial m_k} = \sum_{i=1}^{c}\sum_{j=1}^{n}\frac{\partial \mu_{ij}^b \left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2}{\partial m_i} - \frac{\partial}{\partial m_k}\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}\mu_{ij} - 1\right)$$

注意到:

$$\sum_{i=1}^{c}\mu_i\left(\mathbf{x}_j\right) = 1, \quad j = 1, 2, \ldots, n$$

则:

$$\begin{aligned}\frac{\partial J'}{\partial m_k} &= \sum_{i=1}^{c}\sum_{j=1}^{n}\frac{\partial \mu_{ij}^b \left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2}{\partial m_k}\\ &= \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^b\frac{\partial}{\partial m_k}\left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2\\ &= \sum_{j=1}^{n}\mu_{kj}^b\frac{\partial}{\partial m_k}\left\|\mathbf{x}_j - \mathbf{m}_k\right\|^2\\ &= -2\sum_{j=1}^{n}\mu_{kj}^b\left(x_j - m_k\right) = 0\end{aligned}$$

改换下标$k = i$,化简得到聚类中心的优化结果表达式:

$$\mathbf{m}_i = \frac{\sum_{j=1}^{n}\left[\mu_i\left(\mathbf{x}_j\right)\right]^b \mathbf{x}_j}{\sum_{j=1}^{n}\left[\mu_i\left(\mathbf{x}_j\right)\right]^b}$$

对隶属度$\mu_i\left(\mathbf{x}_j\right)$(以下简写$\mu_{ij}$)求偏导,可以采用分部分求导的策略,对表达式左侧求导有:

$$\frac{\partial J'}{\partial \mu_{ij}} = \frac{\partial}{\partial \mu_{ij}}\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^b\left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2$$

不妨记:$d_{ij} = (x_j - m_i)$,有:

$$\frac{\partial J'}{\partial \mu_{ij}} = \sum_{j=1}^{n}\sum_{i=1}^{c}\frac{\partial}{\partial \mu_{ij}}\mu_{ij}^b d_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{c}b\mu_{ij}^{b-1}d_{ij}$$

对表达式右侧求导有:

$$\begin{aligned}\frac{\partial J'}{\partial \mu_{ij}} &= \frac{\partial}{\partial \mu_{ij}}\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}\mu_{ij} - 1\right)\\ &= \frac{\partial}{\partial \mu_{ij}}\sum_{j=1}^{n}\left(\sum_{i=1}^{c}\lambda_j\mu_{ij} - \lambda_j\right)\\ &= \sum_{j=1}^{n}\sum_{i=1}^{c}\frac{\partial}{\partial \mu_{ij}}\lambda_j\mu_{ij}\\ &= \sum_{j=1}^{n}c\lambda_j\end{aligned}$$

合并两侧求导结果,可得:

$$\begin{aligned}\frac{\partial J'}{\partial \mu_{ij}} &= \sum_{j=1}^{n}\sum_{i=1}^{c}b\mu_{ij}^{b-1}d_{ij} + \sum_{j=1}^{n}c\lambda_j\\ &= \sum_{j=1}^{n}\left(\sum_{i=1}^{c}b\mu_{ij}^{b-1}d_{ij} + c\lambda_j\right)\\ &= \sum_{j=1}^{n}\sum_{i=1}^{c}\left(b\mu_{ij}^{b-1}d_{ij} + \lambda_j\right) = 0\end{aligned}$$

考虑初始条件:

$$b\mu_{ij}^{b-1}d_{ij} + \lambda_j \geq 0$$
$$\therefore b\mu_{ij}^{b-1}d_{ij} + \lambda_j = 0, \forall i = 1\ldots c$$

可以化简:

$$\mu_{ij} = \left(\frac{-\lambda_j}{bd_{ij}}\right)^{\frac{1}{b-1}}$$

考虑到:

$$\sum_{k=1}^{c} \mu_{kj} = 1 \quad \forall j = 1, \ldots, n$$

可以计算:

$$\sum_{k=1}^{c} \left(\frac{-\lambda_j}{bd_{kj}}\right)^{\frac{1}{b-1}} = \left(\frac{-\lambda_j}{b}\right)^{\frac{1}{b-1}} \sum_{k=1}^{c} \left(\frac{1}{d_{kj}}\right)^{\frac{1}{b-1}} = 1$$

$$\Rightarrow \left(\frac{-\lambda_j}{b}\right)^{\frac{1}{b-1}} = \frac{1}{\sum_{k=1}^{c} \left(\frac{1}{d_{kj}}\right)^{\frac{1}{b-1}}}$$

带回到μ_{ij}表达式中,化简可得:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{1}{d_{kj}}\right)^{\frac{1}{m-1}}} \cdot \left(\frac{1}{d_{ij}}\right)^{\frac{1}{m-1}}$$

