ENGINEERING TRIPOS PART IIB

Wednesday 9 May 2007 2.30 to 4

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.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS
Engineering Data Book
CUED approved calculator allowed

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

1 A two-class pattern recognition problem for d-dimensional data has prior probabilities P_1 and P_2. The class conditional probability density functions are Gaussians with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2.

(a) What is the minimum error-rate classifier for this problem in terms of the Gaussian parameters? [15%]

(b) Show that a point, $\mathbf{x}$, that lies on the decision boundary satisfies an expression of the form

$$\mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{x} + c = 0$$

and find expressions for $\mathbf{A}$, $\mathbf{b}$ and c. [25%]

(c) The *Fisher linear discriminant* is to be derived.

(i) What is meant by the Fisher linear discriminant? Include in your answer the cost function that is optimised and the direction of the discriminant vector. [15%]

(ii) How can a classification rule be generated from the Fisher discriminant? [10%]

(iii) For the case that $P_1 = P_2$, $\mu_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\mu_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$, calculate the Fisher discriminant and give an appropriate classification rule. [20%]

(iv) Under what conditions does the classification rule with the Fisher discriminant and the Gaussian classifier in part (b) give the same decision boundary? [15%]

2 An M-component *Gaussian Mixture Model* (GMM) is to be estimated for d-dimensional data from n independent training samples $\mathbf{x}_1 \ldots \mathbf{x}_n$. Maximum likelihood estimation of the model parameters is to be performed using an Expectation-Maximisation (EM) approach.

(a) Write down the log-likelihood function, $l(\theta)$, for the mixture model in terms of the individual mixture component likelihoods where θ denotes the model parameters. [10%]

(b) Briefly explain why a direct closed-form maximum likelihood estimation of the mixture model parameters cannot be used. [10%]

(c) By using Jensen's inequality, show that

$$l(\hat{\theta}) - l(\theta) \geq Q(\theta, \hat{\theta}) - Q(\theta, \theta)$$

where the auxiliary function, $Q(\theta, \hat{\theta})$, is defined as

$$Q(\theta, \hat{\theta}) = \sum_{i=1}^{n} \sum_{m=1}^{M} P(\omega_m | \mathbf{x}_i, \theta) \log \left(p(\mathbf{x}_i, \omega_m | \hat{\theta}) \right).$$

θ denotes the current estimates of the model parameters, $\hat{\theta}$ the new parameter values to be estimated, and ω_m the m^{th} mixture component. What is the significance of this result? [40%]

(d) By maximising the auxiliary function value, derive an expression for a new estimate, $\hat{\mu}_m$, of the mean vector of the m^{th} mixture component of a GMM. [20%]

(e) Outline a practical procedure for estimating GMM parameters using EM. What are the main issues that need to be addressed? [20%]

3 A classifier is to be trained for a K-class classification problem. Supervised training data is available, with m independent, d-dimensional, training samples for each class. For class ω_k there are training examples $\mathbf{x}_1^{(k)}$ to $\mathbf{x}_m^{(k)}$. The decision rule used in the classifier is that sample $\mathbf{x}$ is classified as class ω_k if

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

where $\mathbf{w}^{(k)}$ and $b^{(k)}$ for $k = 1, \ldots, K$ are the parameters of the classifier.

(a) Initially consider a three class problem, $K = 3$, and a 2-dimensional feature vector, $d = 2$. The following classifier parameters are given.

$$\mathbf{w}^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad b^{(1)} = 0, \quad \mathbf{w}^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad b^{(2)} = 0, \quad \mathbf{w}^{(3)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad b^{(3)} = 0.$$

(i) For each of the three sets of parameters draw on the same diagram the lines where $\mathbf{x}$ satisfies $\mathbf{w}^{(k)\prime}\mathbf{x} + b^{(k)} = 0$. [20%]

(ii) Sketch on the diagram the decision boundaries and associated class labels for these classifier parameter values. [30%]

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

$$\tilde{\mathbf{w}}'\mathbf{z}_j > 0, \quad \text{for all } j \neq k$$

where $\tilde{\mathbf{w}}$ is the $K \times (d+1)$-dimensional vector

$$\tilde{\mathbf{w}}' = \begin{bmatrix} b^{(1)} & \mathbf{w}^{(1)\prime} & \ldots & b^{(K)} & \mathbf{w}^{(K)\prime} \end{bmatrix}$$

Clearly show the form of the extended vector $\mathbf{z}_j$. [15%]

(c) The parameters of the classifier are to be trained using a *Support Vector Machine* (SVM). You may assume that all training samples can be correctly classified with the appropriate classifier parameters.

