ENGINEERING TRIPOS PART IIB

Tuesday 7 May 2013 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.*

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Single-sided script paper	Engineering Data Books
	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

wjb04

1 (a) The class-conditional distribution in a classifier for d-dimensional data is to be either an M-component Gaussian mixture model (GMM) with each Gaussian component using a diagonal covariance matrix, or a full covariance Gaussian model.

(i) Compare the two alternate forms of distribution in terms of the nature of data that they can model. Discuss how the generalisation of the classifier would be expected to vary as d increases for each of these distribution types. [20%]

(ii) Compare the computational cost of calculating the log likelihood for each of these distributions, stating any assumptions made. [15%]

(b) An M-component mixture model is to be estimated using n samples of single dimensional data $x_1, \ldots, x_n$. The probability density function associated with the m^{th} component, ω_m, is $p(x|\omega_m)$ and the component prior is c_m.

(i) Write down the log likelihood $\mathscr{L}(\theta)$ of the training data where θ is the parameter vector of the model. [10%]

(ii) Show that the partial derivative of $\mathscr{L}(\theta)$ with respect to a particular parameter associated only with ω_m, θ_m, is [10%]

$$\frac{\partial \mathscr{L}(\theta)}{\partial \theta_m} = \sum_{i=1}^{n} P(\omega_m|x_i) \frac{\partial \ln [p(x_i|\omega_m)c_m]}{\partial \theta_m}$$

(iii) If $p(x|\omega_m)$ is Gaussian, find the gradient of the log likelihood with respect to both the mean and the standard deviation. Hence state how the maximum likelihood estimates of the mean and variance parameters can be obtained using a *gradient descent* procedure. [25%]

(iv) Compare the use of gradient descent and *Expectation-Maximisation* (EM) for the maximum likelihood estimation of the parameters of a Gaussian mixture model. [20%]

2 A *Multi-Layer Perceptron* (MLP) is to be trained using error back-propagation using a least squares error criterion. For each input pattern, $\mathbf{x}$, the target vector is $t(\mathbf{x})$. The MLP has L layers, with $N^{(k)}$ units in layer k, and the input to the MLP is d-dimensional.

The nodes use a hyperbolic tangent activation function of the form

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

(a) Find the differential of the node activation function in terms of the output of the node. [15%]

(b) Give an advantage of using the hyperbolic tangent activation function over a sigmoid logistic regression function for the hidden nodes. [10%]

(c) Find the partial derivative of the output error with respect to a weight going from the final hidden layer to the output layer. [20%]

(d) Show how the partial derivative of the output error with respect to a weight going to the final hidden layer can be computed. [25%]

(e) A gradient descent scheme is to be used to update the weights of the network.

(i) What factors need to be considered in setting the learning rate? [10%]

(ii) What are the differences between *batch* updates, *sequential* weight updates, and the use of a *mini-batch*? What are the advantages and disadvantages of each method? Consider both the number of MLP weights and the training set size. [20%]

3 A classifier is to be built for a two class problem. There are n, d-dimensional training samples, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with class labels, $y_1, \ldots, y_n$. If x_i belongs to class C_1 then $y_i = 1$, and if it belongs to C_0, then $y_i = -1$.

(a) Assume the labelled training samples are generated under the distribution $P(X,Y)$ with density $p(\mathbf{x},y)$. Give an expression for the expected risk, $R(\theta)$, for a classifier $f(x,\theta)$. [10%]

(b) In most practical situations the true distribution $P(X,Y)$ is not known. Give an expression for the empirical risk $R_{emp}(\theta)$ and explain how it can be derived from the expected risk. [20%]

(c) Show that for this binary classification problem, the expected risk is the same as the probability of misclassification: $R(\theta) = P(f(\mathbf{x},\theta) \neq \mathbf{y})$. [10%]

(d) A Support Vector Machine (SVM) is to be built for this two class problem.

(i) Assuming the training data are linearly separable, what conditions must be satisfied by the trained SVM for all the data samples? [20%]

(ii) Show how *slack variables* can be introduced to extend this condition to the case of non-separable training samples. Assuming that the SVM has been estimated to have a margin of ± 1, explain how the slack variables can be used to obtain a simple upper bound on the number of training errors made by the SVM. [20%]

(iii) For the case of non-separable training data, give the training objective for the soft-margin classifier in its primal and dual forms. Explain how to determine the bias b. [20%]

4 A decision tree is to be built for a classification problem. Each training 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.

(a) Explain why a decision tree built from this data can be described as a *non-metric* classifier. Give an example of a problem in which the application of a non-metric classifier would be appropriate. [20%]

(b) The decision tree training process requires a *node impurity* function.

(i) List the desirable properties of a node impurity function. [20%]

(ii) Give the equation for the *entropy* purity function and explain how it satisfies the attributes described in (b)(i). Describe how the entropy purity function is calculated for a particular decision tree node. [20%]

(iii) Explain how the purity function is used to build a decision tree. [20%]

(c) A *regression tree* is to be built. The training data for this task consist of m, d-dimensional training observations $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Each training sample $\mathbf{x}_i$ is associated with a real output value, y_i , as well as the binary attributes described above. Each node in the regression tree is to have its own linear regression function. Explain how the regression tree can be built using the least squares cost function as the purity measure. Give equations for the linear regression function associated with each node. [20%]

5 The Parzen window density estimate $\tilde{p}(x)$ for the 1-dimensional vector x is

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

where the training data consists of samples x_1 to x_n. Assume that the window function $\phi(x)$ is a valid probability distribution.

(a) Show that $\tilde{p}(x)$ will be a valid probability density function. [15%]

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

(c) For a probability distribution with density $p(x)$, define the cumulative distribution function as $F_p(a) = \int_{-\infty}^{a} p(x)dx$. Show that $F_{\tilde{p}}(a) = \frac{1}{n}\sum_{i=1}^{n} F_\phi(\frac{a-x_i}{h})$. [20%]

(d) Suppose the training samples are drawn from a Gaussian distribution with mean μ and variance σ^2. Assume that the form of the window function is also Gaussian. Show that the expected value of the Parzen window density estimate is

$$\mathscr{E}\{\tilde{p}(x)\} = \mathscr{N}(x;\mu,\sigma^2+h^2)$$

Comment on the implication of this result. Note the following result may be useful:

$$\int_{-\infty}^{\infty}\mathscr{N}(x;\nu,\sigma_1^2)\,\mathscr{N}(\nu;\mu,\sigma_2^2)d\nu = \mathscr{N}(x;\mu,\sigma_1^2+\sigma_2^2)$$ [25%]

(e) Suppose the training samples are drawn from a distribution $P(X)$ with probability density function $p(x)$. A Parzen window density estimate is based on the uniform window function

$$\phi(x) = \begin{cases} \frac{1}{2} & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Show that $\mathscr{E}\{\tilde{p}(x)\} = \frac{1}{2h}(F_p(x+h) - F_p(x-h))$. [25%]

END OF PAPER

wjb04