带入$d_{ij} = (x_j - m_i)$,可以得到最终的隶属度优化结果表达式:

$$\mu_i(\mathbf{x}_j) = \frac{\left(1/\|\mathbf{x}_j - \mathbf{m}_i\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{c} \left(1/\|\mathbf{x}_j - \mathbf{m}_k\|^2\right)^{1/(b-1)}}$$

Q4: 证明:在 K 均值聚类中,在某次迭代的时候,将属于第 i 类的样本点移到第 j 类之后, 属于第 i 类的样本点对应的误差平方和将变为:

$$J_i^* = J_i - \frac{n_i\|\hat{x} - m_i\|^2}{n_i - 1}$$

A: 由类中心变化:

$$m_i^* = m_i - \frac{\hat{x} - m_j}{n_i - 1}$$

则此时的误差平方和可以表示为:

$$\begin{aligned}
J_i^* &= \sum_{x\in D_i} \|x - m_i^*\|^2 - \|\hat{x} - m_i^*\|^2 \\
&= \sum_{x\in D_j} \left(\|x - m_i\|^2 + \frac{2}{n_i - 1}(\hat{x} - m_i)^T(x - m_i)\right) \\
&+ \sum_{x\in D_j} \frac{\|\hat{x} - m_i\|^2}{(n_i - 1)^2} - \|\frac{n_i}{n_i - 1}(\hat{x} - m_i)\|^2 \\
&= J_i + \frac{2}{n_i - 1}(\hat{x} - m_i)^T \left(\sum_{x\in D_i} x - \sum_{x\in D_i} m_i\right) \\
&+ \left[\frac{n_i\|\hat{x} - m_i\|^2}{(n_i - 1)^2} - \frac{n_i^2\|\hat{x} - m_i\|^2}{(n_i - 1)^2}\right] \\
&= J_i + \frac{2}{n_i - 1}(\hat{x} - m_i)^T (n_i m_i - n_i m_i) - \frac{n_i\|\hat{x} - m_i\|^2}{(n_i - 1)} \\
&= J_i - \frac{n_i\|\hat{x} - m_i\|^2}{(n_i - 1)}
\end{aligned}$$

综上,可以得到属于第i类的样本点引起的误差平方和的减小n量为:

$$J_i^* = \sum_{x\in D_i} \|x - m_i^*\|^2 - \|\hat{x} - m_i^*\|^2 \quad = J_i - \frac{n_i\|\hat{x} - m_i\|^2}{n_i - 1}$$

Q5: **A1**:原问题为:

$$\begin{gathered}
\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \\
s.t. \quad y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, i = 1, 2; N \\
\xi_i \geq 0, i = 1, 2, ; N
\end{gathered}$$

引入拉格朗日乘子,得到拉格朗日函数:

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i +$$

$$\sum_{i=1}^{N} \alpha_i \left(1 - \xi_i - y_i\left(w^T x_i + b\right)\right) - \sum_{i=1}^{N} \mu_i \xi_i$$

可以得到原问题的对偶问题:

$$\begin{gathered}
\max_{\alpha} \min_{w,b} L(w, b, \xi, \alpha, \mu) \\
s.t. \quad \alpha_i \geq 0, \quad \mu_i \geq 0
\end{gathered}$$

A2:如图$Fig2$所示,当C取值很大时($C \rightarrow +\infty$时,线性支持向量机退化为线性可分的支持向量机),SVM更加关注分类准确度,因此其分界面间隔很小,基本没有错分现象出现; 同样,当C取值很小甚至接近于0时,SVM更加关注分类间隔问题,倾向于产生尽量大的分类间隔,从而产生小部分错分点。

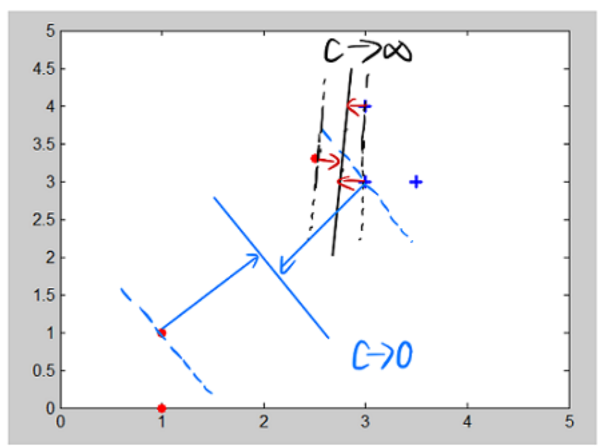

Fig. 2: Segmentation Surface under Different C Values

Q6:分别阐述线性可分向量机中的支持向量的几何意义和代数意义。

A1:从图像上出发,线性可分向量机的支持向量对应的是关于分割面等距的、最近的且在不错分前提下使得类间的几何距离相对最大的这样一批样本点。

A2:从代数上出发,对于线性可分样本,支持向量是指的使下列等式成立(约束条件)的样本点:

$$\begin{cases} w^T x_i + b = -1, & y_i = -1 \\ w^T x_i + b = 1, & y_i = 1 \end{cases}$$

