模式识别导论第四次作业——简述题

杨登天-202028015926089-微电子研究所

1. 请描述使用高斯混合模型进行数据聚类的过程。

解答：

步骤一、定义相关参数——$P(w_j)$，指j类别的已知的先验概率，其中$j \in \{1,2,3 \dots C\}$。$P(x|w_j,\theta_j)$，指j类别的条件概率密度函数，其形式已知，其中x表示待分类样本，θ_j为高斯分类模型的未知向量参数，且有$\theta = \{\theta_j | j \in \{1,2,3 \dots C\}\}$。样本分别来自于$C$个类别，但具体标签未知。

步骤二、确定目标函数——首先通过类先验概率$P(w_j)$随机选择一个类别，然后通过类条件概率密度$P(x|w_j,\theta_j)$随机选择一个样本。设总体样本的概率密度函数为

$$P(x|\theta) = \sum_{j=1}^{C} P(x|w_j,\theta_j)P(w_j)$$

上述密度函数是混合密度，称条件概率密度函数$P(x|w_j,\theta_j)$为成分密度，称先验概率为混合参数。

给定一个包含n个无类别标签的数据集$D = \{x_1, x_2, x_3 \dots x_n\}$，假定这些样本独立地从总体样本的概率密度函数模型中采样得到，那么根据这些样本，采用最大似然估计对θ进行估计，因此有如下对数似然函数

$$f(\theta) = \ln\big(p(D|\theta)\big) = \sum_{k=1}^{n} \ln\left(p(x_k|\theta)\right) = \sum_{k=1}^{n} \ln\left(\sum_{j=1}^{C} P(x|w_j,\theta_j)P(w_j)\right)$$

步骤三、求解参数——令$f(\theta)$对θ的梯度为 0，即可得到待估计的$\hat{\theta}$。

步骤四、对样本表达其后验概率以完成聚类——

$$P(w_j|x,\theta) = \frac{P(x|w_j,\theta)P(w_j)}{\sum_{j=1}^{C} P(x|w_j,\theta)P(w_j)}, j \in \{1,2,3 \dots C\}$$

那么对于剩下所有未分类的样本仅需代入C个类别的后验概率密度函数进行判断，如下

$$P(w_k|x,\theta) > P(w_i|x,\theta), in\ which\ i \neq k\ \&\&\ i \in \{1,2,3 \dots C\}$$

那么将上述未分类样本x分类到第k类。

解答完毕

2. 对于数据：

$$x_1 = (4,5)^T, x_2 = (1,4)^T, x_3 = (0,1)^T$$
$$x_4 = (5,0)^T, x_5 = (4,1)^T, x_6 = (0,6)^T$$

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）$\{x_1, x_2, x_3, x_6\}, \{x_4, x_5\}$

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？给出计算过程。

解答：此处采用误差平方和的最小平方差准则，那么就有

$$J_e = \sum_{i=1}^{C} J_i$$

其中：$J_i = \sum_{x \in D_i} ||x - m_i||^2, m_i = \frac{1}{n_i} \sum_{x \in D_i} x$

对第一种聚类方式求解：

$$m_1 = \frac{1}{n_1} \sum_{x \in D_1} x = \frac{1}{3} \times [(4,5)^T + (1,4)^T + (0,6)^T] = \left(\frac{5}{3}, 5\right)^T$$

$$J_1 = \sum_{x \in D_1} ||x - m_1||^2 = (\frac{7}{3})^2 + \left(1 + \left(\frac{2}{3}\right)^2\right) + \left(1 + \left(\frac{5}{3}\right)^2\right) = \frac{96}{9}$$

$$m_2 = \frac{1}{n_2} \sum_{x \in D_2} x = \frac{1}{3} \times [(0,1)^T + (5,0)^T + (4,1)^T] = \left(3, \frac{2}{3}\right)^T$$

$$J_2 = \sum_{x \in D_2} ||x - m_2||^2 = \left(9 + \left(\frac{1}{3}\right)^2\right) + \left(4 + \left(\frac{2}{3}\right)^2\right) + \left(1 + \left(\frac{1}{3}\right)^2\right) = \frac{132}{9}$$

