Q1 You are given a dataset comprising the facial images of ten persons. Each person has 100 images and each image has size 128×128 pixels. Assume that for each image, a facial vector is obtained by stacking the columns of the image.

(a) What will be the dimension of the resulting vectors?

(2 marks)

(b) Assume that you train a set of one-versus-rest support vector machines (SVMs) to construct an SVM classifier for classifying the ten persons. Draw a block diagram illustrating the architecture of your SVM classifier.

(5 marks)

(c) Suggest the most appropriate kernel function for the SVMs. Briefly explain your choice.

(5 marks)

(d) State two disadvantages of converting a facial image to a facial vector by stacking its columns.

(4 marks)

(e) Assume that principal component analysis (PCA) is applied to project the facial vectors onto a lower dimensional space. What will be the maximum dimension of the PCA-projected vectors? You may assume that all of the images (facial vectors) are independent.

(4 marks)

Q2 A dataset comprises the fingerprints of K persons. Each person in the dataset has N fingerprints collected independently from the same finger and each fingerprint is stored as a 200×100 pixel image. For the k-th person, his/her fingerprints are denoted as $\mathcal{X}_k = \{\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,N}\}$, where $\mathbf{x}_{k,i}$ is a fingerprint vector formed by stacking the rows of the corresponding image. Assume that the fingerprint vectors of the K persons share a global covariance matrix:

$$\boldsymbol{\Sigma} = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} (\mathbf{x}_{k,i} - \boldsymbol{\mu})(\mathbf{x}_{k,i} - \boldsymbol{\mu})^{\mathsf{T}},$$

where

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

Assume also that a Gaussian classifier based on $\boldsymbol{\Sigma}$ is used for classifying these K persons.

(a) If $K = 10$, what is the theoretical minimum value of N for the Gaussian classifier to recognize the persons in this dataset? Show how you calculate this value.

(4 marks)

(b) Determine the theoretical minimum value of N if we have the following relationships:

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

where $k = 1, \ldots, K$, $\odot$ denotes element-wise multiplication, and $\mathbf{1}$ is a vector containing all 1's. Briefly explain your answer.

(6 marks)

(c) Are the decision boundaries produced by this Gaussian classifier linear or non-linear? Briefly explain your answer.

(4 marks)

(d) Assume that the fingerprint vectors of each person are modelled by a Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th person, respectively. Without increasing the value of N as determined in Q2(a), how would you ensure that the new Gaussian classifier (with person-dependent covariance matrices) can recognize the fingerprints of these 10 persons? Briefly explain your answer.

(6 marks)

(e) Denote $P(C_k)$ as the prior probability of the k-th person. Discuss how you would classify these 10 persons based on the Bayes' theorem. Your answer should include an equation comprising $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $P(C_k)$.

(5 marks)

Q3 You are given a dataset comprising the images of handwritten digits. Each digit has 10,000 images and each image has size 28×28 pixels.

(a) Assume that for each image, a vector is obtained by stacking the columns of the image. Assume also that you train a deep neural network (DNN) to classify the handwritten digits into 10 classes.

(i) State the number of inputs and the number of outputs of the network.

(2 marks)

(ii) What is the purpose of the bias terms in the hidden layers of the DNN?

(3 marks)

(iii) Suggest a non-linear function for the network outputs and an objective function (loss function) for training the DNN. Briefly explain your suggestion.
(5 marks)

(iv) Typically, non-linear activation functions (e.g., sigmoid, hyperbolic tangent, ReLU, etc.) are used in the hidden nodes. Explain why it is a bad idea to use a linear activation function for all of the hidden nodes in a DNN. Your answer should include an equation expressing the relationship between the network output $\mathbf{y}$, the network input $\mathbf{x}$, and the hidden-layer weight matrices (including bias terms) $\{\mathbf{W}_l\}_{l=1}^{L}$, where L is the number of hidden layers.
(7 marks)

(b) Assume that a convolutional neural network (CNN) with a number of convolutional layers, max-pooling layers, and fully connected layers is applied to classify the handwritten digits.

(i) If a 3×3 kernel is used in the first convolutional layer and the number of feature maps (also called output channels) is 64, what will be the number of shared weights in the first convolutional layer? You may assume that the images are black and white.
(2 marks)

(ii) State the purpose of the convolutional layers.
(2 marks)

(iii) State the purpose of the max-pooling layers.
(2 marks)

(iv) State the purpose of the fully connected layers.
(2 marks)

Q4 Fig. Q4 shows the waveform and spectrogram of the phoneme [iː] in the word "speech" and "each". Also shown is the waveform and spectrogram of the phoneme [ɪ] in the word "it".

