The University of Hong Kong

Department of Electrical and Electronic Engineering

ELEC 6008 Pattern recognition and machine learning (2018-2019)

Assignment **2**: Written Assignment

Student: *ZHAO Yunqing*
University ID: *3035541156*

Question 1:

a) Firstly, write down the respective likelihood as follow:

$$p(x|\omega_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+2)^2}{2}} \qquad p(x|\omega_2)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-5)^2}{8}}$$

The priori:

$$P(\omega_1)=0.8 \text{ and } P(\omega_2)=0.2$$

i) The Maximum Likelihood Classifier:

Typically, we decide the result is ω_1 if $p(x|\omega_1)>p(x|\omega_2)$, otherwise decide ω_2.

So, first we take logarithm on both sides of the likelihood:

$$\log p(x|\omega_1)=\log p(x|\omega_2)$$

We got:

$$\log\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+2)^2}{2}}=\log\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-5)^2}{8}}$$

$$\Rightarrow \log\frac{1}{\sqrt{2\pi}}-\frac{1}{2}(x+2)^2=\log\frac{1}{2\sqrt{2\pi}}-\frac{1}{8}(x-5)^2$$

$$\Rightarrow \log\frac{1}{\sqrt{2\pi}}-\log\frac{1}{2\sqrt{2\pi}}=\frac{1}{2}(x+2)^2-\frac{1}{8}(x-5)^2$$

$$\Rightarrow 3x^2+26x-9-8\log 2=0$$

$$\Rightarrow x_1=\frac{-13+\sqrt{169+3(9+8\log 2)}}{3}, x_2=\frac{-13-\sqrt{169+3(9+8\log 2)}}{3}$$

We decide the ω_1 if the equality is **greater** than 0 otherwise we decide ω_2.

Hence, we get Maximum Likelihood Classifier and two boundaries .

ii) Using Bayes rule to find the MAP Classifier:

$$p(\omega_i|x)=\frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

First, we take logarithm to this function:

$$\log P(\omega_i \mid x) = \log p(x \mid \omega_i) + \log P(\omega_i) - \log(p(x))$$

The MAP Classifier is :

$$\log(P(\omega_1 \mid x)) = \log(P(\omega_2 \mid x))$$

Combining the above two equalities, we get:

$$\log(p(x \mid \omega_1)) + \log(P(\omega_1)) = \log(p(x \mid \omega_2)) + \log(P(\omega_2))$$

$$\Rightarrow \log \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+2)^2}{2}} + \log 0.8 = \log \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-5)^2}{8}} + \log 0.2$$

$$\Rightarrow \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(x+2)^2 + \log 0.8 = \log \frac{1}{2\sqrt{2\pi}} - \frac{1}{8}(x-5)^2 + \log 0.2$$

$$\Rightarrow \log 8 = \frac{1}{2}(x+2)^2 - \frac{1}{8}(x-5)^2$$

$$\Rightarrow 3x^2 + 26x - 8\log 8 - 9 = 0$$

$$\Rightarrow x_1 = \frac{-13 - \sqrt{169 + 3(9 + 8\log 8)}}{3}, x_2 = \frac{-13 + \sqrt{169 + 3(9 + 8\log 8)}}{3}$$

We decide the ω_1 if the equality is **greater** than 0 otherwise we decide ω_2.

Hence, we get the functions of MAP Classifiers and two boundaries respectively.

iii) From the given loss functions, we get:

$$\lambda(\alpha_1 \mid \omega_1)=0, \lambda(\alpha_1 \mid \omega_2)=5, \lambda(\alpha_2 \mid \omega_1)=1, \lambda(\alpha_2 \mid \omega_2)=0$$

iv) In order to find the Bayes Minimum Risk Classifier, we should consider the loss function in iii):

$$\lambda(\alpha_1 \mid \omega_1)=0, \lambda(\alpha_1 \mid \omega_2)=5, \lambda(\alpha_2 \mid \omega_1)=1, \lambda(\alpha_2 \mid \omega_2)=0$$

Then, because:

$$\begin{aligned} R(\alpha_1 \mid x) &= \lambda(\alpha_1 \mid \omega_1)P(\omega_1 \mid x) + \lambda(\alpha_1 \mid \omega_2)P(\omega_2 \mid x) \\ &= 0 \cdot P(\omega_1 \mid x) + 5 \cdot P(\omega_2 \mid x) = 5 \cdot P(\omega_2 \mid x) \end{aligned}$$