合并得到:

$$y_i\left(w^T x_i+b\right)=1$$

引入拉格朗日函数后,解决对偶问题所引出的约束条件如下,使得该条件成立的样本点便称作支持向量:

$$\alpha_i^* \geq 0, \quad y_i\left(w^{*T} x_i+b^*\right)=1$$

Q7:结合图例,阐述线性可分支持向量机中的分类间隔的含义

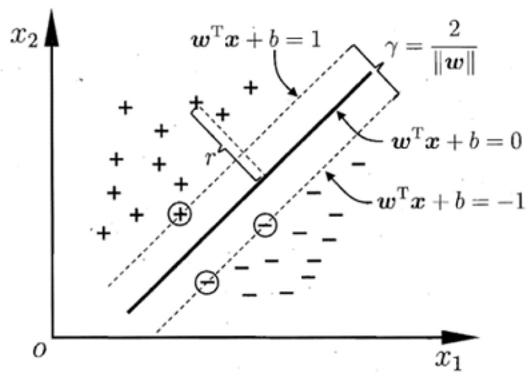

Fig. 3: Classification Interval of Linear Separable Vector Machines

A:对线性可分支持向量机,如上图所示,分类间隔描述了两类样本点集被分类面分隔开后,两类样本点与分类面之间的间隙大小和,即图中虚线间隙距离。从代数上,其实就是下式:

$$\gamma=\frac{2}{\|w\|_2}$$

我们的初始优化目标就是使分类间隔即得该式取得最大值。

Q8:请描述使用交叉验证对线性支持向量机的参数C进行设置的过程 。

A:交叉验证方法本质上就是划分部分样本作为验证集(validation),以实现对训练参数不断快速优化的过程。如下图所示:

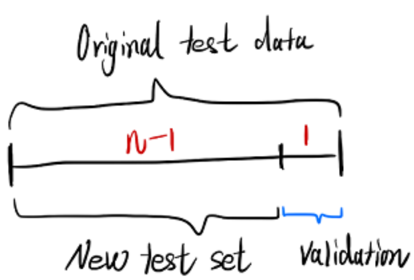

Fig. 4: Cross-Validation Diagram

1) 将原训练集划分为n份,随机取其中$n-1$份样本做为训练集训练分类器(SVM模型),剩下的一份样本作为验证集(validation),用来测试训练成果。
2) 取不同C值并不断重复1)中过程,取正确率最高的参数值作为C值。
3) 将2)部筛选出的C值最为参数值,在所有训练样本上重新学习SVM模型,获得相应的模型参数。

Q9: 将支持向量机对应的优化问题进行对偶化之后,有什么优势?

A:进行对偶化操作后,可以使得优化问题更容易求解,原问题的优化是一个二次规划问题,求解较麻烦,用拉格朗日乘子法转换后可以用smo等算法更简单地优化。

并且对偶化操作后,更加容易引入核函数,可以推广到非线性分类问题的求解。由于转换后的假设函数主要由内积运算构成,可以使用核函数简化特征映射到高维空间后的内积运算,高效地求解非线性问题。

II. Programming

Q1:对如下的 30 个数据进行 K-均值聚类,聚类个数设置为 K=4

1) 指出所使用的初始聚类中心,并报告在此条件下得到的最终聚类结果以及需要的迭代次数,对应的误差平方和。
2) 重新选择 3 组不同的初始聚类中心,给出对应的聚类结果和误差平方和。

A1:我们采用了基于$Python$3.9的模块化设计,代码均由$Python$编译。初始的聚类中心由程序随机筛选得出,程序运行的结果图如下图所示:

随机取得的初始聚类中心为:

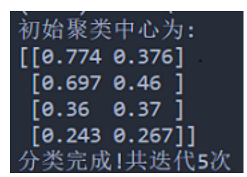

Fig. 5: Initial Cluster Center

初始分布情况为:

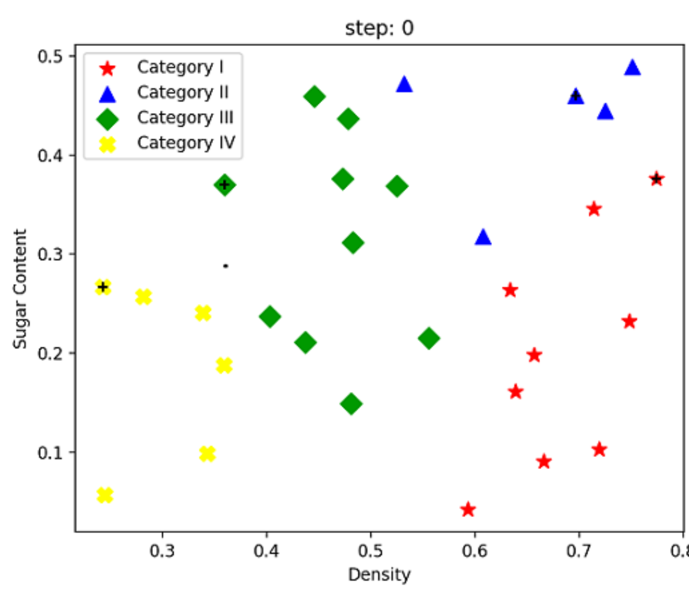

Fig. 6: Initial Situation

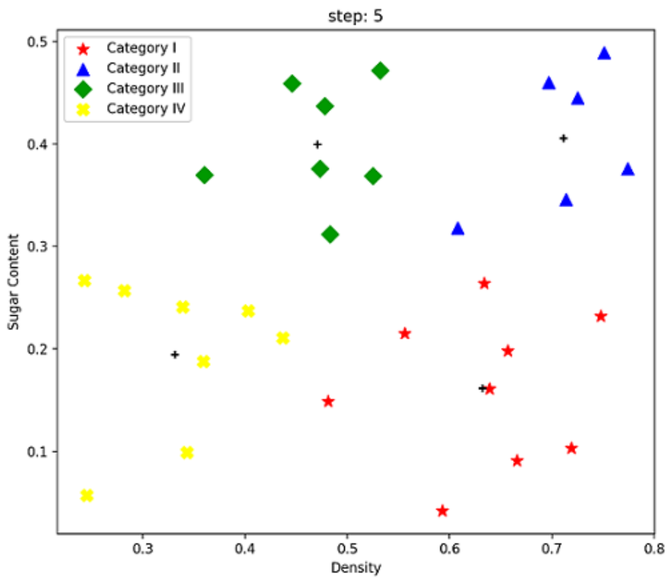

Fig. 7: Final Clustering Results

最终聚类的结果为如上图所示。可见,共需迭代5次,相应的最终聚类结果数据和误差平方和如图$Fig9$所示:

```
共分为4类,最终分类的结果为:
第 1 类最终聚类中心为:
[0.63255556 0.16166667]
误差平方和为:
0.09545422222222222
各类样本在原数组的编号分别是: [3, 5, 7, 9, 13, 14, 16, 17, 21]
第 2 类最终聚类中心为:
[0.7115     0.40566667]
误差平方和为:
0.040146833333333354
各类样本在原数组的编号分别是: [1, 2, 4, 22, 26, 29]
第 3 类最终聚类中心为:
[0.471      0.39928571]
误差平方和为:
0.03999142857142857
各类样本在原数组的编号分别是: [15, 23, 24, 25, 27, 2, 30]
第 4 类最终聚类中心为:
[0.331375 0.194625]
误差平方和为:
0.07642375000000001
各类样本在原数组的编号分别是: [6, 8, 10, 11, 12, 18, 19, 20]
```

Fig. 8: Final clustering Results[numerical value]

A2:随机选取不同的聚类中心数据如下:

初始聚类中心:

I	II	III
[[0.748 0.232]	[[0.657 0.198]	[[0.593 0.042]
[0.437 0.211]	[0.556 0.215]	[0.725 0.445]
[0.403 0.237]	[0.774 0.376]	[0.532 0.472]
[0.634 0.264]]	[0.245 0.057]]	[0.245 0.057]]

Fig. 9: Three groups of different initial cluster centers

他们最终的聚类结果由图$fig10 \sim fig12$展示,具体的运行结果反映在附录中。

```
分类完成!共迭代8次
共分为4类,最终分类的结果为:
第 1 类最终聚类中心为:
[0.6515  0.16325]
误差平方和为:
0.0694335
各类样本在原数组的编号分别是: [3, 5, 9, 13, 14, 16, 17, 21]
第 2 类最终聚类中心为:
[0.348      0.18955556]
误差平方和为:
0.09817422222222223
各类样本在原数组的编号分别是: [6, 7, 8, 10, 11, 12, 18, 19, 20]
第 3 类最终聚类中心为:
[0.471      0.39928571]
误差平方和为:
0.03999142857142857
各类样本在原数组的编号分别是: [15, 23, 24, 25, 27, 2, 30]
第 4 类最终聚类中心为:
[0.7115     0.40566667]
误差平方和为:
0.040146833333333354
各类样本在原数组的编号分别是: [1, 2, 4, 22, 26, 29]
```

