Version MJFG/5

EGT3
ENGINEERING TRIPOS PART IIB

Monday 24 April 2017 9.30 to 11

Module 4F10

STATISTICAL PATTERN PROCESSING

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

1 A generative classifier is to be built for a two-class problem. The observation vectors for this task are d-dimensional. The class conditional distributions for the two classes are Gaussian. The parameters for class ω_1 are μ_1 and Σ_1 and those for class ω_2 are μ_2 and Σ_2. The priors for the two classes are $P(\omega_1)$ and $P(\omega_2)$.

(a) State Bayes' decision rule for this task. Under what conditions will this form of generative classifier yield a classifier with the minimum probability of error? [15%]

(b) The covariance matrices for the two classes are constrained to be the same and diagonal, $\Sigma_1 = \Sigma_2 = \Sigma$, and the priors for the two classes are equal.

(i) The model parameters are trained on n training observations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with class labels $y_1, \ldots, y_n$. If observation $\mathbf{x}_i$ belongs to class ω_1 then $y_i = 1$, and if it belongs to class ω_2 then $y_i = 0$. The log-likelihood of the training data can be expressed as

$$\mathscr{L}(\mu_1, \mu_2, \Sigma) = \sum_{i=1}^{n} \left(y_i \log(\mathscr{N}(\mathbf{x}_i; \mu_1, \Sigma)) + (1 - y_i) \log(\mathscr{N}(\mathbf{x}_i; \mu_2, \Sigma)) \right)$$

What is the maximum-likelihood estimate of the covariance matrix Σ? [20%]

(ii) Show that the posterior probability for class ω_1 given an observation $\mathbf{x}$ can be expressed in the form

$$P(\omega_1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}'\phi(\mathbf{x}) + a)}$$

where $\phi(\mathbf{x})$ is a function of $\mathbf{x}$ that yields a d-dimensional vector. Find expressions for $\mathbf{w}$, $\phi(\mathbf{x})$ and a. [25%]

(c) The means of the two class-conditional distributions are now constrained to be zero, $\mu_1 = \mu_2 = \mathbf{0}$. The covariance matrices for the two classes are allowed to be different, but constrained to be diagonal. Again the priors for the two classes are equal. Show that the posterior probability for class ω_1 can be expressed in the same form as in part (b)(ii) and find expressions for $\mathbf{w}$, $\phi(\mathbf{x})$ and a in this case. [25%]

(d) Compare the decision boundaries that result from the two forms of classifier in parts (b) and (c). [15%]

2 A set of K, M-component, Gaussian mixture models (GMMs) are to be used to extract a representation for a particular speaker. The component priors, $c_1,\ldots,c_M$, are the same for all K GMMs. The component mean vectors for GMM k are given by, $\mu_1^{(k)},\ldots,\mu_M^{(k)}$ and are GMM specific. All component covariance matrices are identity matrices. There are N d-dimensional training examples, $\mathbf{x}_1,\ldots,\mathbf{x}_N$, to estimate the representation of a particular speaker.

(a) Find an expression for the log-likelihood of the data for the speaker using only the k^{th} GMM. This should be expressed in terms of the parameters of the model. [10%]

(b) The mean for the speaker is formed by combining the means of all the components. Thus for component m of the speaker

$$\mu_m = \sum_{k=1}^{K} \lambda_k \mu_m^{(k)} = \left[\begin{array}{ccc} \mu_m^{(1)} & \ldots & \mu_m^{(K)} \end{array} \right] \lambda$$

where λ_k is the "weight" assigned to GMM k to represent the speaker. λ is the K-dimensional vector formed from λ_1 to λ_K. Expectation-Maximisation (EM) is to be used to find the vector λ. The auxiliary function for this problem has the form

$$Q(\lambda,\hat{\lambda}) = z + \sum_{i=1}^{N}\sum_{m=1}^{M} P(\omega_m|\mathbf{x}_i,\lambda)\log(p(\mathbf{x}_i|\omega_m,\hat{\lambda}))$$

(i) Describe how EM can be used to find the value of λ. You should clearly describe the terms in the auxiliary function shown above. [20%]

(ii) Show that the auxiliary function can be expressed in the following form

$$Q(\lambda,\hat{\lambda}) = e + \hat{\lambda}'\mathbf{b} + \hat{\lambda}'\mathbf{C}\hat{\lambda}$$

Find expressions for the scalar e, vector $\mathbf{b}$ and matrix $\mathbf{C}$ in terms of z, $P(\omega_m|\mathbf{x}_i,\lambda)$ and the parameters of the K GMMs. [35%]

(iii) Hence derive an expression to estimate the optimal value of $\hat{\lambda}$ in terms of $\mathbf{b}$ and $\mathbf{C}$. What practical issues may occur with this expression? [15%]

(c) The vector λ is to be used in a speaker recognition system. Briefly describe how this vector could be used in such a system. You should mention the computational cost of the approach if there are a total of S possible speakers. [20%]

3 A classifier based on support vector machines (SVMs) is to be used for a K-class classification problem. There are a total of m training samples, $\mathbf{x}_1$ to $\mathbf{x}_m$, with associated class labels, y_1 to y_m. $\phi(\mathbf{x})$ is the mapping from the *input-space* to the *feature-space*. In this feature-space the training examples are linearly separable for all classes.

(a) Initially a set of binary SVMs are trained. An SVM is trained for each possible pair of classes.

