DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2016/ 2017)

COURSE: EIE4105 YEAR: 4
SUBJECT: Multimodal Human Computer Interaction Technology

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

1. (a) The minimum value of N is $(260)(360)/100 + 1 = 93600/100 + 1 = 937$.

(5 marks, A)

(b) The minimum value of N is $937 + 1 = 938$. This is because $\mathbf{x}_{k,2}$ depends on $\mathbf{x}_{k,1}$, meaning that $\mathbf{x}_{k,2}$ is redundant if we already have $\mathbf{x}_{k,1}$. As a result, we need one extra sample for each person to make sure that $\text{rank}(\boldsymbol{\Sigma}) = 93600$.

(5 marks, E)

(c) We use $\mathcal{X}_k$ to estimate the mean vector $\boldsymbol{\mu}_k$ of person k, where $k = 1, \ldots, 100$. Specifically, we compute

$$\boldsymbol{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{k,i}.$$

Then, a Gaussian classifier can be constructed by using the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\}_{k=1}^{100}$.

(5 marks, K)

(d) The decision boundaries are linear because the covariance matrices of all classes (Gaussian models) are the same.

(5 marks, AE)

[The following equations are optional] Assuming equal priors, for Persons r and s, the discriminant function given by the Gaussian classifier is

$$\begin{aligned} & \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_r, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}) \\ \Rightarrow \quad & \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_r)\right\} = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_s)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_s)\right\} \end{aligned}$$

Taking logarithm on both side, we have

$$\begin{aligned} & (\mathbf{x}-\boldsymbol{\mu}_r)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_r) = (\mathbf{x}-\boldsymbol{\mu}_s)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_s) \\ \Rightarrow \quad & 2(\boldsymbol{\mu}_s-\boldsymbol{\mu}_r)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} = \boldsymbol{\mu}_s^{\mathsf{T}}\boldsymbol{\mu}_s - \boldsymbol{\mu}_r^{\mathsf{T}}\boldsymbol{\mu}_r. \end{aligned}$$

The discriminant function is linear in $\mathbf{x}$.

(e) We train a DNN with 93600 inputs and 100 outputs. The outputs should be a softmax function of the outputs of the last hidden layer:

$$P(\text{Person} = k|\mathbf{x}) = y_k(\mathbf{x}) = \frac{\exp(a_k(\mathbf{x}))}{\sum_j \exp(a_j(\mathbf{x}))}$$

or

$$P(\text{Person} = k|\mathbf{x}) = y_k(\mathbf{x}) = \frac{\exp(o_k(\mathbf{x}))}{\sum_j \exp(o_j(\mathbf{x}))}$$

where $a_k(\mathbf{x})$ and $o_k(\mathbf{x})$ are the activation and output of the k-th node in the last hidden, respectively. Note that the last hidden layer should have 100 nodes.

(5 marks, KA)

COURSE: EIE4105 YEAR: 4
SUBJECT: Multimodal Human Computer Interaction Technology

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

2. (a) We compute the covariance matrix using all training data in $\mathcal{X}$. Then, we compute the eigenvectors and eigenvalues of the covariance matrix. Then, we select the top-5 eigenvectors whose eigenvalues are the largest. Then, we project the vectors using

$$\mathbf{y}_i^{(g)} = \mathbf{W}^{\mathsf{T}}(\mathbf{x}_i^{(g)} - \boldsymbol{\mu}), \qquad g \in \{m, f\},$$

where $\mathbf{W}$ comprises the top-5 eigenvectors in its columns and $\boldsymbol{\mu}$ is the global mean of $\mathcal{X}$.

(6 marks, K)

(b) Fig. Q2(b) is the eigenface corresponding to the last PC. This is because the variation in the image is very small when compared with that in Fig. Q2(a).

(4 marks, A)

(c) Max. no. of eigenfaces with positive eigenvalues is 4096 because the dimension of $\mathbf{x}_i^{(g)}$ is 4096 and $2N = 10000 > 4096$. This means that we have enough data to compute all of the eigenvectors.

(5 marks, AE)

(d) Max. no. of eigenfaces with positive eigenvalues is 1. This is because there are only 2 classes (genders) and the solution of LDA comprises the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$, where $\mathbf{S}_W$ and $\mathbf{S}_B$ are the within-class and between-class scatter matrices, respectively. For 2-class problems, the rank of $\mathbf{S}_B$ is 1 so that there is only one eigenvector with positive eigenvalue.

