DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME ~~(2017/2018)~~ 2019/2020

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

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

1. (a) The minimum number of independent samples is 2. This is because diagonal covariance matrices are to be used in the Gaussian classifier. The inverse of a diagonal covariance matrix will exists as long as all of its diagonal elements are non-zero. To meet this requirement, all we need is to have two different samples (vectors) for each class. Also, none of the feature components in the two vectors can be identical; otherwise, the corresponding variance will be zero, causing infinite reciprocal.

(5 marks, AE)

(b) The outcomes of rolling a die are discrete events, i.e., they can only be $1, 2, \ldots, 6$. The probability mass function (PMF) comprises impulses at $1, 2, \ldots, 6$ and zero everywhere. In fact, probabilities are only defined at discrete values of the feature space. On the other hand, a Gaussian density function or a Gaussian mixture density function is for modeling continuous random variables in which the full feature space are defined (although in many regions, the likelihood could be very close to zero). If a Gaussian density or a GMM is used to model the distribution of discrete events, the probability mass between the discrete intervals will be incorrect.

(5 marks, AE)

(c) (i) Ten GMMs are required, one for each digit. Given an unknown image $\mathbf{x}$, we compute the likelihood of $\mathbf{x}$ for each digit using $p(\mathbf{x}|C_k)$, where $k = 0, \ldots, 9$. Then, we assign the digit to $\mathbf{x}$ according to

$$d = \arg\max_{k=0}^{9} P(C_k|\mathbf{x}).$$

(4 marks, AE)

(ii) $P(C_k)$ is the prior probability of digit k, which is the probability of getting digit k without observing the query image $\mathbf{x}$. It plays the role of scaling up or down the likelihoods according to how often a digit will occur in an application. For example, in bank-cheque recognition, the digit '0' will appear more often than the others. Then, its prior probability will also be higher, which scales up the likelihood $p(\mathbf{x}|C_0)$. Therefore, even if $p(\mathbf{x}|C_6)$ is slightly larger than $p(\mathbf{x}|C_0)$, after the scaling, we still recognise the query image as digit '0' after scaling because $P(C_0|\mathbf{x}) > P(C_6|\mathbf{x})$.

(6 marks, AE)

(iii) $p(\mathbf{x})$ is the marginal likelihood of $\mathbf{x}$. Its purpose is to scale the numerator so that the ratio $P(C_k|\mathbf{x}) = \frac{P(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$ is bounded between 0 and 1 and that $P(C_k|\mathbf{x})$ is the posterior probability. Because it is independent of the digits, it is not necessary to compute its value for classification.

(5 marks, AE)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME ~~(2017/2018)~~ 2019/2020

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

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

2. (a) When the feature dimension is larger than the number of training samples, linear SVMs could perfectly separate the two classes. Therefore, using linear SVMs has less chance of overfitting the data. Another situation is that when there is only one or two training samples from one class, using linear SVMs will be more desirable as the chance of overfitting the minority class is smaller.

(5 marks, AE)

(b) (i) These support vectors correspond to the persons whose voice can hardly tell their gender, i.e., their voice is confusing in terms of femininess and masculinity. However, the SVM is still able to identify their gender correctly.

(4 marks, AE)

(ii) These support vectors correspond to the persons whose voice sounds more like their opposite gender than their own gender. These vectors cause the gender identification problem non-linearly separable. The linear SVM needs to use non-zero slack variables ($\xi_i > 1$) to handle these vectors.

(4 marks, AE)

(c) The maximum number of eigenfaces is 999. Because the feature dimension (20,000) is larger than the number of training vectors (1000), the maximum rank of the covariance matrix is $1000 - 1 = 999$. The assumption is that these 1000 images are all different and independent.

(5 marks, AE)

(d) We first convert the images into 784-dimensional vectors by stacking the columns (or rows) of the images. Then, we compute the within-class scatter matrix $\mathbf{S}_w$ and between-class scatter matrix $\mathbf{S}_b$. Then, we compute the LDA project matrix by finding the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$. Then we project each vector by the LDA project matrix and train 10 one-vs-rest SVMs (one for each digit) using the LDA-projected vectors. *Pros*: The LDA-projected vectors are class discriminative because the projection matrix is found by maximizing the between-class scatter and minimizing the within-class scatter. *Cons*: The maximum dimension of the LDA-projected vectors is 9, which may be too small for the SVMs.

(7 marks, AE)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME ~~(2017/2018)~~ 2019/2020

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

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

3. (a) The pooling operation in CNNs serves two purposes: (1) it reduces the size of the feature maps in the subsequent layers and (2) it reduces the resolution of the feature maps so that the responses of the maps will become less dependent on the position of the object to be recognized/classified. Omitting the pooling operation will lead to large feature maps at the final convolutional layer, which in turn leads to a large number of inputs to the fully-connected layers and a large number of weights.