The same as above:

$$\begin{aligned} R(\alpha_2 \mid x) &= \lambda(\alpha_2 \mid \omega_2)P(\omega_2 \mid x) + \lambda(\alpha_2 \mid \omega_1)P(\omega_1 \mid x) \\ &= 0 \cdot P(\omega_2 \mid x) + 1 \cdot P(\omega_1 \mid x) = P(\omega_1 \mid x) \end{aligned}$$

Let $R(\alpha_1 \mid x)=R(\alpha_2 \mid x)$ and taking logarithms on both sides, we get:

$$\log P(\omega_1 \mid x) = \log(5 \cdot P(\omega_2 \mid x))$$

$$\Rightarrow \log(p(x \mid \omega_1)) + \log(P(\omega_1)) - \log(p(x)) = \log 5 + \log(p(x \mid \omega_2)) + \log(P(\omega_2)) - \log(p(x))$$

$$\Rightarrow \log \frac{8}{5} = \frac{1}{2}(x+2)^2 - \frac{1}{8}(x-5)^2$$

$$\Rightarrow 3x^2 + 26x - 8\log \frac{8}{5} - 9 = 0$$

Then we get:

$$x_1 = \frac{-13 - \sqrt{169 + 3(9 + 8\log\frac{8}{5})}}{3}, x_2 = \frac{-13 + \sqrt{169 + 3(9 + 8\log\frac{8}{5})}}{3}$$

We decide the ω_1 if the equality **less** than 0 otherwise we decide ω_2.

Now, we get the functions of the Bayes Minimum Risk Classifier and the corresponding boundaries.

v) ***No.*** Since it not reasonable to regard all kinds of errors as the same weight in some special situation like cancer cells recognition or diagnose of the patients, which means we cannot refer to minimum error rate any more.

So, every kind of decision should be multiplied by a weight, depending on the importance of potential recognition error, this is the so called "Bayes Minimum Risk Classifier" in this problem, rather than the minimum error rate.

b)

i) By choosing to minimize

$$J_q(a) = \sum_{y \in Y_C} -a^T \tilde{y}$$

Y_C contains the mis-classified samples such that $\mathbf{a}^T \tilde{y} < 0$.

The Gradient Descent of the loss function:

$$-\nabla_a J_a(a^{(k)}) = \sum_{y \in Y_C} \tilde{y}$$

Hence, we have the update rule for ***a*** iteratively:

$$\begin{aligned} a^{(k+1)} &= a^{(k)} - \rho^{(k)} \nabla_a J_a(a^{(k)}) \\ &= a^{(k)} + \rho^{(k)} \sum_{y \in Y_C} \tilde{y} \end{aligned}$$

ii) Firstly, by multiplying -1 to those labelled Class 2, we find:

$$\tilde{Y} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 4 & 3 & -6 & -9 \\ 1 & 2 & -8 & -9 \end{bmatrix}^T$$

By initializing $a^{(1)} = 0$, and use $a^{(k+1)} = a^{(k)} + \rho^{(k)} \sum_{y \in Y_C} \tilde{y}$, we have:

$$\begin{aligned} a^{(2)} &= \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T + 1 \cdot \left(\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ -6 \\ -8 \end{bmatrix} + \begin{bmatrix} -1 \\ -9 \\ -9 \end{bmatrix} \right) \\ &= \begin{bmatrix} 0 & -8 & -14 \end{bmatrix}^T \end{aligned}$$

Then we verify the classification accuracy:

$$\tilde{Y}a^{(2)} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 4 & 3 & -6 & -9 \\ 1 & 2 & -8 & -9 \end{bmatrix}^T \cdot \begin{bmatrix} 0 & -8 & -14 \end{bmatrix}^T$$
$$= \begin{bmatrix} -46 & -52 & 160 & 198 \end{bmatrix}^T$$

Apparently, there being misclassified samples: $\tilde{y}_1\ \tilde{y}_2$, so we update $a^{(k)}$:

$$a^{(3)} = a^{(2)} + \tilde{y}_1\ \tilde{y}_2$$
$$= \begin{bmatrix} 0 \\ -8 \\ -14 \end{bmatrix} + \left(\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \right)$$
$$= \begin{bmatrix} 2 & -1 & -11 \end{bmatrix}^T$$

iii) The object function of the soft-margin SVM is:

$$\min_{\omega,\omega_0} \|\omega\|_2^2 + C\sum_{i=1}^{N} \max(0, 1 - z_i(\omega^T x_i + \omega_0)), z_i = 1\ or\ -1\ (can\ be\ swapped)$$

