DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2015/ 2016)

COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

1. (a) The maximum value of M is $K-1$ because the rank of $\mathbf{S}_B$ is at most $K-1$.
(5 marks, K)

(b) Consider the term $\mathbf{z}_n = \mathbf{x}_n - \boldsymbol{\mu} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}) + (\mathbf{x}_n - \boldsymbol{\mu}_k)$. Then, we have

$$\begin{aligned}
\mathbf{z}_n\mathbf{z}_n^{\mathsf{T}} &= [(\boldsymbol{\mu}_k - \boldsymbol{\mu}) + (\mathbf{x}_n - \boldsymbol{\mu}_k)]\,[(\boldsymbol{\mu}_k - \boldsymbol{\mu}) + (\mathbf{x}_n - \boldsymbol{\mu}_k)]^{\mathsf{T}} \\
&= (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}} + (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}} + \\
&\quad (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}} + (\mathbf{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbf{S}_T &= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k} \mathbf{z}_n\mathbf{z}_n^{\mathsf{T}} \\
&= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}} + \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}} \\
&\quad + \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}} + \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}} \\
&= \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}} + \mathbf{S}_W + \sum_{k=1}^{K}(\boldsymbol{\mu}_k - \boldsymbol{\mu})\sum_{n\in\mathcal{C}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}} \\
&\quad + \sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^{\mathsf{T}} + \cdots + \sum_{n\in\mathcal{C}_K}(\mathbf{x}_n - \boldsymbol{\mu}_K)(\boldsymbol{\mu}_K - \boldsymbol{\mu})^{\mathsf{T}} \\
&= \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}} + \mathbf{S}_W + \sum_{k=1}^{K}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(N_k\boldsymbol{\mu}_k - N_k\boldsymbol{\mu}_k)^{\mathsf{T}} \\
&\quad + (N_1\boldsymbol{\mu}_1 - N_1\boldsymbol{\mu}_1)(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^{\mathsf{T}} + \cdots + (N_K\boldsymbol{\mu}_K - N_K\boldsymbol{\mu}_K)(\boldsymbol{\mu}_K - \boldsymbol{\mu})^{\mathsf{T}} \\
&= \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathsf{T}} + \mathbf{S}_W.
\end{aligned}$$

Because $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$, the first term in the equation above must be $\mathbf{S}_B$.
(10 marks, E)

(c) (**Option 1**) The Lagrangian function can be written as

$$\begin{aligned}
L(\mathbf{W}, \{\lambda_j\}) &= \mathrm{Tr}\{\mathbf{W}^{\mathsf{T}}\mathbf{S}_B\mathbf{W}\} - \sum_{j=1}^{M}\lambda_j(\mathbf{w}_j^{\mathsf{T}}\mathbf{S}_W\mathbf{w}_j - 1) \\
&= \mathrm{Tr}\{\mathbf{W}^{\mathsf{T}}\mathbf{S}_B\mathbf{W}\} - \mathrm{Tr}\{\mathbf{W}^{\mathsf{T}}\mathbf{S}_W\mathbf{W}\boldsymbol{\Lambda}_M - \boldsymbol{\Lambda}_M\}
\end{aligned}$$

where $\mathbf{\Lambda}_M = \text{diag}\{\lambda_1, \ldots, \lambda_M\}$ comprises M Lagrange multipliers. Setting $\frac{\partial L}{\partial \mathbf{W}} = \mathbf{0}$ and using the property of matrix trace derivative, we obtain

$$\begin{aligned} &\mathbf{S}_B\mathbf{W} - \mathbf{S}_W\mathbf{W}\mathbf{\Lambda}_M = \mathbf{0} \\ &\Longrightarrow \mathbf{S}_B\mathbf{W} = \mathbf{S}_W\mathbf{W}\mathbf{\Lambda}_M \\ &\Longrightarrow (\mathbf{S}_W^{-1}\mathbf{S}_B)\mathbf{W} = \mathbf{W}\mathbf{\Lambda}_M \end{aligned}$$

Therefore, $\mathbf{W}$ comprises the first M eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ and $\mathbf{\Lambda}_M$ comprises M eigenvalues in its diagonal elements.

(**Option 2**) To find $\mathbf{w}_j$, we write the Lagrangian function as:

$$L(\mathbf{w}_j, \lambda_j) = \mathbf{w}_j^{\mathsf{T}}\mathbf{S}_B\mathbf{w}_j - \lambda_j(\mathbf{w}_j^{\mathsf{T}}\mathbf{S}_W\mathbf{w}_j - 1)$$

Setting $\frac{\partial L}{\partial \mathbf{w}} = 0$, we obtain

$$\begin{aligned} &\mathbf{S}_B\mathbf{w} - \lambda\mathbf{S}_W\mathbf{w} = 0 \\ &\Longrightarrow \mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w} \\ &\Longrightarrow (\mathbf{S}_W^{-1}\mathbf{S}_B)\mathbf{w} = \lambda\mathbf{w} \end{aligned} \tag{1}$$

Therefore, the optimal solution of $\mathbf{w}_j$ satisfies

$$(\mathbf{S}_W^{-1}\mathbf{S}_B)\mathbf{w}_j = \lambda_j\mathbf{w}_j$$

Therefore, $\mathbf{W}$ comprises the first M eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

(10 marks, A)

COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

2. (a) (i) Consider input vectors $\mathbf{x} = [x_1 \;\; x_2]^\mathsf{T}$ and $\mathbf{y} = [y_1 \;\; y_2]^\mathsf{T}$. Then, we have

$$
\begin{aligned}
K(\mathbf{x},\mathbf{y}) &= (1+\mathbf{x}^\mathsf{T}\mathbf{y})^2 \\
&= \left(1 + [x_1 \;\; x_2]\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right)^2 \\
&= (1 + x_1y_1 + x_2y_2)(1 + x_1y_1 + x_2y_2) \\
&= 1 + 2x_1y_1 + 2x_2y_2 + 2x_1y_1x_2y_2 + x_1^2y_1^2 + x_2^2y_2^2 \\
&= \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1x_2 & x_1^2 & x_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \\ \sqrt{2}y_1y_2 \\ y_1^2 \\ y_2^2 \end{bmatrix} \\
&= \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{y}).
\end{aligned}
$$

Therefore, vector $\mathbf{x}$ is mapped to $\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1x_2 & x_1^2 & x_2^2 \end{bmatrix}^\mathsf{T}$, which is a 6-dimensional vector. The decision boundary becomes linear because the output of the SVM can now be written as:

$$f(\mathbf{x}) = \sum_{i\in\mathcal{S}} a_i \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_i) + b,$$

which is linearly related to $\phi(\mathbf{x})$. (10 marks)

(ii) Because of the nonlinear cross-product terms x_iy_i, $x_1y_1x_2y_2$, and $x_i^2y_i^2$ in $K(\mathbf{x},\mathbf{y})$, $f(\mathbf{x})$ is nonlinear function of $\mathbf{x}$.

(3 marks)

(b) Define $a_1^{(L)} = \sum_j w_j^{(L)} o_j^{(L-1)}$ as the activation of the output neuron.[1] Then, the instantaneous error gradident can be written as

$$\frac{\partial E}{\partial w_j^{(L)}} = \frac{\partial E}{\partial a_1^{(L)}} \frac{\partial a_1^{(L)}}{\partial w_j^{(L)}} = \delta_1^{(L)} o_j^{(L-1)}$$

[1] $w_j^{(L)}$ also includes the bias term

where

$$
\begin{aligned}
\delta_1^{(L)} &= \frac{\partial E}{\partial a_1^{(L)}} = \frac{\partial E}{\partial o_1^{(L)}}\frac{\partial o_1^{(L)}}{\partial a_1^{(L)}} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial a_1^{(L)}} \\
&= \left[-\frac{t}{y} + \frac{1-t}{1-y}\right] h'(a_1^{(L)}) \qquad\qquad h(z) = \frac{1}{1+e^{-z}} \\
&= \frac{-t(1-y)+(1-t)y}{y(1-y)} h(a_1^{(L)}) \left[1 - h(a_1^{(L)})\right] \\
&= y - t \qquad\qquad \because y = h(a_1^{(L)})
\end{aligned}
$$

$$
\Rightarrow \frac{\partial E}{\partial w_j^{(L)}} = (y-t)o_j^{(L-1)}
$$

(6 marks, E)

(c) Consider the posterior density:

$$
\begin{aligned}
&p(\mathbf{z}_i|\mathbf{x}_i,\boldsymbol{\omega}) \\
&\propto p(\mathbf{x}_i|\mathbf{z}_i,\boldsymbol{\omega})p(\mathbf{z}_i) \\
&= \mathcal{N}(\mathbf{x}_i|\mathbf{m}+\mathbf{V}\mathbf{z}_i,\boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I}) \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mathbf{m}-\mathbf{V}\mathbf{z}_i)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m}-\mathbf{V}\mathbf{z}_i) - \frac{1}{2}\mathbf{z}_i^{\mathsf{T}}\mathbf{z}_i\right\} \\
&= \exp\left\{\mathbf{z}_i^{\mathsf{T}}\mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m}) - \frac{1}{2}\mathbf{z}_i^{\mathsf{T}}\left(\mathbf{I}+\mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V}\right)\mathbf{z}_i\right\}.
\end{aligned}
$$

Comparing the terms of this equation with

$$
\begin{aligned}
\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z,\mathbf{C}_z) &\propto \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_z)^{\mathsf{T}}\mathbf{C}_z^{-1}(\mathbf{z}-\boldsymbol{\mu}_z)\right\} \\
&\propto \exp\left\{\mathbf{z}^{\mathsf{T}}\mathbf{C}_z^{-1}\boldsymbol{\mu}_z - \frac{1}{2}\mathbf{z}^{\mathsf{T}}\mathbf{C}_z^{-1}\mathbf{z}\right\},
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathbf{C}_z^{-1} &= \mathbf{I}+\mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V} = \mathbf{L} \\
\mathbf{C}_z^{-1}\boldsymbol{\mu}_z &= \mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m}) \\
\Longrightarrow \boldsymbol{\mu}_z &= \mathbf{C}_z\mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m}) \\
&= \mathbf{L}^{-1}\mathbf{V}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m}) \\
&= \langle\mathbf{z}_i|\mathbf{x}_i\rangle
\end{aligned}
$$

Using the definition of covariance matrix: $\text{cov}(\mathbf{z},\mathbf{z}) = \langle \mathbf{x}\mathbf{x}^{\mathsf{T}}\rangle - \langle \mathbf{x}\rangle\langle \mathbf{x}^{\mathsf{T}}\rangle$ and noting that $\mathbf{L}$ is the posterior precision matrix, we have

$$\langle \mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathbf{x}_i\rangle = \mathbf{L}^{-1} + \langle \mathbf{z}_i|\mathcal{X}\rangle\langle \mathbf{z}_i^{\mathsf{T}}|\mathbf{x}_i\rangle.$$

(6 marks, A)