$$J_{e1} = J_1 + J_2 = \frac{228}{9} = 25.33$$

对第二种聚类方式求解：

$$m_1 = \frac{1}{n_1} \sum_{x \in D_1} x = \frac{1}{3} \times [(4,5)^T + (5,0)^T + (4,1)^T] = \left(\frac{13}{3}, 2\right)^T$$

$$J_1 = \sum_{x \in D_1} ||x - m_1||^2 = \left(9 + \left(\frac{1}{3}\right)^2\right) + \left(4 + \left(\frac{2}{3}\right)^2\right) + \left(1 + \left(\frac{1}{3}\right)^2\right) = \frac{132}{9}$$

$$m_2 = \frac{1}{n_2} \sum_{x \in D_2} x = \frac{1}{3} \times [(1,4)^T + (0,1)^T + (0,6)^T] = \left(\frac{1}{3}, \frac{11}{3}\right)^T$$

$$J_2 = \sum_{x \in D_2} ||x - m_2||^2 = \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) + \left(\left(\frac{1}{3}\right)^2 + \left(\frac{8}{3}\right)^2\right) + \left(\left(\frac{1}{3}\right)^2 + \left(\frac{7}{3}\right)^2\right) = \frac{120}{9}$$

$$J_{e2} = J_1 + J_2 = \frac{252}{9} = 28$$

对第三种聚类方式求解：

$$m_1 = \frac{1}{n_1} \sum_{x \in D_1} x = \frac{1}{4} \times [(4,5)^T + (1,4)^T + (0,1)^T + (0,6)^T] = \left(\frac{5}{4}, 4\right)^T$$

$$J_1 = \sum_{x \in D_1} ||x - m_1||^2 = \left(1 + \left(\frac{11}{4}\right)^2\right) + \left(0 + \left(\frac{1}{4}\right)^2\right) + \left(9 + \left(\frac{5}{4}\right)^2\right) + \left(4 + \left(\frac{5}{4}\right)^2\right) = \frac{99}{4}$$

$$m_2 = \frac{1}{n_2} \sum_{x \in D_2} x = \frac{1}{2} \times [(5,0)^T + (4,1)^T] = \left(\frac{9}{2}, \frac{1}{2}\right)^T$$

$$J_2 = \sum_{x \in D_2} ||x - m_2||^2 = \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) + \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 1$$

$$J_{e3} = J_1 + J_2 = \frac{103}{4} = 25.75$$

结论：由于$J_{e1} < J_{e3} < J_{e2}$，所以按照最小平方和误差，选择第一种聚类划分更好。

解答完毕

3. 请阐述K均值聚类和模糊K均值聚类的异同。

   解答：

   从原理上看:

   1、 都是依据距离最近准则划分类别，但模糊K均值聚类中的距离是指待分类点到所有分类中心的距离并以某种隶属关系的量化值加权，K均值聚类则是以十分肯定的隶属关系分类。
   2、 都是需要初始化聚类中心，但模糊K均值聚类额外增加样本隶属度参数。
   3、 都是需要更新聚类中心，但模糊K均值聚类需要通过样本隶属度参数和样本集合更新聚类中心，而K均值聚类则仅需要当前各分类样本求均值或者其他方式。在复杂度上，模糊K均值聚类比K均值聚类更复杂，且运算量更大。
   4、 都能完成样本的分类，但模糊K均值聚类同时输出样本点归属于各类的隶属度，因此还需要比较隶属度大小才能确定样本点的类别。而K均值聚类无需此步骤。

   从算法性能上看:

   1、 都对初始值敏感
   2、 都需要知道类别数
   3、 都需要处理不好的点
   4、 都可以很好得处理密集簇和大数据集
   5、 模糊K均值聚类比K均值聚类鲁棒性更好