(5 marks, AE)

(b) The convolutional filters in the first hidden layer aim to detect primitive features in the input, where the input could be an image or a waveform. The primitive features could be edges in various orientations. For 1-D convolutional filters on waveform, the primitive features could be the strength of the waveform at different frequency bands. Because the kernel size is typically very small, e.g., 3×3, each filter could only focus on one primitive features. To detect more features in complex signals, we need a number of filters.

(5 marks, AE)

(c) The fully connected layers act as a classifier. The vectors at the final convolutional/pooling layer are reshaped to form the input to this classifier. Typically, the softmax activation function is used for the output layer because it makes the outputs sum to a 1.0. As a result, we may consider the outputs as the posterior probabilities of individual classes.

(5 marks, AE)

(d) For the mean squared error function to have no local minimum, we need it to be a quadratic function with respect to the weights. The only case this happens is that all processing nodes are linear. However, deep neural networks typically have multiple layers of non-linear processing nodes. The combination of these non-linear nodes makes the mean squared error highly non-linear with respect to the weights, which cause a lots of local minimum in the error function.

(5 marks, AE)

(e) Because the number of syllables varies from word-to-word and each state in an HMM can only model a sub-part of a syllable, the required number of states for modeling a completed word varies from word-to-word. The rule of thumb is that the longer the word (number of letters), the larger the number of syllables. As a result, we need more states for modeling longer words. Each syllable requires at least 3 states to model. For examples, for words comprising three syllables, we need 9 states plus some short pause states.

(5 marks, AE)

4. (a) The i-vector of an utterance is the posterior mean of the latent factor of a factor analysis model:

$$\boldsymbol{\mu} = \boldsymbol{\mu}^{(b)} + \mathbf{T}\mathbf{w},$$

where $\mathbf{T}$ is the total variability matrix and $\mathbf{w}_i$ is the latent factor. Given an utterance with T acoustic vectors $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, the corresponding i-vector is given by

$$\mathbf{x} \equiv \langle \mathbf{w} | \mathcal{O}_i \rangle = \mathbf{L}^{-1} \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \left(\boldsymbol{\Sigma}_c^{(b)}\right)^{-1} \tilde{\mathbf{f}}_c$$

where

$$\mathbf{L} = \mathbf{I} + \sum_{c=1}^{C} N_c \mathbf{T}_c^{\mathsf{T}} (\boldsymbol{\Sigma}_c^{(b)})^{-1} \mathbf{T}_c,$$

$$N_c \equiv \sum_{t=1}^{T} \gamma(\ell_{t,c}) \quad \text{and} \quad \tilde{\mathbf{f}}_c \equiv \sum_{t=1}^{T} \gamma(\ell_{t,c})(\mathbf{o}_t - \boldsymbol{\mu}_c^{(b)}),$$

where $\gamma(\ell_{t,c})$ is the posterior probability of the c-th mixture of the UBM. In this set of equations, the number of acoustic vectors, T, depends on the utterance length. However, because $\tilde{\mathbf{f}}_c$ is obtained by summing over all frames (from $t = 1$ to $t = T$), the dimension of $\tilde{\mathbf{f}}_c$ is fixed, so as the dimension of $\mathbf{L}$. Therefore, the dimension of $\mathbf{x}$ is independent of T.

(10 marks, E)

(b) When the decision threshold is small, both true-speakers and impostors will be accepted by the system, which causes high false acceptance rate but low false rejection rate. On the other hand, when the decision threshold is large, both true-speakers and impostors will be rejected by the system, which causes high false rejection rate but low false acceptance rate. That's why there is a tradeoff between FAR and FRR. If security is a major concern, we should set the decision threshold very high.

(5 marks, AE)

(c) (i) The phone-specific acoustic models in LVCSR are to compute the likelihood of acoustic vector sequences given various phonetic units (can be phones or triphones). Given a hypothesized word sequence, we use a dictionary to find the corresponding phone sequence. Then, an acoustic model is formed by joining the phone-specific acoustic models corresponding to the phones in the sequence. Then, the acoustic vectors are aligned to the internal states of the acoustic model, from which the likelihood of the whole acoustic vector sequence can be computed from the joined acoustic model.

(6 marks, AE)

(ii) GMM-HMM: The states (which are GMMs) can be easily shared across different phone-specific HMMs. It requires less data to train when compared with DNN-

HMM.

DNN-HMM: It uses discriminative training to determine the weights of the DNN, which produces more accurate posterior probabilities of phones. It performs better than GMM-HMM provided that sufficient data are available for training the DNN. The input to the DNN contains multiple contextual frames instead of a single frame as in GMM, which allows the DNN to capture the dynamic of acoustic vectors.

(4 marks, E)