First, $k = 1$. We find $\ell_i(\tilde{\omega})$ and $\partial\ell_i(\tilde{\omega})$:

$$\ell(\tilde{\omega}^{(1)}) = \max(0, 1 - 0) = 1, \quad Then: \partial\ell_1(\tilde{\omega}^{(1)}) = -\begin{bmatrix} 1 & 4 & 1 \end{bmatrix}^T$$
$$\partial\ell_2(\tilde{\omega}^{(1)}) = -\begin{bmatrix} 1 & 3 & 2 \end{bmatrix}^T, \partial\ell_3(\tilde{\omega}^{(1)}) = \begin{bmatrix} 1 & 6 & 8 \end{bmatrix}^T, \partial\ell_4(\tilde{\omega}^{(1)}) = \begin{bmatrix} 1 & 9 & 9 \end{bmatrix}^T$$

Since:

$$\tilde{\omega}^{(k+1)} = \tilde{\omega}^{(k)} + (\rho(-\partial L(\tilde{\omega})))$$

So, we have:

$$\tilde{\omega}^{(2)} = \tilde{\omega}^{(1)} + \rho(-\partial L(\tilde{\omega}))$$
$$= 0 - \rho \begin{bmatrix} 0 \\ 2\omega^{(1)} \end{bmatrix} - \rho C \sum_{i=1}^{N} \partial\ell_i(\tilde{\omega})$$
$$= 0 - 0.1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \times 10 \times \left(\begin{bmatrix} -1 \\ -4 \\ -1 \end{bmatrix} + \begin{bmatrix} -1 \\ -3 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ 6 \\ 8 \end{bmatrix} + \begin{bmatrix} 1 \\ 9 \\ 9 \end{bmatrix} \right)$$
$$= \begin{bmatrix} 0 & -8 & -14 \end{bmatrix}^T$$

Similarly, now we have updated $\tilde{\omega}^{(2)}$, then we calculate the Hinge loss again:

By using $\ell_i(\tilde{\omega}^{(2)}) = \max(0, 1 - z_i(\omega^{(2)} x_i + \omega_0))$, we have:

$$\ell_1(\tilde{\omega}^{(2)}) = \max(0,1-(\omega^{(2)}x_1+\omega_0))=\max(0,1-\begin{bmatrix}-8 & -14\end{bmatrix}\begin{bmatrix}4 & 1\end{bmatrix}^T+0)=47>0$$

$$\ell_2(\tilde{\omega}^{(2)}) = \max(0,1-(\omega^{(2)}x_2+\omega_0))=\max(0,1-\begin{bmatrix}-8 & -14\end{bmatrix}\begin{bmatrix}3 & 2\end{bmatrix}^T+0)=53>0$$

$$\ell_3(\tilde{\omega}^{(2)}) = \max(0,1-(\omega^{(2)}x_3+\omega_0))=\max(0,1+\begin{bmatrix}-8 & -14\end{bmatrix}\begin{bmatrix}6 & 8\end{bmatrix}^T+0)=0$$

$$\ell_4(\tilde{\omega}^{(2)}) = \max(0,1-(\omega^{(2)}x_4+\omega_0))=\max(0,1+\begin{bmatrix}-8 & -14\end{bmatrix}\begin{bmatrix}9 & 9\end{bmatrix}^T+0)=0$$

And $\partial\ell_i(\tilde{\omega}^{(2)})$:

$$\partial\ell_1(\tilde{\omega}^{(2)}) = -z_1\begin{bmatrix}1 & x_1^T\end{bmatrix}^T = -\begin{bmatrix}1 & 4 & 1\end{bmatrix}^T$$

$$\partial\ell_2(\tilde{\omega}^{(2)}) = -z_2\begin{bmatrix}1 & x_2^T\end{bmatrix}^T = -\begin{bmatrix}1 & 3 & 2\end{bmatrix}^T,\ \partial\ell_3(\tilde{\omega}^{(2)}) = \partial\ell_4(\tilde{\omega}^{(2)}) = 0$$