Fig. 10: Final clustering Results[Group I]

```
分类完成!共迭代3次
共分为4类,最终分类的结果为:
第 1 类最终聚类中心为:
[0.6515  0.16325]
误差平方和为:
0.0694335
各类样本在原数组的编号分别是: [3, 5, 9, 13, 14, 16, 17, 21]
第 2 类最终聚类中心为:
[0.488125 0.389125]
误差平方和为:
0.06219575
各类样本在原数组的编号分别是: [4, 15, 23, 24, 25, 27, 2, 30]
第 3 类最终聚类中心为:
[0.7322 0.4232]
误差平方和为:
0.018069600000000012
各类样本在原数组的编号分别是: [1, 2, 22, 26, 29]
第 4 类最终聚类中心为:
[0.348      0.18955556]
误差平方和为:
0.09817422222222223
各类样本在原数组的编号分别是: [6, 7, 8, 10, 11, 12, 18, 19, 20]
```

Fig. 11: Final clustering Results[Group II]

```
分类完成!共迭代3次
共分为4类,最终分类的结果为:
第 1 类最终聚类中心为:
[0.63255556 0.16166667]
误差平方和为:
0.09545422222222222
各类样本在原数组的编号分别是: [3, 5, 7, 9, 13, 14, 16, 17, 21]
第 2 类最终聚类中心为:
[0.7115     0.40566667]
误差平方和为:
0.040146833333333354
各类样本在原数组的编号分别是: [1, 2, 4, 22, 26, 29]
第 3 类最终聚类中心为:
[0.471      0.39928571]
误差平方和为:
0.03999142857142857
各类样本在原数组的编号分别是: [15, 23, 24, 25, 27, 2, 30]
第 4 类最终聚类中心为:
[0.331375 0.194625]
误差平方和为:
0.07642375000000001
各类样本在原数组的编号分别是: [6, 8, 10, 11, 12, 18, 19, 20]
```

Fig. 12: Final clustering Results[Group III]

Q2:对上述数据集进行模糊 K-均值聚类,聚类个数设置为 K=4。指出使用的初始聚类中心、初始隶属度,报告在此初始化条件下的聚类结果(即：样本属于不同聚类的隶属度)以及需要的迭代次数

A:我们设定,当隶属度矩阵各个向量变化小于$\epsilon = 1e^{-7}$时,终止迭代。随机初始化的聚类中心由下图(限于篇幅,初始隶属度数据在附录中展示)给出:

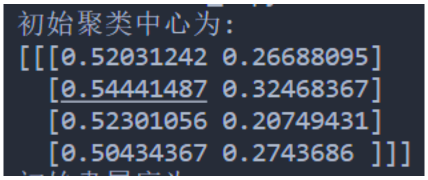
初始聚类中心为:
[[[0.52031242 0.26688095]
[0.54441487 0.32468367]
[0.52301056 0.20749431]
[0.50434367 0.2743686]]]

Fig. 13: Initial data [Fuzzy K-means]

最终分类结果如$Fig14$所示,可以看出,程序共进行了28次的迭代:

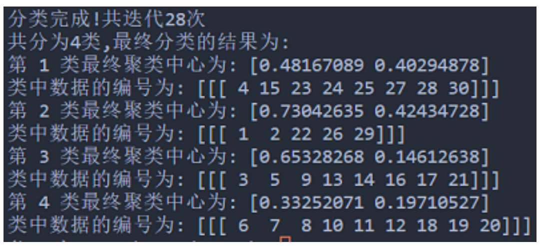
分类完成!共迭代28次
共分为4类,最终分类的结果为:
第 1 类最终聚类中心为: [0.48167089 0.40294878]
类中数据的编号为: [[[4 15 23 24 25 27 28 30]]]
第 2 类最终聚类中心为: [0.73042635 0.42434728]
类中数据的编号为: [[[1 2 22 26 29]]]
第 3 类最终聚类中心为: [0.65328268 0.14612638]
类中数据的编号为: [[[3 5 9 13 14 16 17 21]]]
第 4 类最终聚类中心为: [0.33252071 0.19710527]
类中数据的编号为: [[[6 7 8 10 11 12 18 19 20]]]

Fig. 14: Final clustering Results[Fuzzy K-means]

初始的分类图像:

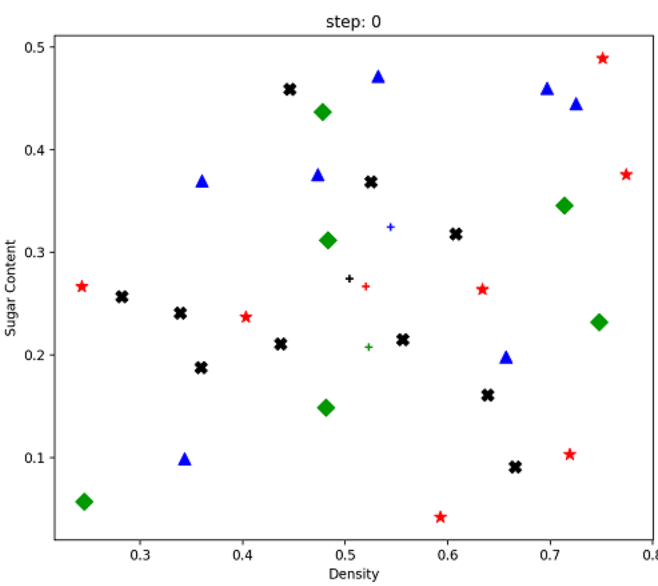

Fig. 15: Initial Situation

最终的分类情况:

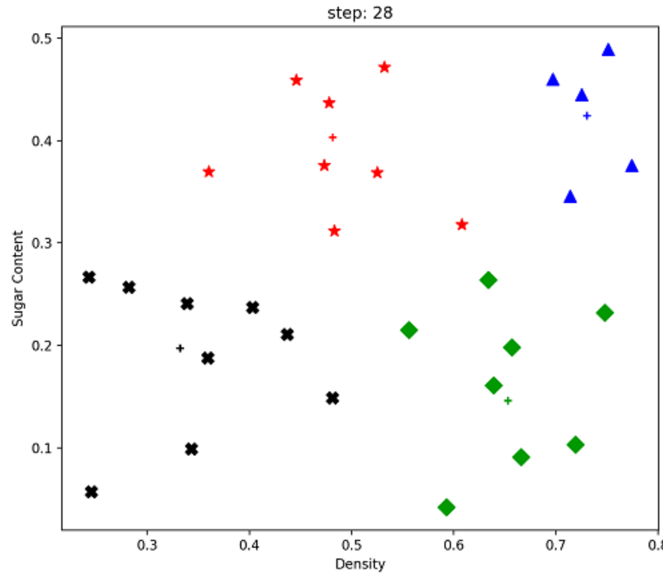

Fig. 16: Final Clustering Situation

Thinking:事实上,当程序迭代至$Step8$次时,聚类中心和样本隶属度虽然还在变化,但变化幅度过小,导致样本分布(几乎)不发生变化,原本隶属情况事实上已经不再更新。后续的计算虽然 可以在一定程度上精进参数准确度,但对于实际问题的解决几乎没有作用,浪费了算力。

调节初始参数和收敛条件ϵ值大小或许可以优化这一问题。

III. Appendix

本次作业中,所采用的拟合计算代码均是基于Matlab和Python3.9,相关的源码已经被开源于Github上: https://github.com/Alexiopro/First-year-of-UCAS/tree/main/UCAS/Source%20Code%20of%20Pattern%20Classification 供读者查用。

A1:三组不同初始聚类中心的样本点分布情况(初始态和最终聚类分布):

Group1:

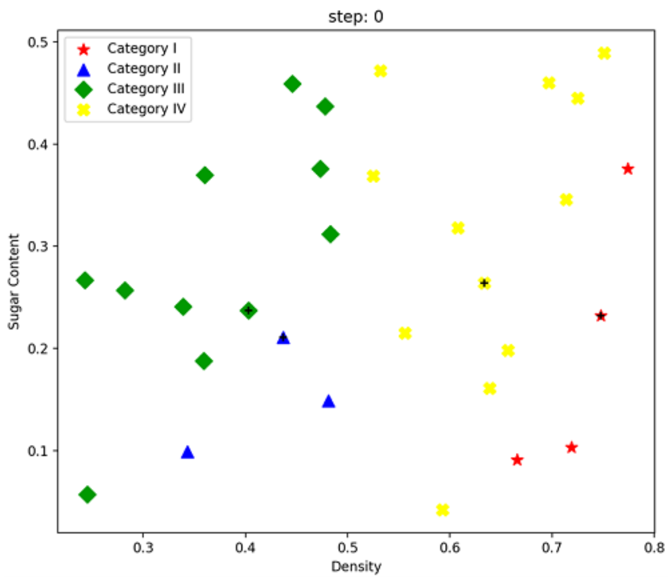

Fig. 17: Initial Situation

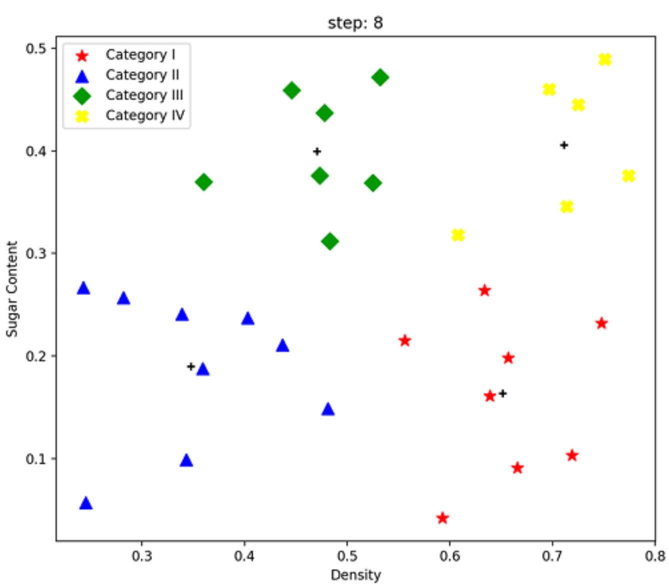

Fig. 18: Final Clustering Results

Group2:

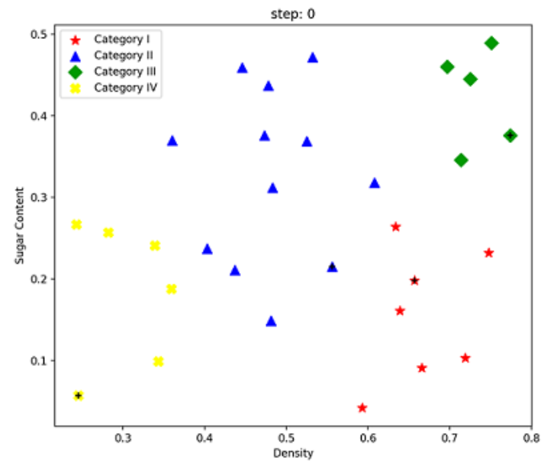

Fig. 19: Initial Situation

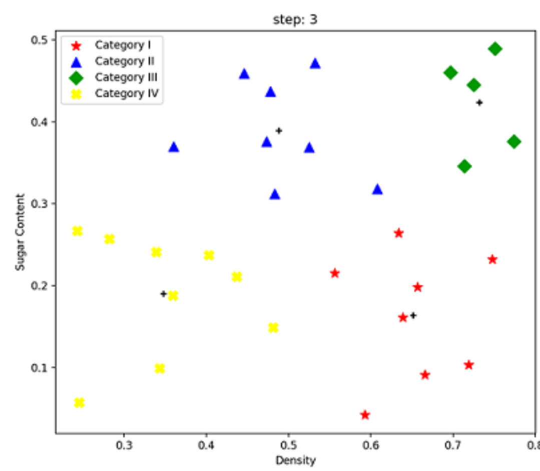

Fig. 20: Final Clustering Results

Group3:

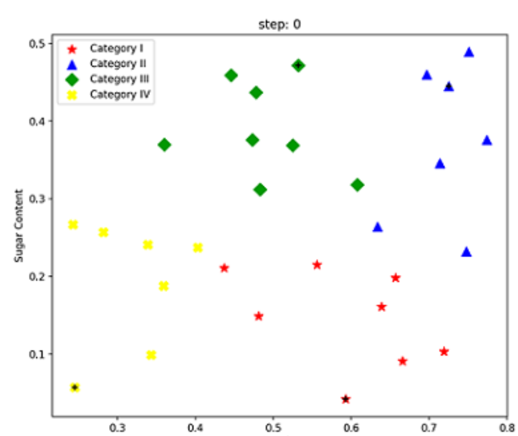

Fig. 21: Initial Situation

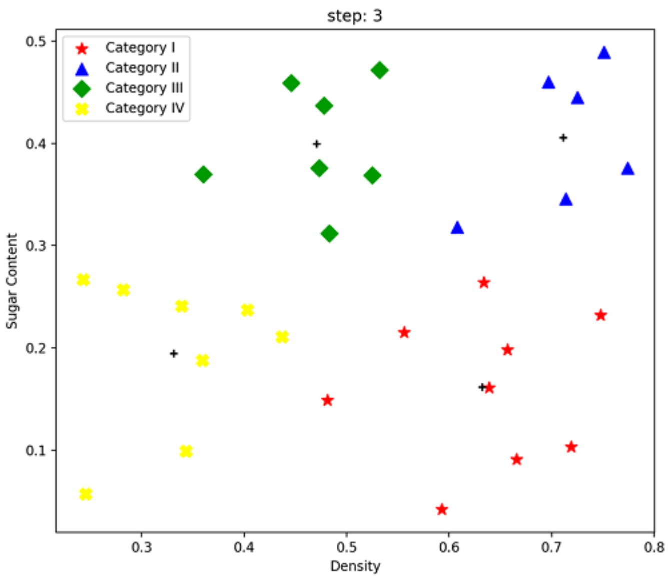

Fig. 22: Final Clustering Results

$\mathbb{A}2$:采用模糊$k-means$聚类的初始和优化后的隶属度矩阵如下图所示:

```
初始隶属度为:
[[0.07490854 0.42355264 0.07810232 0.4234365 ]
 [0.3756864  0.19359493 0.06096437 0.3697543 ]
 [0.36017524 0.24816811 0.21901291 0.17264374]
 [0.02925402 0.29790346 0.25633837 0.41650414]
 [0.29082569 0.02293578 0.32058237 0.36565616]
 [0.73631416 0.043752   0.15014406 0.06978978]
 [0.24154026 0.08395259 0.50862863 0.16587852]
 [0.41869305 0.1022416  0.01587568 0.46318967]
 [0.32443852 0.12194002 0.20631057 0.34731089]
 [0.39522649 0.231169   0.04048505 0.33311947]
 [0.21482459 0.10574626 0.60650069 0.07292845]
 [0.26790596 0.35105054 0.27126455 0.10977896]
 [0.00365705 0.21893539 0.16684275 0.61056481]
 [0.25685432 0.41173184 0.3159003  0.01551354]
 [0.27902177 0.31678877 0.13115262 0.27303684]
 [0.37514925 0.26071717 0.18784424 0.17628933]
 [0.40519315 0.07495214 0.40141364 0.11844107]
 [0.01776456 0.27653248 0.21332723 0.49237572]
 [0.26669603 0.15784837 0.1434785  0.4319771 ]
 [0.06903576 0.11619134 0.21472763 0.60004527]
 [0.10480584 0.08489043 0.56993464 0.24036909]
 [0.04297968 0.25578035 0.37819318 0.3230468 ]
 [0.22329982 0.23543762 0.28901648 0.25224609]
 [0.25977044 0.30258433 0.32437856 0.11326667]
 [0.1944773  0.2234368  0.17530693 0.40677897]
 [0.48293867 0.20701574 0.03934402 0.27070157]
 [0.36073583 0.36213171 0.22194526 0.0551872 ]
 [0.26181322 0.47513333 0.20017363 0.06287982]
 [0.30047024 0.31101022 0.19815145 0.19036809]
 [0.04181058 0.18487548 0.31757862 0.45573532]]
```

Fig. 23: Initial Membership matrix

```
[[1.70478845e-02 9.73669388e-01 6.65923585e-03 2.62349119e-03]
 [1.71815677e-02 9.54253705e-01 2.38387606e-02 4.72596626e-03]
 [1.44423649e-01 1.87094606e-01 6.19301690e-01 4.91800555e-02]
 [3.74652581e-01 3.16544120e-01 2.47881773e-01 6.09215266e-02]
 [1.58788385e-01 7.15813935e-02 6.49200157e-01 1.20430064e-01]
 [9.62098400e-02 1.41115170e-02 3.57288012e-02 8.53949842e-01]
 [1.27584333e-01 4.62552815e-02 3.58910127e-01 4.67250259e-01]
 [1.38915497e-01 2.72924011e-02 9.66444437e-02 7.37147658e-01]
 [6.90829185e-03 8.21703989e-03 9.77294969e-01 7.57969907e-03]
 [8.30917301e-02 1.57729627e-02 2.55066170e-02 8.75628690e-01]
 [6.96630595e-02 2.57531532e-02 7.02249628e-02 8.34358824e-01]
 [3.52520771e-02 1.16574935e-02 4.16472102e-02 9.11443219e-01]
 [8.76792598e-04 9.61742513e-04 9.97428109e-01 7.33355813e-04]
 [1.19631453e-02 1.67072370e-02 9.64022586e-01 7.30703151e-03]
 [6.54277993e-01 3.58951761e-02 3.73207746e-02 2.72506057e-01]
 [4.03999702e-02 3.32619640e-02 8.53717594e-01 7.26204715e-02]
 [1.39900741e-02 2.22224972e-02 9.51185581e-01 1.26018473e-02]
 [2.97811138e-03 6.41065178e-04 1.82717575e-03 9.94553648e-01]
 [1.44068173e-02 2.26081879e-03 4.70805837e-03 9.78624305e-01]
 [4.36455583e-02 7.50427261e-03 1.31818669e-02 9.35668302e-01]
 [6.08907471e-02 2.27197350e-01 6.82721209e-01 2.91906936e-02]
 [4.71728895e-02 8.73873289e-01 6.77062485e-02 1.12475732e-02]
 [7.85692711e-01 4.24390610e-02 6.06268700e-02 1.11241358e-01]
 [9.89398768e-01 4.79296058e-03 2.18014599e-03 3.62812503e-03]
 [9.43558661e-01 2.56458828e-02 1.54699620e-02 1.53254946e-02]
 [2.14044345e-02 9.62627352e-01 1.15336491e-02 4.43456403e-03]
 [8.71809324e-01 8.55469791e-02 2.06548409e-02 2.19888560e-02]
 [9.91578495e-01 2.62821304e-03 1.96447314e-03 3.82881839e-03]
 [1.45858080e-03 9.97456804e-01 8.13649524e-04 2.70965719e-04]
 [9.51767826e-01 1.93141547e-02 9.40358553e-03 1.95144339e-02]]
```

Fig. 24: Final Membership matrix