解答完毕

4. 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：

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

   其中J_i为移动前属于第i类样本的误差平方和

   解答：假设样本$\hat{x}$从类别i到类别j，此时i类的类别中心发生变化为

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

   那么属于第i类的样本点引起的误差平方和将变成

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

$$= \sum_{x\in D_i} \left(\|x - m_i\|^2 + \frac{2}{n_i - 1}(\hat{x} - m_i)^T(x - m_i) + \frac{\|\hat{x} - m_i\|^2}{(n_i - 1)^2} \right)$$

$$- \frac{{n_i}^2\|\hat{x} - m_i\|^2}{(n_i - 1)^2}$$

$$= \sum_{x\in D_i} \|x - m_i\|^2 + \frac{2}{n_i - 1}(\hat{x} - m_i)^T \left(\sum_{x\in D_i} x - \sum_{x\in D_i} m_i \right) + n_i \frac{\|\hat{x} - m_i\|^2}{(n_i - 1)^2}$$

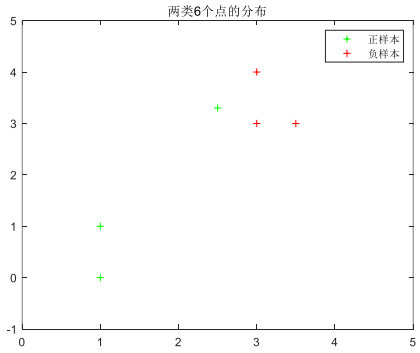
$$- \frac{{n_i}^2\|\hat{x} - m_i\|^2}{(n_i - 1)^2} = J_i + 0 - \frac{n_i\|\hat{x} - m_i\|^2}{n_i - 1}$$

所以最后得到

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

解答完毕

5. 已知正样本点$x_1 = (1,1)^T, x_2 = (1,0)^T, x_3 = (2.5,3.3)^T$，负样本点$x_4 = (3,3)^T, x_5 = (3,4)^T, x_6 = (3.5,3)^T$，它们的分布如下图所示



(1) 请写出线性支持向量机需要求解的原问题和对偶问题
(2) 当C取值很大（比如$C \to +\infty$）时，定性画出会得到的决策面，并解释原因
(3) 当C取值很小（比如$C \to 0$）时，定性画出会得到的决策面，并解释原因

解答：

（1） 请写出线性支持向量机需要求解的原问题和对偶问题

其原问题如下：

对于训练数据集：$(x_i, y_i), i = 1,2,\dots,N$

其最大间隔分类面(w^*, b^*)通过求解如下问题获得：

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

对于其约束条件为

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1,2,\dots,N$$
$$\xi_i \geq 0, i = 1,2,\dots,N$$

其对偶问题如下：

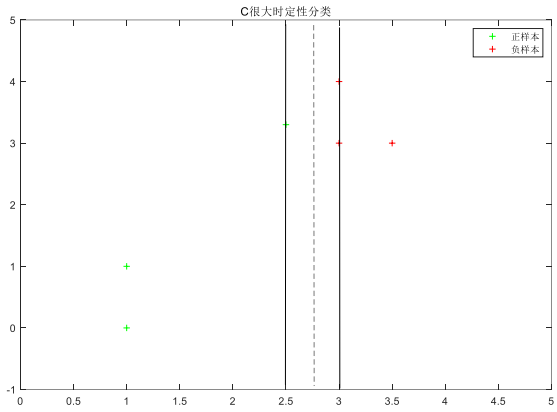
$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