Hence,

$$\tilde{\omega}^{(3)} = \tilde{\omega}^{(2)} + \rho(-\partial L(\tilde{\omega}))$$

$$= \begin{bmatrix}0\\-8\\-14\end{bmatrix} - \rho\begin{bmatrix}0\\2\omega^{(2)}\end{bmatrix} - \rho C\sum_{i=1}^{N}\partial\ell_i(\tilde{\omega})$$

$$= \begin{bmatrix}0\\-8\\-14\end{bmatrix} - 0.1\begin{bmatrix}0\\-16\\-28\end{bmatrix} - 0.1\times 10\times\left(\begin{bmatrix}-1\\-4\\-1\end{bmatrix} + \begin{bmatrix}-1\\-3\\-2\end{bmatrix} + \begin{bmatrix}0\\0\\0\end{bmatrix} + \begin{bmatrix}0\\0\\0\end{bmatrix}\right)$$

$$= \begin{bmatrix}0\\-8\\-14\end{bmatrix} + \begin{bmatrix}0\\1.6\\2.8\end{bmatrix} + \begin{bmatrix}1\\4\\1\end{bmatrix} + \begin{bmatrix}1\\3\\2\end{bmatrix}$$

$$= \begin{bmatrix}2 & 0.6 & -8.2\end{bmatrix}^T$$

Question 2:

a)

i) Given the likelihood functions as:

$$p(x\,|\,\theta) = \frac{4}{3\sqrt{\pi}}\theta^{5/2}x^4e^{(-\theta x^2)},\ x \geq 0$$

and independent samples as :{ x_1, x_2, x_3, x_4, }= {2,5,7,11}

Take the logarithm on the likelihood we have:

$$\log(p(x\,|\,\theta)) = \log(\frac{4}{3\sqrt{\pi}}) + \log(\theta^{5/2}) + \log(x^4) - \theta x^2$$

Then, we have:

$$p(x_1, x_2, x_3, x_4 \mid \theta) = \prod_{n=1}^{4} p(x_n \mid \theta)$$

$$\Rightarrow \log(p(x_1, x_2, x_3, x_4 \mid \theta)) = \sum_{n=1}^{4} \log(p(x_n \mid \theta))$$

$$\Rightarrow \log(p(x_1, x_2, x_3, x_4 \mid \theta)) = \sum_{n=1}^{4} \log(\frac{4}{3\sqrt{\pi}}) + \log(\theta^{5/2}) + \log(x_n^4) - \theta x_n^2$$

Then, minimize the negative log likelihood:

$$\min. -(\sum_{n=1}^{4} \log(\frac{4}{3\sqrt{\pi}}) + \frac{5}{2}\log(\theta) + \log(x_n^4) - \theta x_n^2)$$

Drop the constant, then the above function equals:

$$\min. -(\sum_{n=1}^{4} \frac{5}{2}\log(\theta) - \theta x_n^2)$$

Taking derivative on θ and set it to zero:

$$-\frac{\partial \log(p(x_1, x_2, x_3, x_4 \mid \theta))}{\partial \theta}$$

$$= -\frac{\partial(\sum_{n=1}^{4} \frac{5}{2}\log(\theta) - \theta x_n^2)}{\partial \theta}$$

$$= -\sum_{n=1}^{4}(\frac{5}{2\theta} - x_n^2) = 0$$

We get: $\theta = \frac{10}{199}$.

However, wo don't know whether it is the maximum value or minimum value of the negative likelihood. So, we must make a proof. Since we have:

$$-\frac{\partial \log(p(x_1, x_2, x_3, x_4 \mid \theta))}{\partial \theta}$$

$$= \sum_{n=1}^{4} x_n^2 - \frac{10}{\theta}$$

$$= 199 - \frac{10}{\theta}$$

Hence, when $\theta < \frac{10}{199}$, we have

$$\frac{\partial \log(p(x_1, x_2, x_3, x_4 \mid \theta))}{\partial \theta} < 0,$$

when $\theta > \frac{10}{199}$, we have $-\frac{\partial \log(p(x_1, x_2, x_3, x_4 \mid \theta))}{\partial \theta} > 0$.

Therefore, $\theta = \frac{10}{199}$ is the minimum of the negative log likelihood.

ii) From the equality as below:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}$$

$$= \frac{p(x_1, x_2, x_3, x_4 \mid \theta)p(\theta)}{\int_{-\infty}^{+\infty} p(x_1, x_2, x_3, x|\theta)p(\theta)d\theta}$$

$$= \frac{\prod_{n=1}^{4}(\frac{4}{3\sqrt{\pi}}\theta^{5/2}x_n^4 e^{(-\theta x_n^2)}) \times \frac{1}{2}[\delta(\theta-2)+\delta(\theta-3)]}{\int_{-\infty}^{+\infty}\prod_{n=1}^{4}(\frac{4}{3\sqrt{\pi}}\theta^{5/2}x_n^4 e^{(-\theta x_n^2)}) \times \frac{1}{2}[\delta(\theta-2)+\delta(\theta-3)]d\theta}$$

$$= \frac{\prod_{n=1}^{4}(\theta^{5/2}x_n^4 e^{(-\theta x_n^2)}) \times [\delta(\theta-2)+\delta(\theta-3)]}{\left[\prod_{n=1}^{4}(2^{5/2}x_n^4 e^{(-2x_n^2)})\right] + \left[\prod_{n=1}^{4}(3^{5/2}x_n^4 e^{(-3x_n^2)})\right]}$$

