ENGINEERING TRIPOS PART IIB

Tuesday 24 April 2012 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.*

There are no attachments.

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

mjfg04

1 An interesting family of probability distributions for one-dimensional data may be described by the following equation

$$p(x|\alpha) = \frac{1}{Z}\exp\left(\alpha'\mathbf{f}(x)\right)$$

where α is the vector of parameters associated with the distribution and $\mathbf{f}(x)$ is a function of the data point x that returns a vector of the same dimension as α.

(a) What expression must be satisfied by Z for this expression to be a valid probability density function? [10%]

(b) Show that if

$$\mathbf{f}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

then a univariate Gaussian distribution may be expressed in this form. Find expressions for Z and the elements of the vector α in terms of the mean, μ, and variance, σ^2, of the Gaussian distribution in this case. [25%]

(c) Rather than using a single distribution, a mixture of distributions is to be used based on the function $\mathbf{f}(x)$ in (b). To reduce the total number of parameters one of the elements of the parameter vector is constrained to be the same for all components. Thus

$$p(x|\alpha_1,\ldots,\alpha_M,\lambda) = \sum_{m=1}^{M} c_m \frac{1}{Z_m}\exp\left(\begin{bmatrix} \alpha_m & \lambda \end{bmatrix}\mathbf{f}(x)\right)$$

The parameters of the distribution, $\alpha_1,\ldots,\alpha_m,\lambda$, are to be trained on N independent samples of data, $x_1,\ldots,x_N$. The priors, c_1 to c_M, are known and not re-estimated. Maximum Likelihood (ML) training is used to estimate the model parameters.

(i) Derive an expression for Z_m in terms of α_m and λ. [15%]

(ii) The first element of the parameter vector for component m, α_m, is to be trained using Expectation Maximisation (EM). The following form of auxiliary function is to be used

$$Q(\alpha_1,\ldots,\alpha_M,\lambda,\hat{\alpha}_1,\ldots,\hat{\alpha}_M,\hat{\lambda}) = \sum_{m=1}^{M}\sum_{i=1}^{n} P(m|x_i,\alpha_1,\ldots,\alpha_M,\lambda)\log(p(x_i|m,\hat{\alpha}_m,\hat{\lambda}))$$

Derive an update formula for the value of $\hat{\alpha}_m$. [30%]

(iii) All of the model parameters (excluding the priors), $\hat{\alpha}_1, \ldots, \hat{\alpha}_m, \hat{\lambda}$, are now to be estimated. Derive an update formula for $\hat{\lambda}$. [20%]

2 For a two-class problem a Bayes' minimum error rate classifier is to be used. The class-conditional probability density functions (PDFs) are $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ and the prior probabilities are $P(\omega_1)$ and $P(\omega_2)$ for classes ω_1 and ω_2 respectively. The feature vector is d-dimensional.

(a) What is the general form of Bayes' decision rule for a two class problem? [10%]

(b) The classifier divides the feature-space into two regions. The data are classified as ω_1 in region $\mathscr{R}_1$ and ω_2 in region $\mathscr{R}_2$. Find an expression for the probability of error in terms of the class-conditional PDFs, the class priors and the regions $\mathscr{R}_1$ and $\mathscr{R}_2$. [15%]

(c) The two class-conditional PDFs are known to be multivariate Gaussians, both with an identity covariance matrix, $\mathbf{I}$. The mean for class ω_1 is μ_1 and for ω_2 is μ_2. The priors are known to be equal. Bayes' decision rule is used to design the classifier.

(i) Find an expression in terms of μ_1 and μ_2, that is satisfied by a point $\mathbf{x}$ that lies on the decision boundary. [25%]

(ii) By considering the distributions in the direction of the line joining the class means, show that the probability of error, P_e, can be expressed as

$$P_e = \int_a^\infty \mathcal{N}(v;0,1)dv$$

and find an expression for a. [30%]

(d) Why do practical classifiers of this form normally have an error rate higher than the value of the Bayes' minimum error classifier? [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 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 classifier has the form

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

where $\phi(\mathbf{x})$ is a transformation of $\mathbf{x}$ that yields a p-dimensional vector.

(a) The parameters of the classifier, $\mathbf{b}$, are to be trained by maximising the log-probability, $\mathscr{L}(\mathbf{b})$.