对于其约束条件为

$$\sum_{j=1}^{N}\alpha_i y_i=0$$

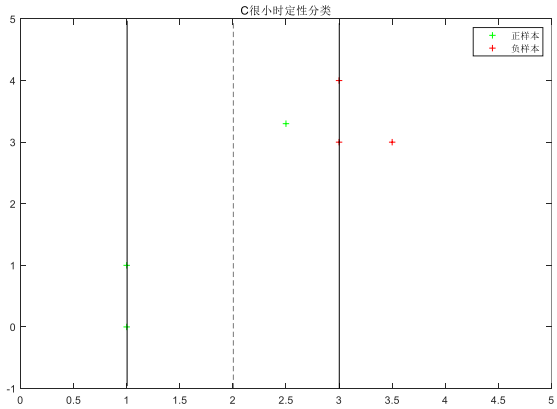
$$0\le\alpha_i\le C, i=1,2,\dots,N$$

(2) 当C取值很大（比如$C\to+\infty$）时，定性画出会得到的决策面，并解释原因



因为C取值很大更关心错误样本，倾向于产生没有错分样本的分界面。

(3) 当C取值很小（比如$C\to 0$）时，定性画出会得到的决策面，并解释原因

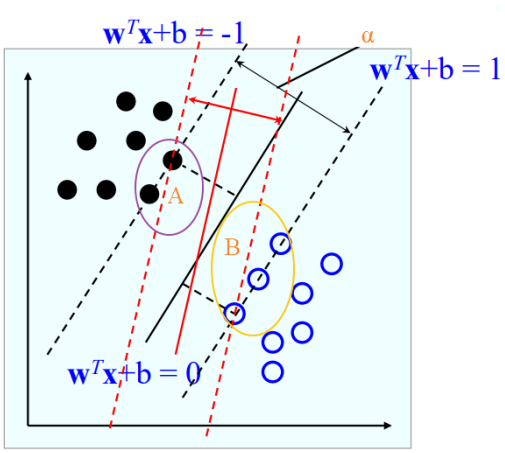


因为C取值很小更加关心分类间隔，倾向于产生大间隔的分界面。

解答完毕

6. 结合图例，阐述线性可分支持向量机中的支持向量的概念。

解答：下图来自樊彬老师《支持向量机》一节 PPT10/81.



上图中椭圆区域 A 和 B 内的两组点都是“支持向量”。

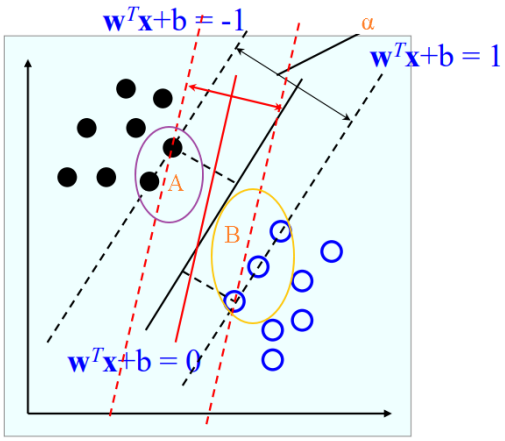
$$\begin{cases} w^T x_i + b = -1, y_i = -1 & if\ points\ in\ A \\ w^T x_i + b = +1, y_i = +1 & if\ points\ in\ B \end{cases}$$

上述两组点均满足对应的等式，其中y_i是人为赋予的点的分类号，假定黑点是负类，白点为正类。

解答完毕

7. 结合图例，阐述线性可分支持向量机中的分类间隔的涵义。

解答：下图来自樊彬老师《支持向量机》一节 PPT10/81.



上图中的α就是最大分类间隔，所谓间隔就是两组点各自到判别直线最短距离之和。

解答完毕

8. 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。

解答：

步骤一、对不同参数 C 操作步骤二、三和四。

步骤二、将训练集分成 n 等份，依次进行 n 次分类器学习-分类器测试；

步骤三、每次选择 n-1 份数据训练分类器，在剩下的 1 份数据集上进行测试；

步骤四、交叉验证的正确率为 n 次测试的平均结果。

步骤五、取步骤一中最高正确率的 C 作为最终的参数 C

解答完毕

9. 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？

解答：

1. 将原问题转变为更容易解决的问题——对偶化问题之后的结果，使得支持向量机学习问题转变为求解凸二次规划问题。
2. 减少样本参与计算数量——对偶化问题仅考虑少数“支持向量”，少量样本可以学习得到很强的分类模型。
3. 此外在非线性支持向量机模型中，求解对偶问题只与特征空间中样本的内积有关，通过对偶问题的解构造支持向量机分类平面也只与特征空间中样本的内积有关。

解答完毕