(i) Discuss how the SVM parameters can be trained using the extended vector, $\mathbf{z}_j$, in part (b). The training criterion should be clearly stated. Why is an SVM a suitable classifier? [25%]

(ii) A Gaussian kernel is to be used with the SVM classifier. State the form of the decision boundary that can be efficiently used in this case. [10%]

4 A *binary decision-tree classifier* is to be constructed on a set of independent d-dimensional samples for a K-class classification problem. The elements of the feature vector associated with each sample are continuous-valued. In addition, each sample has a number of binary attributes and a class label associated with it. A set of Q questions, $\{q_1,\ldots,q_Q\}$, related to the binary attributes of the observations are specified for building the decision tree. The decision tree is to be constructed to maximise the log-likelihood of the training data. Therefore the question selected to split the data associated with a node is the one that yields the greatest increase in log-likelihood. Each node of the decision tree is represented by a multivariate Gaussian distribution, whose parameters are trained using Maximum Likelihood (ML) estimation on the data associated with that node. Diagonal covariance matrices are used for each distribution.

(a) There are m independent data samples, $\mathbf{X}^{(p)} = \{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$, associated with node N_p.

(i) What is the log-likelihood of the data $\mathbf{X}^{(p)}$ in terms of the Gaussian distribution, $\mathcal{N}(\mathbf{x};\mu^{(p)},\Sigma^{(p)})$, representing node N_p? [10%]

(ii) Show that the log-likelihood of the data $\mathbf{X}^{(p)}$ can be re-written as

$$\log\left(p(\mathbf{X}^{(p)}|\mu^{(p)},\Sigma^{(p)})\right) = \alpha + \beta\log\left(|\Sigma^{(p)}|\right)$$

when the ML estimates of the parameters of the multivariate Gaussian, $\mu^{(p)}$ and $\Sigma^{(p)}$, trained using $\mathbf{X}^{(p)}$ are used. Find expressions for the constants α and β. [30%]

(b) Derive an expression for the change in log-likelihood from applying question q to node N_p to form nodes N_L and N_R. You may assume that the question splits the data from node N_p so that the first r samples are assigned to node N_R. [20%]

(c) Discuss the form of stopping criterion that may be used to decide when to stop constructing the decision tree and how the decision tree would be used for classification. [20%]

(d) Contrast the use of this log-likelihood measure with using an Entropy impurity measure for training the decision tree. You should consider the form of the purity measure, computational cost, and how well you feel the trees will perform for classification. You should state the form of the Entropy purity measure. [20%]

5 (a) Define what is meant by a *Gaussian Process*. [10%]

(b) An interpolation function using basis functions and a linear model of the form

$$y = f(x) + \varepsilon, \quad f(x) = \sum_{k=1}^{H} w_k \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)$$

with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is to be trained. There are n training examples consisting of the 1-dimensional observations, $x_1, \ldots, x_n$, and target values $y_1, \ldots, y_n$.

(i) Briefly discuss why this is a more powerful interpolation model than linear regression. What effect does the value of σ^2 have on the interpolation process? [20%]

(ii) Each of the weights has a Gaussian prior with a mean of zero and variance σ_w^2. Find an expression for the n-dimensional mean and $n \times n$ covariance matrix of the distribution of the target values, $p(\mathbf{y})$, where $\mathbf{y} = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}'$, in terms of the observations $x_1, \ldots, x_n$. [25%]

(iii) The number of basis functions is increased so that $H \to \infty$. The positions of centres of the basis functions, μ_k, are Gaussian distributed with a mean of zero and variance of σ_h^2, and $\sigma_h^2 \gg \sigma^2$. The variance of the weights' prior scales linearly with σ_h^2 and inversely with the number of basis functions, so $\sigma_w^2 = \sigma_h^2/H$. Show that the covariance matrix of $p(\mathbf{y})$ can be approximated as a covariance function with elements of the form

$$C(x_i, x_j) \approx \alpha \exp\left(-\beta(x_i - x_j)^2\right) + \gamma\delta(i-j)$$

where

$$\delta(i-j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

What are the values of α, β and γ? [30%]

(iv) Discuss how the effective number of weight parameters varies as $H \to \infty$. [15%]

END OF PAPER