Since the given data samples are:

$\{x_1, x_2, x_3, x_4\} = \{2,5,7,11\}$ Hence:

$$p(\theta \mid X) = \frac{\prod_{n=1}^{4}(\theta^{5/2}x_n^4 e^{(-\theta x_n^2)}) \times [\delta(\theta-2)+\delta(\theta-3)]}{\left[\prod_{n=1}^{4}(2^{5/2}x_n^4 e^{(-2x_n^2)})\right] + \left[\prod_{n=1}^{4}(3^{5/2}x_n^4 e^{(-3x_n^2)})\right]}$$

$$= \frac{\prod_{n=1}^{4}(\theta^{5/2}e^{(-\theta x_n^2)}) \times [\delta(\theta-2)+\delta(\theta-3)]}{\left[\prod_{n=1}^{4}(2^{5/2}e^{(-2x_n^2)})\right] + \left[\prod_{n=1}^{4}(3^{5/2}e^{(-3x_n^2)})\right]}$$

$$= \frac{(\theta^{10}e^{-199\theta}) \times [\delta(\theta-2)+\delta(\theta-3)]}{\left[2^{10} \times e^{-2\times 199}\right] + \left[3^{10} \cdot e^{-3\times 199}\right]}$$

Since the impulse of priori, there exists value only when $\theta = 2$ or 3. Thus,

$$p(\theta \mid X) = \frac{(\theta^{10}e^{-199\theta}) \times [\delta(\theta-2)+\delta(\theta-3)]}{\left[2^{10} \times e^{-2\times 199}\right] + \left[3^{10} \cdot e^{-3\times 199}\right]}$$

$$= \begin{cases} \dfrac{2^{10}e^{-2\times 199}\delta(0)}{\left[2^{10} \times e^{-2\times 199}\right] + \left[3^{10} \cdot e^{-3\times 199}\right]} & for\ \theta = 2 \\ \dfrac{3^{10}e^{-3\times 199}\delta(0)}{\left[2^{10} \times e^{-2\times 199}\right] + \left[3^{10} \cdot e^{-3\times 199}\right]} & for\ \theta = 3 \\ 0 & otherwise \end{cases}$$

iii) Then we find the MAP estimate of θ.

MAP of θ:

$$\tilde{\theta} = \arg \max p(\theta \mid X)$$

depending on the results from ii), we have

$$p(2 \mid X) > p(3 \mid X) \quad , \text{So} \quad \tilde{\theta} = 2$$

b)

$$X = \{1,2,2,4,5,7,8,9,9\}, \; N = 9$$

i) This is a one-dimension dataset.

Refer to the $p(x)$ function, we know that:

$$p_N(x) \cong \frac{k_N}{NV_N} = \frac{1}{Nh_N^d} \sum_{n=1}^{N} K\left(\frac{(x - x^{(n)})}{h_N}\right)$$

By substituting $N=9$, $d=1$ and the rectangular window, we have:

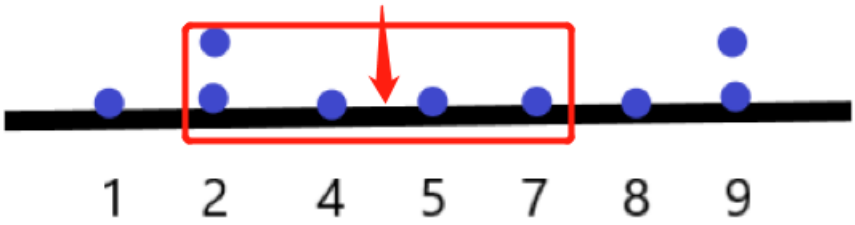
$$p(4.5) = \frac{1}{9 \times 2} \sum_{n=1}^{9} K\left(\frac{(4.5 - x^{(n)})}{2}\right) = \frac{1}{9}$$

ii) Small width leads to spiky pdf, and large width leads to over-smooth pdf, masked the structure of real pdf. When distribution is of **Gaussian**, the Optimal Bandwidth is :

$$h_N = 1.06\sigma N^{-1/5}$$

If the X isn't of normal distribution, then we can use ***cross-validation*** to find unknow bandwidth, by maximizing the average log likelihood of testing samples.

