DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017/ 2018)

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

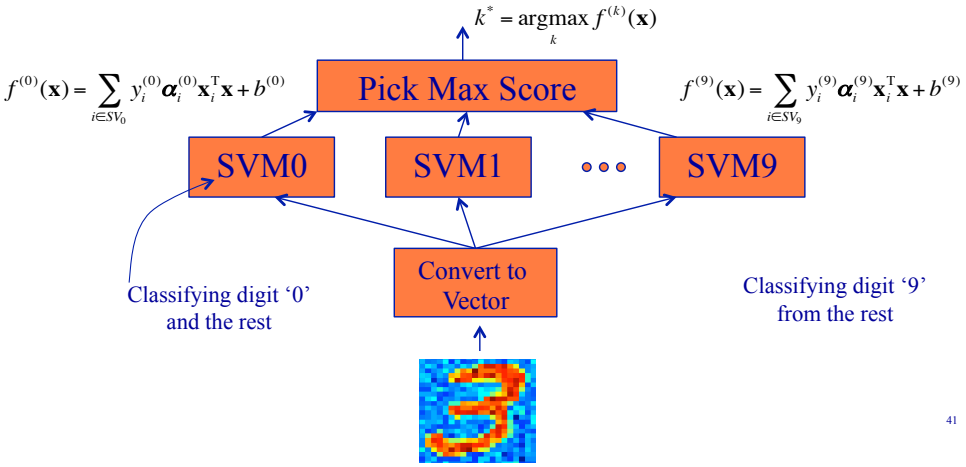
	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

1. (a) Dimension = 16384

(2 marks, K)

(b) The structure of the SVM classifier is as follows (replace Digit by Person):

(5 marks, KA)



(c) Linear kernel is the most appropriate. This is because the dimension of features (16384) is larger than the number of training samples (1000).

(5 marks, A)

(d) (i) Stacking the columns will destroy the local spatial information of the images.
(ii) This method produces feature vectors of very large dimension.

(4 marks, E)

(e) The maximum dimension is 999. Note that the 1000-th eigenvalue is 0.

(4 marks, E)

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

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

2. (a) The minimum value of N can be computed as follows:

$$K(N-1) \geq 200 \times 100$$
$$\Rightarrow N \geq \frac{200 \times 100}{10} + 1 = 2001$$

(4 marks, A)

(b) Substituting the 2nd equation into the first, we obtain

$$\mathbf{x}_{k,5} = 3\mathbf{x}_{k,3} + 2\mathbf{x}_{k,2} - \mathbf{x}_{k,2} \odot \mathbf{x}_{k,1} + 3\mathbf{1}$$
$$\mathbf{x}_{k,4} = \mathbf{x}_{k,3} + \mathbf{x}_{k,2} + \mathbf{1}.$$

This means that $\mathbf{x}_{k,5}$ depends on $\mathbf{x}_{k,1}$, $\mathbf{x}_{k,2}$ and $\mathbf{x}_{k,3}$ and that $\mathbf{x}_{k,4}$ depends on $\mathbf{x}_{k,2}$ and $\mathbf{x}_{k,3}$. Therefore, $\mathbf{x}_{k,4}$ and $\mathbf{x}_{k,5}$ are redundant if we know $\mathbf{x}_{k,1}$, $\mathbf{x}_{k,2}$ and $\mathbf{x}_{k,3}$. As a result, we need **two** extra samples for each person to make sure that $\texttt{rank}(\mathbf{\Sigma}) = 20000$. Therefore, the minimum value of N is $2001 + 2 = 2003$.

An alternative way to find this value is to change the formula in Q1(a) as follows:

$$K(N-2-1) \geq 200 \times 100$$
$$\Rightarrow N \geq \frac{200 \times 100}{10} + 3 = 2003$$

(6 marks, E)

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

(4 marks, KA)

(d) We may constrain the covariance matrices $\{\mathbf{\Sigma}_k\}_{k=1}^K$ to be diagonal. When $\mathbf{\Sigma}_k$'s are diagonal matrices, we only need to make sure that all of the diagonal elements are non-zero. As 2001 samples per person are more than enough to meet this requirement, the covariance matrices will have valid inverse. An alternative approach is to add a small constant to the diagonal elements of the full covariance matrices, i.e., $\mathbf{\Sigma}_k \leftarrow \mathbf{\Sigma}_k + \epsilon\mathbf{I}$. This method can turn a singular matrix into non-singular one.

