模式识别导论第四次作业

翁晨阳

1. 请描述使用EM算法进行高斯混合模型聚类的过程

输入：观测数据集Y，高斯混和模型

1. 取参数的初始值开始迭代；
2. E步，依据当前模型参数，计算模型k对y_j响应度

$$\widehat{\gamma}_{jk} = \frac{\alpha_k \varphi(y_j|\theta_k^{(i)})}{\sum_{k=1}^{K} \alpha_k \varphi(y_j|\theta_k^{(i)})} \\ j = 1, 2, \ldots, m;\ k = 1, 2, \ldots, K \tag{1.1}$$

3. M步，计算$i+1$次迭代的模型参数

$$\begin{aligned} \mu_k^{(i+1)} &= \frac{\sum_{j=1}^{m} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{m} \hat{\gamma}_{jk}}; \\ (\sigma_k^2)^{(i+1)} &= \frac{\sum_{j=1}^{m} \hat{\gamma}_{jk} (y_j - \mu_k^{(i+1)})^2}{\sum_{j=1}^{m} \hat{\gamma}_{jk}}; \\ \alpha_k^{(i+1)} &= \frac{\sum_{j=1}^{m} \hat{\gamma}_{jk}}{m}; \end{aligned} \tag{1.2}$$

4. 重复第二第三步，直到收敛。

2. 对于数据：$\mathbf{x}_1 = (4,5)^T$,$\mathbf{x}_2 = (1,4)^T$,$\mathbf{x}_3 = (0,1)^T$,$\mathbf{x}_4 = (5,0)^T$,$\mathbf{x}_5 = (4,1)^T$, $\mathbf{x}_6 = (0,6)^T$,现有以下三种聚类划分，使用最小平方和误差准则下哪种划分更好?

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

计算聚类中心：$m_1 = (\frac{5}{3}, 5)^T,_2 = (3, \frac{2}{3})^T$

平方和误差：
$J_1 = ||(x_1 - m_1)||^2 + ||(x_2 - m_1)||^2 + ||(x_6 - m_1)||^2 + ||(x_3 - m_2)||^2 + ||(x_4 - m_2)||^2 + ||(x_5 - m_2)||^2 = 25.33$

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

计算聚类中心：$m_1 = (\frac{13}{3}, 2)^T, m_2 = (\frac{1}{3}, \frac{11}{3})^T$

平方和误差：
$J_2 = ||(x_1 - m_1)||^2 + ||(x_4 - m_1)||^2 + ||(x_5 - m_1)||^2 + ||(x_2 - m_2)||^2 + ||(x_3 - m_2)||^2 + ||(x_6 - m_2)||^2 = 28$

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

计算聚类中心：$m_1 = (1.25, 4)^T, m_2 = (4.5, 0.5)^T$

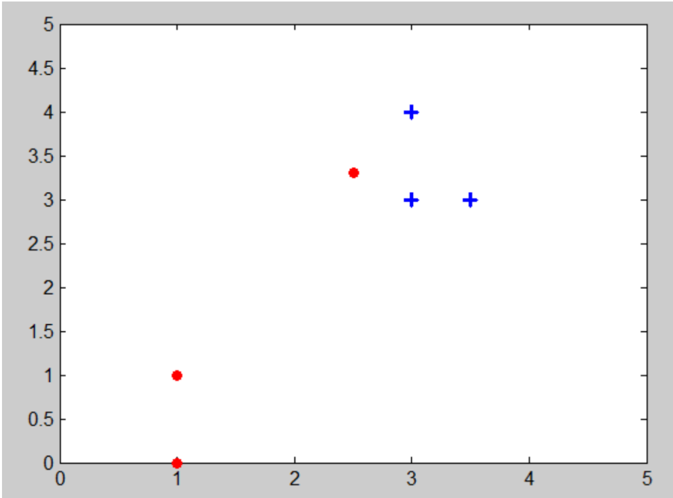
平方和误差：
$J_3 = ||(x_1 - m_1)||^2 + ||(x_2 - m_1)||^2 + ||(x_3 - m_1)||^2 + ||(x_6 - m_1)||^2 + ||(x_4 - m_2)||^2 + ||(x_5 - m_2)||^2 = 25.75$

因为$j_1 < j_3 < j_2$，所以第一钟聚类划分更好。

3. 请阐述K均值聚类和模糊K均值聚类的关系

K均值聚类就是模糊K均值聚类的特殊情况，在K均值聚类情况下，样本对离样本最近的簇隶属度为1，对其它样本隶属度为0.