iii) By using *KNN* method, with $k_N = 3$, we have:



We may clearly see that the volume extends to 5, and there are 5 points counted in the below function:

$$p(x = 4.5) = \frac{k_N}{NV_N}, \text{ then by using:}$$

$$V_N^{(2k+1)} = \frac{2k!(4\pi)^k}{(2k+1)!} R^{2k+1} \text{ with } k = 0, \text{ we have } V_N^1 = 2R$$

Hence, calculate out the answer:

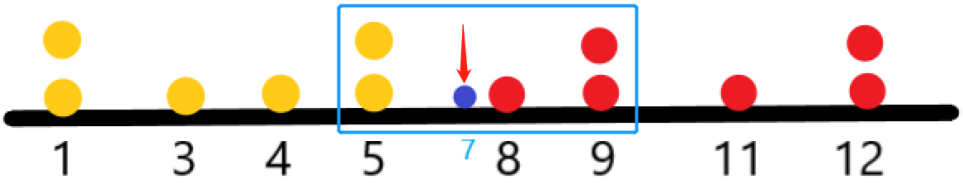
$$p(x=4.5)=\frac{3}{9\times 2\times(7-4.5)}=\frac{1}{15}$$

So, this answer is less accurate since it contains much more points than expected.

iv) In this case, we have:

$$X_1=\{1,1,3,4,5,5\}\ and\ X_2=\{8,9,9,11,12,12\}$$

By using *KNN* method and x=7, we can plot the points:



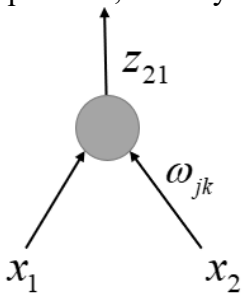
The number containing samples of the volume is 5, bigger than the expected 3. Number of Class 2 : **3** > Number of Class 1: **2**. Hence, choose **Class 2**.

v) In iv), we predict the unlabeled point using *KNN* with an odd k_N, recognize it by simply choosing the class which have larger number in the volume. However, if we use even k_N, we may meet the situation that we have same number of both class points in the certain volume, thus we cannot make decisive classification.

Question3

a)

i) For this non-linear separable problem, we may need **1** unit at least.



In this case, we use logistic function : $f(z)=\frac{1}{1+e^{-z}}$ as our activator.

ii) Denote that $g(x_n)=\omega^T x_n, z_n=f(\omega,x_n)=f(g(x_n)), f(z)=\frac{1}{1+e^{-z}}$

$$R(t,x)=\frac{1}{N}\sum_{n=1}^{N}\lambda(z_n,t_n)$$

$$=\frac{1}{10}\sum_{n=1}^{10}\frac{1}{2}\|t_n-f(g(x_n))\|_2^2$$

$$=\frac{1}{10}\sum_{n=1}^{10}\frac{1}{2}\left\|t_n-\frac{1}{1+e^{-\omega^T x_n}}\right\|_2^2$$

iii) Dimension: **2**, because the number of features is 2.

iv) Firstly, denote: $g_n = w_{jk}^T z_{p,j,n}$, $prediction = f_k(g_n) = f_k(u)$. Since:

$$\frac{\partial R}{\partial w_{jk}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \lambda_n}{\partial g_n}\cdot\frac{\partial g_n}{\partial w_{jk}}$$

We first have:

$$\begin{aligned}\frac{\partial \lambda_n}{\partial g_n} &= \frac{\partial}{\partial g_n}\left(\frac{1}{2}\left\| prediction - t_n \right\|_2^2\right)\\ &= \frac{\partial}{\partial g_n}\left(\frac{1}{2}\left\| f_k(u) - t_n \right\|_2^2\right)\\ &= (f_k(u) - t_n)\, f_k'(u)\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial R}{\partial w_{jk}} &= \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \lambda_n}{\partial g_n}\cdot\frac{\partial g_n}{\partial w_{jk}}\\ &= \frac{1}{N}\sum_{n=1}^{N}(f_k(u) - t_n)\cdot f_k'(u)\cdot\frac{\partial g_n}{\partial w_{jk}}\\ &= \frac{1}{N}\sum_{n=1}^{N}(f_k(u) - t_n)\cdot f_k'(u)\cdot z_{p,j,n}\end{aligned}$$

v)

$$\min_{\omega,\omega_0}\left\|\omega\right\|_2^2 \ \ s.t.\ \ z_i(\omega^T x_n + \omega_0) \geq 1$$