(6 marks, E)

(e) We compute the posterior probabilities of the 10 classes given an unknown fingerprint $\mathbf{x}$. Then, we identify the person according to the maximum posterior probability. Specifically,

$$l = \arg\max_k P(C_k|\mathbf{x}) = \arg\max_k \frac{P(C_k)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^K P(C_j)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)}$$

(5 marks, A)

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

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

3. (a) (i) No. of inputs = 784
No. of outputs = 10

(2 marks, K)

(ii) The bias terms are to allow the network to produce decision boundaries that lie on any regions of the feature space, not necessarily passing through the origin.

(3 marks, A)

(iii) The softmax function is applied to the network output. This is because it is a classification problem and we would like the network to produce posterior probabilities of individual classes given an unknown digit.
The objective function is cross-entropy because the function is differentiable and is closest to the classification error. It also emphasizes the error on the correct class.

(5 marks, KA)

(iv) If all hidden nodes are linear, the multiple layers reduce to a single layer and the advantages of deep architecture is lost. Mathematically, denote $\mathbf{x}$ as the input and $\mathbf{W}_l$, $l = 1, \ldots, L$ as the weights (including bias terms) of the hidden layers. When all hidden neurons are linear, then the network outputs become

$$\mathbf{y} = \text{softmax}(\mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1 \mathbf{x}) = \text{softmax}(\mathbf{W}\mathbf{x})$$

where $\mathbf{W} = \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1$. This is equivalent to multi-class logistic regression.

(7 marks, E)

(b) (i) No. of weights = $(3 \times 3 + 1) \times 64 = 640$.

(2 marks, K)

(ii) The convolutional layers are used for extracting spatial patterns (features) that are relevant to the classification task.

(2 marks, K)

(iii) The max-pooling layers perform subsampling, which can reduce the number of parameters in the convolutional layers and the fully connected layers.

(2 marks, K)

(iv) The fully connected layers act as a non-linear classifier that classifies the features extracted from the last convolutional layer or last max-pooling layer of the CNN.

(2 marks, K)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017/ 2018)

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

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

4. (a) The probability of transiting from State 1 to State 2 in the HMM is $1 - 0.6 = 0.4$.
(3 marks, K)

(b) The probability will be larger than 0 but smaller than 0.6. This is because the first section of the phoneme ɪ is shorter. As a result, the chance of staying in State 1 becomes smaller.
(6 marks, E)

(c) Condition 1 is true because the MFCCs of iː match the HMM of iː on the left-hand-side of the inequality. On the other hand, there is a mismatch between the MFCCs and HMM on the right-hand-side of the inequality.

Condition 2 is false because the MFCCs match the HMM on the left. Its likelihood should be larger than the one on the right in which the MFCCs do not match the HMM.

Condition 3 is likely to be false. Because speech is stochastic, it is very unlikely that the two likelihoods are identical. (6 marks, A)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017/ 2018)

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

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

5. (a) Because the matrix $\mathbf{T}$ is rectangular (low-rank) with the number of columns smaller than the number of rows, the dimension of $\mathbf{w}_s$ is smaller than that of $\vec{\boldsymbol{\mu}}$.

(4 marks, K)

(b) Because the TV matrix $\mathbf{T}$ is trained by using the speech of many speakers and there is no restriction on what channel conditions to collect the speech. Moreover, when training $\mathbf{T}$, each utterance is considered to be spoken by a different speaker. As a result, the covariance matrix $\mathbf{T}\mathbf{T}^{\mathsf{T}}$ reflects not only speaker variability but also channel variability.

(4 marks, A)

(c) Because $\mathbf{T}$ defines the total variability space, the covariance matrix $\mathbf{T}\mathbf{T}^{\mathsf{T}}$ models the channel variability as well as speaker variability. Therefore, the i-vectors contain not only speaker information but also channel information. The cosine distance scores will therefore be affect by the channel information, causing poor verification performance. The problem can be addressed by applying LDA or LDA+WCCN on the i-vectors before computing the cosine distance scores. These pre-precessing methods suppress within-speaker variability but emphasize between-speaker variability, which has the effect of mitigating the channel effect on cosine-distance scores.

(7 marks, AE)