4. 已知正样本点 $\mathbf{x}_1=(1,1)^T$，$\mathbf{x}_2=(1,0)^T$，$\mathbf{x}_3=(2.5,3.3)^T$，负样本点 $\mathbf{x}_4=(3,3)^T$，$\mathbf{x}_5=(3,4)^T$，$\mathbf{x}_6=(3.5,3)^T$， 它们的分布如下图所示



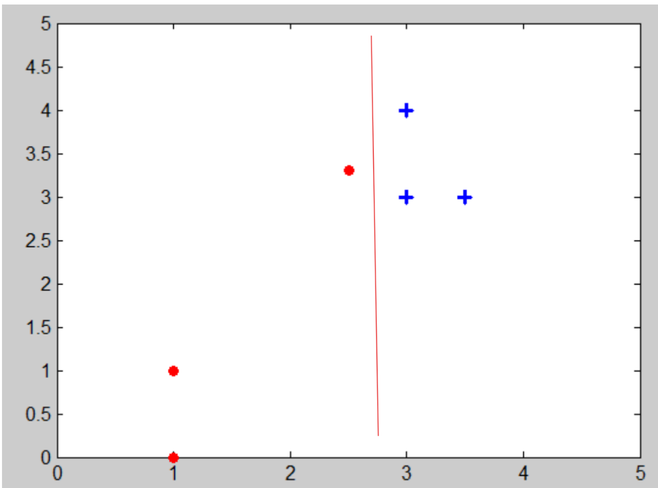
1. **线性支持向量机需要求解的原问题和对偶问题**
    1. 原问题：

$$
\begin{gathered}
\min_{W,b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \\
s.t.\ y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i,\ i = 1,2,\dots,N \\
\xi_i \geq 0;\ i = 1,2,\dots,N
\end{gathered} \tag{4.1}
$$

    2. 对偶问题：

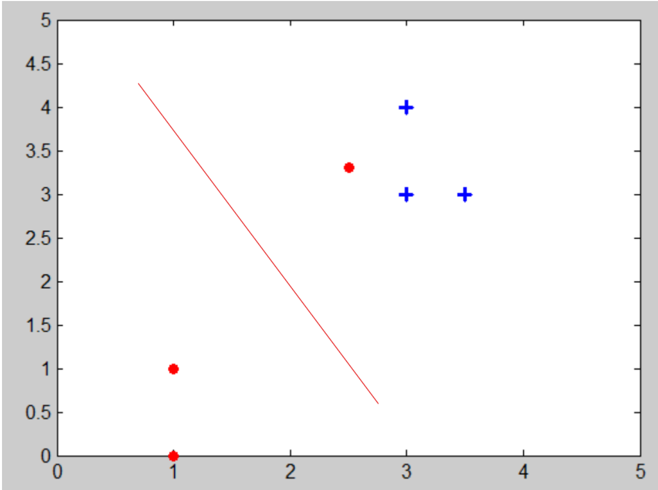
$$
\begin{gathered}
\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\
s.t.\ \sum_{i=1}^{N}\alpha_i y_j = 0,\ 0 \leq \alpha_i \leq C,\ i = 1,2,\dots,N
\end{gathered} \tag{4.2}
$$

2. **当 C 取值很大（比如 C->+∞）时，定性画出会得到的决策面，并解释原因**



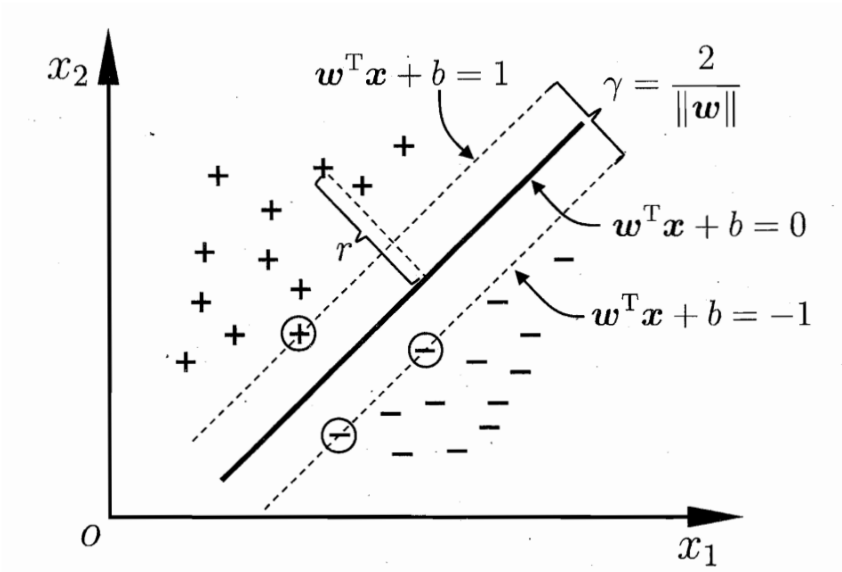
当C取值很大的时候，决策面会过分考虑错分样本，倾向于产生没有错分样本的决策面，容易产生过拟合。

3. **当 C 取值很小（比如 C->0）时，定性画出会得到的决策面，并解释原因**



当C取值很小的时候，决策面会更考虑分类间隔，会产生分类间隔很大的决策面

5. 结合图例，阐述线性可分支持向量机中的支持向量的概念。



如图所示，距离超平面最近的这几个训练样本点使式（5.1）的等号成立，称为“支持向量”，支持向量距离决策面最近。

$$\begin{cases}\mathbf{w}^T\mathbf{x}_i+b\geq+1,\ y_i=+1;\\ \mathbf{w}^T\mathbf{x}_i+b\leq-1,\ y_i=-1;\end{cases}\tag{5.1}$$