(i) Show that the log-probability of the training data may be expressed as

$$\mathscr{L}(\mathbf{b}) = \sum_{i=1}^{n} (y_i \log(P(\omega_1|\mathbf{x}_i, \mathbf{b})) + (1 - y_i) \log(1 - P(\omega_1|\mathbf{x}_i, \mathbf{b}))$$

What form of decision boundary will this type of classifier yield? [15%]

(ii) Derive an expression for the derivative of $\mathscr{L}(\mathbf{b})$ with respect to $\mathbf{b}$. How can this derivative be used to find the model parameters? [30%]

(b) The classifier is to be used to solve a problem where the training and test data from the two classes comprise the following sets of points.

$$\begin{array}{lll} \omega_1: & [\ 1, -1]' & [-1,\ 1]' \\ \omega_2: & [\ 1,\ 1]' & [-1, -1]' \end{array}$$

The prior probability of each of the observations is equal. The update formula derived in (a)(ii) is used to obtain $\mathbf{b}$.

(i) Initially a transformation of the form $\phi(\mathbf{x}) = \mathbf{x}$ is used ($p = d$). If the parameter vector $\mathbf{b}$ has the form, $\mathbf{b} = [\alpha, 0]'$, what value of α maximises the log-probability? Does this yield a reasonable classifier? [25%]

(ii) The transformation is modified to have the following form:

$$\phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}$$

Find a solution for $\mathbf{b}$. Comment on the performance of this classifier. [30%]

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

(a) 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%]

(b) Each of the weights, w_k, has a Gaussian prior with a mean of zero and variance σ_w^2. For the distribution of the target values, $p(\mathbf{y})$, where $\mathbf{y} = [y_1, \ldots, y_N]'$, show that the mean vector is zero and element i, j of the covariance matrix, Σ_y, can be written as

$$[\Sigma_y]_{ij} = \sigma_w^2 \sum_{k=1}^{H} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma^2}\right) \exp\left(-\frac{(x_j-\mu_k)^2}{2\sigma^2}\right) + \sigma_n^2 \delta(i-j)$$

where

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

[35%]

(c) The number of basis functions is increased so that $H \to \infty$. The positions of the centres of the basis functions, $\mu_1, \ldots, \mu_H$, are Gaussian distributed with a mean of zero and variance of σ_n^2, and $\sigma_w^2 \gg \sigma^2$. The variance of the weights' prior scales linearly with σ_n^2 and inversely with the number of basis functions, so $\sigma_w^2 = \sigma_n^2/H$. Discuss what impact the weights' prior will have in this situation and hence show that the covariance matrix of $p(\mathbf{y})$ can be approximated as a covariance function with elements of the form

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

What are the values of α, β, and γ? [35%]

(d) Discuss how the effective number of weight parameters varies as $H \to \infty$. [10%]

The following equality may be useful for this question

$$\int \mathcal{N}(x;\mu_1,\sigma^2)\mathcal{N}(x;\mu_2,\sigma^2)dx = \frac{1}{2\sqrt{\pi\sigma^2}}\exp\left(-\frac{1}{4\sigma^2}(\mu_1-\mu_2)^2\right)$$

5 A decision tree is to be built for a classification problem.

(a) As part of the training process for a decision tree a *node impurity* function is required.

(i) What general attributes should be satisfied by a node impurity function? How are node impurity functions used in building a decision tree? [20%]

(ii) For a K-class problem, the Gini impurity measure may be expressed in either of the two following ways:

$$\sum_{i \neq j} P(\omega_i)P(\omega_j); \quad \text{or} \quad 1 - \sum_{i=1}^{K} (P(\omega_i))^2$$

Describe how $P(\omega_i)$ should be calculated for a particular decision tree node. Show that these two expressions are equivalent and satisfy the attributes described in part (a)(i). [25%]

(b) The Gini impurity measure is to be used to train a decision tree for a two-class problem with 4-dimensional, binary valued, data. The training data for the two classes, ω_1 and ω_2, are shown below.

$$\begin{array}{lllll} \omega_1: & [0,0,0,0]' & [1,0,1,0]' & [1,1,0,0]' & [0,0,1,1]' \\ \omega_2: & [1,1,0,0]' & [1,1,1,1]' & [1,1,1,0]' & [0,1,1,1]' \end{array}$$

(i) Using changes in the Gini impurity measure, determine which of the first two elements of the feature-vector would be better used for the initial split. What is the change in the impurity measure in this case? [20%]

(ii) What is the lowest impurity measure that can be obtained using this data? [15%]

(c) For a particular task, the cost of misclassifying class ω_2 is twice that of class ω_1. How would this alter the decision tree training process? [20%]

END OF PAPER

mjfg04