(i) For a particular pair of classes ω_p and ω_q, what condition must be satisfied by all training examples of these classes for the trained SVM? [15%]

(ii) For the class pairing ω_p and ω_q, any point $\mathbf{x}$ on the SVM decision boundary can be expressed in the following forms

$$\mathbf{w}^{(pq)\prime}\phi(\mathbf{x}) + b^{(pq)} = \sum_{i=1}^{m} \alpha_i^{(pq)} \phi(\mathbf{x}_i)^\prime \phi(\mathbf{x}) + b^{(pq)} = 0$$

Discuss how the values of $\alpha_i^{(pq)}$ and $b^{(pq)}$ can be found for the class pairing ω_p and ω_q. [20%]

(iii) Briefly describe one scheme for combining the multiple binary SVM classifiers together for K-class classification. You should comment on the computational cost and any issues associated with the proposed scheme. [20%]

(b) The SVM classifier is extended to directly handle the K-class classification problem. A *single* SVM is to be used for the multi-class classification. For this classifier sample $\mathbf{x}$ is classified as class ω_k if

$$\mathbf{w}^{(k)\prime}\phi(\mathbf{x}) + b^{(k)} - \mathbf{w}^{(j)\prime}\phi(\mathbf{x}) - b^{(j)} > 0 \text{ for all } j \neq k$$

where $\mathbf{w}^{(k)}$ and $b^{(k)}$ are the parameters associated with class ω_k.

(i) Show that this decision rule for classifying sample $\mathbf{x}$ as ω_k can be written as

$$\tilde{\mathbf{w}}^\prime \mathbf{z}_j > 0 \text{ for all } j \neq k$$

where $\tilde{\mathbf{w}}^\prime = \begin{bmatrix} b^{(1)} & \mathbf{w}^{(1)\prime} & \ldots & b^{(K)} & \mathbf{w}^{(K)\prime} \end{bmatrix}$. Clearly state the form of $\mathbf{z}_j$. [15%]

(ii) Discuss how the SVM could be trained in this case. [15%]

(iii) Compare the computational cost of classification with this approach to the scheme described in part a(iii). [15%]

4 The parameters of a deep neural network (multi-layer perceptron) are to be trained using a quadratic approximation to the error surface. The set of weights associated with the classifier are denoted as the vector θ. An iterative procedure is used to estimate the parameters where at iteration $\tau+1$

$$\theta^{(\tau+1)} = \theta^{(\tau)} + \Delta\theta^{(\tau)}$$

and $\theta^{(\tau)}$ is the estimate of the model parameters at iteration τ. The value of the cost function with model parameters θ is $E(\theta)$.

(a) The following quadratic approximation is to be used to estimate the weights

$$E(\theta) \approx E(\theta^{(\tau)}) + (\theta - \theta^{(\tau)})'\mathbf{b} + \frac{1}{2}(\theta - \theta^{(\tau)})'\mathbf{A}(\theta - \theta^{(\tau)})$$

(i) By considering a second-order Taylor series expansion about the point $\theta^{(\tau)}$ find expressions for $\mathbf{b}$ and $\mathbf{A}$. [10%]

(ii) Derive an expression for the value of θ that will minimise this quadratic approximation. Hence obtain an expression for $\Delta\theta^{(\tau)}$. [20%]

(b) The network has been trained to a local minimum to yield parameters $\theta^\star$. The aim is to reduce the number of parameters by removing links, effectively setting the weights of particular links to zero. The remaining parameters are then updated to yield $\hat{\theta}$ using the quadratic approximation in part (a).

(i) The link associated with element q of the vector of network parameters is to be removed, so that $\hat{\theta}_q = 0$. By using this constraint with a Lagrange multiplier, or otherwise, show that the updated model parameters $\hat{\theta}$ are given by

$$\hat{\theta} = \theta^\star - d\mathbf{A}^{-1}\mathbf{c}$$

where $\mathbf{A}$ is evaluated for the Taylor series expansion about $\theta^\star$. What are the values of the vector $\mathbf{c}$ and the scalar d? [30%]

(ii) What is the change in the cost function, $E(\theta^\star) - E(\hat{\theta})$? [15%]

(iii) Describe how this form of approach can be used to prune multiple weights from the network. [10%]

(iv) For large networks it is proposed that $\mathbf{A}$ is restricted to be diagonal. Discuss the impact that this has on both the computational cost and the estimation of $\hat{\theta}$? [15%]

5 A deep neural network (multi-layer perceptron) is to be trained for a K-class problem, with a d-dimensional observation vector for each of the training samples. Supervised training is to be used. The activation function for the output layer is a *softmax* activation function. All other layers have activation functions of the form

$$\phi(z_i) = \begin{cases} \alpha z_i; & z_i \geq 0 \\ \beta z_i; & z_i < 0 \end{cases}$$

where α and β are the same for all layers. The number of hidden layers for the network is L, with the number of nodes for layer l being N_l.

(a) Discuss what needs to be considered when determining L and the number of nodes for each layer. You should include the total number of model parameters, including the bias term, in the network described. [15%]

(b) A large number of samples, generated from a Gaussian distribution with zero mean and variance σ^2, are passed through the hidden layer activation function.

(i) What is the mean and variance of the data at the output of the activation function? [30%]

(ii) How could this information be used when initialising the parameters of the network? You should consider how both the mean and the variance can be used. [20%]

(c) The values of α and β are set to be the same, such that $\alpha = \beta = \alpha_0$.

(i) How does this alter the choice of the number of layers and the number of model parameters? Does this alter the choice of how to initialise the network? You should justify your answers. [15%]

(ii) Does the choice of α_0 alter the answers to part (b)(ii) and (c)(i)? Will it alter the final performance of the network? [10%]

(d) Rather than setting the values of α and β to be the same for every node in the hidden layers, α and β are trained. Briefly discuss whether this is expected to be useful. [10%]

END OF PAPER