(a) The value 0.6 in Fig. Q4 is the probability of remaining in the first state. Suggest the probability of transiting from State 1 to State 2 in the HMM. Show how you calculate this probability.
(3 marks)

(b) If the vertical dashed lines indicate the three sections of the phonemes. Deduce roughly the probability of remaining in the first state for the HMM that models the phoneme [ɪ] in the word "it". Briefly explain your answer.
(6 marks)

(c) Denote $p(\mathcal{X}_p|\Lambda_q)$ as the likelihood of $\mathcal{X}_p$ given the HMM model Λ_q corresponding to Phoneme q, where $\mathcal{X}_p$ comprises the acoustic vectors (MFCCs) corresponding

to Phoneme p. State if the following conditions are true or false.

$$\begin{aligned}
&\text{Condition 1}: \quad p(X_{\text{i:}}|\Lambda_{\text{i:}}) > p(X_{\text{i:}}|\Lambda_{\text{I}}) \\
&\text{Condition 2}: \quad p(X_{\text{i:}}|\Lambda_{\text{i:}}) < p(X_{\text{I}}|\Lambda_{\text{i:}}) \\
&\text{Condition 3}: \quad p(X_{\text{I}}|\Lambda_{\text{I}}) = p(X_{\text{i:}}|\Lambda_{\text{i:}})
\end{aligned}$$

Briefly explain your answers. (6 marks)

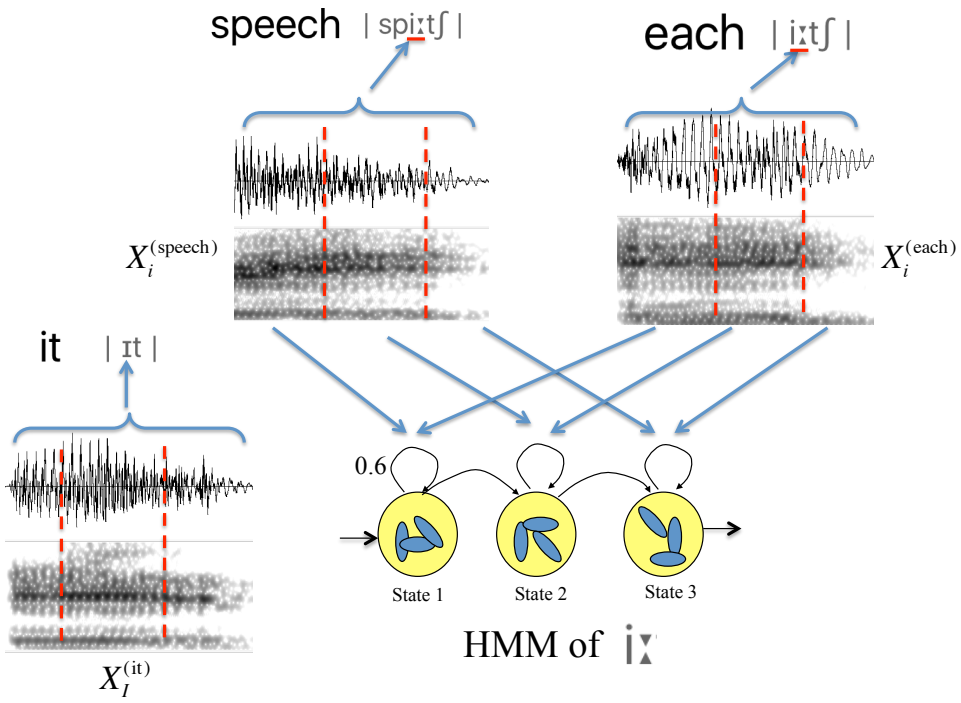

Fig. Q4

Q5 In i–vector based speaker verification, the dimension of the GMM-supervector $\vec{\boldsymbol{\mu}}_s$ corresponding to Speaker s is reduced by the following factor analysis model:

$$\vec{\boldsymbol{\mu}}_s = \vec{\boldsymbol{\mu}} + \mathbf{T}\mathbf{w}_s,$$

where $\mathbf{T}$ is a low-rank matrix called the total variability matrix, $\mathbf{w}_s$ is a speaker factor and $\vec{\boldsymbol{\mu}}$ is a mean supervector formed by stacking the mean vectors of a universal background model.

(a) Explain why this factor analysis model can reduce the dimension of $\vec{\boldsymbol{\mu}}_s$.
(4 marks)

(b) Why does the total variability matrix define both the speaker and session (channel) variability?
(4 marks)

(c) The i-vector $\mathbf{x}_s$ of Speaker s is the posterior mean of $\mathbf{w}_s$, which can be obtained from $\mathbf{T}$ and the acoustic vectors derived from his/her utterance. The i-vector $\mathbf{x}_c$ of a claimant is obtained in the same manner. During a verification session, we are given the i-vectors of Speaker s and Claimant c. A naive way is to accept Claimant c if the cosine-distance score is larger than a decision threshold η, i.e.,

$$S_{\text{cosine}}(\mathbf{x}_s, \mathbf{x}_c) = \frac{\mathbf{x}_s^{\mathsf{T}} \mathbf{x}_c}{\|\mathbf{x}_s\| \|\mathbf{x}_c\|} > \eta,$$

where $^{\mathsf{T}}$ and $\|\cdot\|$ denote vector transpose and vector norm, respectively. Explain why this naive approach is undesirable for speaker verification. Suggest a method to pre-process the i-vectors to remedy this undesirable situation.

(7 marks)

– END OF PAPER –