ENGINEERING TRIPOS PART IIB

Tuesday 20 April 2004 2.30 to 4

Module 4F10

STATISTICAL PATTERN PROCESSING

*Answer not more than **two** 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.*

There are no attachments.

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

(TURN OVER

1 (a) What is the Bayes' minimum error rate classification rule for a multi-class problem? What determines how closely a practical classifier approaches the Bayes' minimum error rate? [15%]

(b) Class-conditional probability density functions (PDFs) are to be estimated for a supervised training task. A 20 dimensional feature vector is used. The class-conditional PDFs are thought to be approximately Gaussian distributed and have class-dependent correlations between the elements of the feature vector. Compare in terms of modelling capabilities, estimation issues, number of parameters, and computational complexity the use of either a single full covariance Gaussian model, or a Gaussian mixture model (GMM) with M diagonal covariance components, for each class-conditional PDF. [20%]

(c) What are the advantages of using expectation maximisation (EM) rather than gradient descent based methods for estimating the parameters of a GMM? [10%]

(d) The component priors of a GMM are to be estimated using maximum likelihood estimation. There are N independent training vectors $\mathbf{x}_1$ to $\mathbf{x}_N$.

(i) Write down the log likelihood function, $l(\theta)$, of a GMM for the training data. [10%]

(ii) Find the partial derivative of $l(\theta)$ with respect to the prior of component m, c_m, in terms of the posterior probability of component occupation for training vector $\mathbf{x}$, $P(m|\mathbf{x})$. [15%]

(iii) Using the method of Lagrange multipliers, find an expression that must be satisfied for the maximum likelihood values of the component priors. [15%]

(iv) Explain, without derivation, how the formula found in (d)(iii) can be used to estimate the component priors using the EM algorithm. Assume that the values of the means and covariance matrices are known, and the component priors are randomly initialised and appropriately normalised. [15%]

2 (a) A multi-layer perceptron (MLP) is to be used for a classification task. There are C possible classes. The MLP has a d-dimensional input, an output layer, L hidden layers, and $N^{(k)}$ units in the k^{th} hidden layer.

(i) What is the purpose of each layer of the MLP? Discuss the issues that must be considered when selecting the number of network layers and nodes per layer for a particular problem. [15%]

(ii) Write down a general expression for the total number of weights (including biases) in the MLP. [10%]

(b) The MLP is to be trained using error-back propagation with a least squares error criterion. For each input pattern $\mathbf{x}_i$ the target vector is $\mathbf{t}(\mathbf{x}_i)$. All nodes use a sigmoid logistic activation function of the form

$$y(z) = \frac{1}{1 + \exp(-z)}$$

(i) Find the differential of this activation function with respect to z in terms of the output of the activation function. Hence, find the partial derivative of the output error with respect to a particular weight from the output of the final hidden layer to the output layer. [25%]

(ii) Show how this partial derivative of the error with respect to the weights can be used in a gradient descent optimisation scheme with learning rate η to find the weights of the final hidden layer. Describe how η should be chosen for fast convergence. [15%]

(iii) It is suggested that a momentum term be added to the gradient descent weight update formula. Explain what a momentum term is and the effects that it has on the training procedure. [15%]

(c) The MLP described in part (b) is to be modified so that the output layer uses the soft-max activation function. What is the form of the soft-max activation function? Discuss how the training procedure must be modified to support this change. [20%]

3 A classifier is required for a two class problem. There are a total of m training samples $\mathbf{x}_1$ to $\mathbf{x}_m$ with associated labels y_1 to y_m where $y_i \in \{-1, 1\}$.

(a) Initially a linear classifier is to be constructed. Contrast the training criteria used to train a support vector machine (SVM) classifier and the perceptron algorithm classifier when the training data is linearly separable. How is the training criterion for the SVM altered for the case when the training data is not separable? [25%]

(b) Discuss how the use of kernel functions may be used to improve the performance of a SVM classifier. What is the general form for a polynomial kernel-function? [15%]

(c) The training samples are 1-dimensional. The following mapping is proposed from the 1-dimensional *input-space* to the $(2N+1)$-dimensional *feature-space*.

$$\mathbf{\Phi}(x) = \left[\begin{array}{ccccccccc} \frac{1}{\sqrt{2}} & \cos(x) & \cos(2x) & \ldots & \cos(Nx) & \sin(x) & \sin(2x) & \ldots & \sin(Nx) \end{array} \right]'$$

where x is the point in the input-space

(i) Show that the kernel-function, the dot-product of two vectors in the feature-space, between two points x_i and x_j for this mapping may be expressed in the following form

$$k(x_i, x_j) = \frac{\sin(a(x_i - x_j))}{2\sin(b(x_i - x_j))}$$

What are the values of a and b? [25%]

(ii) Express the classification rule using the kernel-function in its dual form which is a function of the support vectors. How does the computational cost of classification vary as the number of support vectors, S, number of training samples, m, and N change? [15%]

(d) The SVM classifier is to be extended to handle classification problems with more than two classes. Discuss how the SVM training and classification might be modified to allow a *single* SVM classifier to perform multi-class classification. [20%]

4 A Parzen window is to be used to estimate the class-conditional density for a pattern classification task.

(a) Contrast the use of a Parzen window density estimate with using a single multivariate Gaussian distribution as the class-conditional density. You should comment on memory requirements, computational cost and factors that will affect the performance. [20%]

(b) The form of the Parzen window density estimate $\tilde{p}(\mathbf{x})$ for the the d-dimensional vector $\mathbf{x}$ is given by

$$\tilde{p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

where the training data consists of training samples $\mathbf{x}_1$ to $\mathbf{x}_n$.

(i) Discuss how the value of h affects the Parzen window density estimate. How should the value of h be varied as n changes? [15%]

(ii) Show that if the window function $\phi(\mathbf{x})$ is a valid probability density function, then the Parzen window estimate $\tilde{p}(\mathbf{x})$ will also be a valid probability density function. [15%]

(c) For a particular application the data is one dimensional, $d = 1$, and the form of the window function is a Gaussian.

(i) By using a first order Taylor series expansion based around $\phi(0)$, show that the Parzen window estimate $\tilde{p}(x)$ may be approximated as

$$\tilde{p}(x) \approx b_0 + b_1 x + b_2 x^2$$

where b_0, b_1 and b_2 are only functions of the training data. What are the values of b_0, b_1 and b_2? [25%]

(ii) Discuss how the use of this approximation affects the memory requirements and computational speed of using the Parzen window. [10%]

(iii) What will affect how good this approximation is to the exact Parzen window density estimate? How could the approximation be improved? [15%]

END OF PAPER