(5 marks, AE)

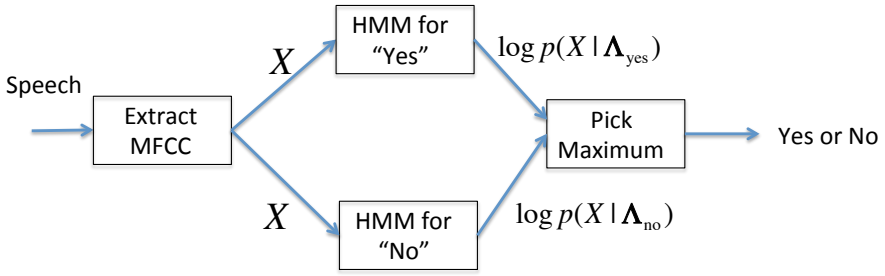
(e) Linear SVM is appropriate because there is only one sample for the male class and the total number of training samples is much less than the feature dimension. Linear SVM provides a strong regularization on the decision boundary to avoid overfitting in such situation.

(5 marks, E)

COURSE: EIE4105 YEAR: 4
SUBJECT: Multimodal Human Computer Interaction Technology

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

3. (a) The block diagram of the Yes-No recognizer is shown below.



(5 marks, K)

(b) 1. Collect many utterances of "Yes" and "No" from many speakers.
2. Extract the acoustic vectors (energy and MFCCs plus their first and second derivative) from the speech regions of these utterances.
3. Use the acoustic vectors from "Yes" to train an HMM to model the spectral and temporal characteristics of "Yes"; similarly for the word "No".

(6 marks, KA)

(c) No. of states for each HMM is 5–10 (must not be less than 3) because each phone may require 3 states to model and each word has 3 phonemes.

(4 marks, A)

(d) We use the Bayes' theorem to compute the posterior probability and determine the class label of X according to the HMM that gives the maximum posterior probability. That is, the predicted label is

$$\begin{aligned} l(\mathcal{X}) &= \underset{i\in\{\text{'Yes'},\text{'No'}\}}{\arg\max} \; P(i|\mathcal{X}) \\ &= \underset{i\in\{\text{'Yes'},\text{'No'}\}}{\arg\max} \; \frac{P(i)p(\mathcal{X}|\Lambda_i)}{0.2p(\mathcal{X}|\Lambda_{\text{yes}}) + 0.8p(\mathcal{X}|\Lambda_{\text{no}})} \\ &= \underset{i\in\{\text{'Yes'},\text{'No'}\}}{\arg\max} \; \frac{P(i)p(\mathcal{X}|\Lambda_i)}{p(\mathcal{X})} \\ &= \underset{i\in\{\text{'Yes'},\text{'No'}\}}{\arg\max} \; P(i)p(\mathcal{X}|\Lambda_i) \end{aligned}$$

where $P(\text{'Yes'}) = 0.8$ and $P(\text{'No'}) = 0.2$.

(6 marks, E)

(e) They are not necessary because this two-word recognition task does not use phone models. Instead, word models are used.

(4 marks, A)

COURSE: EIE4105 YEAR: 4
SUBJECT: Multimodal Human Computer Interaction Technology

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

4. (a) (i) Fig. Q4(b) corresponds to the support vector because it does not look like the digit 0. A support vector is either close to the decision boundary or on the other side of the decision boundary. This means that they are confusable. For Digit '0', the support vectors will not look like a Digit '0'.

(6 marks, AE)

(ii) The max. number of support vectors is 10,000. When $\sigma \to 0$, all training vectors will become support vectors.

(5 marks, E)

(iii) The min. number of SVs per SVM is 2, one from the positive class and one from the negative class.

(3 marks, K)

(b) (i) When the enrollment is very long, $\alpha_j \to 1$ so that the GMM means depend almost on the enrollment utterance. This is reasonable because when there are many acoustic vectors in $\mathcal{X}^{(s)}$, we should believe our observations.

When the utterance is very short, $\alpha_j \to 0$ so that the GMM means are almost the same as the UBMs means. This is reasonable because when there are not many acoustic vectors in $\mathcal{X}^{(s)}$, we better believe the prior, i.e., the UBM means.

(6 marks, KA)

(ii) We cannot directly apply EM to compute $\boldsymbol{\mu}_j^{(s)}$'s because the EM algorithm has no guarantee on the index arrangement in the mixture model. This means that if we apply EM independently on individual speakers when computing their supervectors, the one-to-one correspondence between the subvectors $\boldsymbol{\mu}_j^{(s)}$'s in $\vec{\boldsymbol{\mu}}^{(s)}$ will be lost for different target speakers. This one-to-one correspondence, however, can be guaranteed in MAP adaptation because $\boldsymbol{\mu}_j^{(s)}$'s are computed one by one for each j.

(5 marks, E)