Where, z_i is the label, can be either 1 or -1. ω is the weights and ω_0 the adjustment.

vii) Firstly, we have inequality as below:

$$\min_x . f(\tilde{\omega}) \text{ subject to } b_l \leq c(\tilde{\omega}) \leq b_u,\ \omega = [\omega_1, \omega_2, \omega_3, ..., \omega_I]^T$$

If is a quadratic programming solver, then

$$f(\tilde{\omega}) = \tilde{\omega}^T Q \tilde{\omega},\ c(\tilde{\omega}) = G\tilde{\omega}$$

Where Q, q, G are input matrices/vectors as appropriate dimensions.

viii) If the above function to be used to solve the hard-margin SVM, then:

$$b_l \leq G\tilde{\omega} \leq b_u$$

Meanwhile the primal SVM function can be:

$$\min_{\omega,\omega_0}\left\|\omega\right\|_2^2 \ \ s.t.\ \ z_i(\omega^T x_n + \omega_0) \geq 1$$

Comparing the coefficient, we have:

$$\left\|\omega\right\|_2^2 = \left\|\omega\right\|_2^2 + 0\cdot\omega_0^2 = \tilde{\omega}^T\begin{bmatrix}0 & 0^T\\ 0 & I_{M-1}\end{bmatrix}\tilde{\omega}, thus\ \ Q = \begin{bmatrix}0 & 0^T\\ 0 & I_{M-1}\end{bmatrix}$$

$z_i(\omega^T x_n + \omega_0) \geq 1$ can be rewritten as:

$$(\omega^T x_n + \omega_0) \le -1 \; if \; z = -1, \; and \; (\omega^T x_n + \omega_0) \ge 1 \; if \; z = 1$$

Compared with $b_l \le G\tilde{\omega} \le b_u$,we have:

$$G = \begin{bmatrix} 1 & ... & -1 & ... & -1 \\ x_1^T & ... & -x_{N_1+1}^T & ... & -x_{N_1+N_2}^T \end{bmatrix}^T \; and \; b_u = \begin{bmatrix} -1 & -1 & ... & -1 \end{bmatrix}^T, b_l = -\infty$$

Where
N_1 : No. of samples in Class 1
N_2 : No. of samples in Class 2
$I : N_1 + N_2$

Finally, the conclusion is :

$$\min_{\omega,\omega_0} \tilde{\omega}^T \begin{bmatrix} 0 & 0^T \\ 0 & I_{M-1} \end{bmatrix} \tilde{\omega} \qquad s.t. \; -\infty \le G\tilde{\omega} \le -1$$

where
N_1 : No. of samples in Class 1
N_2 : No. of samples in Class 2, $I : N_1 + N_2$

$$G = \begin{bmatrix} 1 & ... & -1 & ... & -1 \\ x_1^T & ... & -x_{N_1+1}^T & ... & -x_{N_1+N_2}^T \end{bmatrix}^T, \tilde{\omega} = [\omega_0, \omega^T]^T, \omega = [\omega_1, \omega_2, \omega_3, ..., \omega_I]^T$$

b)

i) First, we have:

$$X_1 = \left[[2,4]^T, [3,6]^T, [5,8]^T, [6,6]^T, [8,10]^T \right]^T$$
$$X_2 = \left[[7,4]^T, [8,5]^T, [9,7]^T, [10,6]^T, [11,10]^T \right]^T$$

Since $X = [X_1^T, X_2^T]^T$, we have $\overline{X}^T = [\overline{X}_1^T, \overline{X}_2^T]$,

we can calculate the covariance matrix:

$$C_{xx} = \frac{1}{N-1}(\overline{X}^T \overline{X})$$
$$= \frac{1}{9}\begin{bmatrix} 769/10 & 138/5 \\ 138/5 & 212/5 \end{bmatrix}$$
$$\approx \begin{bmatrix} 8.54 & 3.07 \\ 3.07 & 4.71 \end{bmatrix}$$

Then we can obtain the eigenvalues:

$$P_A(\lambda) = \lambda^2 - tr(C_{xx})\lambda + \det(C_{xx})$$

where: $tr(C_{xx}) = 8.54 + 4.71 = 13.25, \det(C_{xx} = 8.54 \times 4.71 - 3.07 \times 3.07) \approx 30.8$.

Hence, $P_A(\lambda) = \lambda^2 - 13.25\lambda + 30.8$

Then the eigenvalues are $\lambda_1 = 10.245, \lambda_2 = 3.005$.

The eigenvectors: $u_1 = [3.07, 1.705]^T$, $u_2 = [3.07, -5.535]^T$

ii) Firstly, we find the class statistics:

$$\text{For } X: \bar{x}_1 = [4.8, 6.8]^T,\ \bar{x}_2 = [9, 6.4]^T.\ S_w = \sum_{k=1}^{2} (x_k - \bar{x}_k)(x_k - \bar{x}_k)^T$$

$$S_1 = \begin{bmatrix} 22.8 & 18.8 \\ 18.8 & 20.8 \end{bmatrix}, S_2 = \begin{bmatrix} 10 & 13 \\ 13 & 21.2 \end{bmatrix} \text{ Hence,}$$

$$S_\omega = S_1 + S_2 = \begin{bmatrix} 22.8 & 18.8 \\ 18.8 & 20.8 \end{bmatrix} + \begin{bmatrix} 10 & 13 \\ 13 & 21.2 \end{bmatrix} = \begin{bmatrix} 32.8 & 31.8 \\ 31.8 & 42 \end{bmatrix}$$

Then we should calculate ω:

$$\begin{aligned} \omega &= S_\omega^{-1}(\bar{x}_1 - \bar{x}_2) \\ &= \begin{bmatrix} 32.8 & 31.8 \\ 31.8 & 42 \end{bmatrix}^{-1} \begin{bmatrix} -4.2 \\ 0.4 \end{bmatrix} \\ &= \begin{bmatrix} -0.52 \\ 0.4 \end{bmatrix} \end{aligned}$$

After normalization :

$$Normalized(\omega) := \omega / \|\omega\|^2 = \begin{bmatrix} -0.79 \\ 0.61 \end{bmatrix}$$

ii) **Merit for PCA:**

Can make dimension reduction to reduce computational load

There is no ill-conditioning for PCA method.

Cons for PCA:

PCA method does not consider class information/label.

The eigenvectors may not separate the classes well.

Merit for FLD:

Consider the class information and it is to some extent a remedy for PCA.

Cons for FLD:

Computing process may suffer from ill-conditioning.

Projection may lead to performance loss by information loss.

iv) Firstly, the assignment step. Give No. of Classes and initial centroids of the samples(**3** here). Then, assign each sample to the nearest centroid, by calculating the distance between sample and the temporary centroids:

$$\min . \sum_{k=1}^{K} \sum_{x_i \in \omega_k} \|x_i - \mu_k\|_2^2$$

Second step is update step:

After assigning the whole samples to their corresponding classes, we calculate the new centroids using samples with the assigned labels, by:

$$\mu_k = \frac{1}{N_k} \sum_{x_i \in \omega_k} x_i$$

Where x_i are the updated samples belonging to $\boldsymbol{w}_k$, from assignment step, this operation will help us to find the next centroids.

Do the above operations **iteratively** and terminate the algorithm when there is no change in the labels of these samples. At this time, K-Means can be viewed as converged.

v) **Divisive clustering**.
In divisive clustering, it will start with 1 cluster containing all samples, successively split clusters until all clusters contains only 1 sample. In this process, similar clusters are merged and the cluster with largest dis-similarity is split.

vi) In fact, as its name, *Impurity Measure* means the chaos degree of the current state. By using the function:

$$R_i = -\sum p_k \log_2 p_k$$

For all possible value i =1,2, … I of feature A. Then we can know the entropy and use the gain of entropy to be the evidence for choosing appropriate feature to make split. For example, we may choose the feature which is corresponding to maximum information gain:

$$G = R - \sum p_i R_i$$

as the feature to split the samples.

vii) Firstly:

If we do not have a prior knowledge or there are too many features in the data samples, we should take transformation and filtering techniques.

For example, in time series data prediction (Such as energy consumption value prediction) and voice classification of male/female problem, we may transform the data samples into frequency domain by using Fourier Transformation or Wavelet Transformation.

For Fourier Transformation, we may regard it as a kind of **decomposition of features**, projecting the value to frequency domain makes us easier to make filtering or de-noising. What's more, adopting the useful frequencies as the features can help us separate the data.

On the other hand, filtering the data can help to compress the data by removing the irrelevant data by using Low Pass, High Pass or Band Pass filtering, which

may reduce the redundancy and enhance the data quality as well.

Secondly:

2-D Fourier transform is global transformation with computational expensive. For **HARR** transform, it is easier to be implemented and lower complexity, thus it needs less computational resource.

----------END OF ANSWER SHEET----------