ENGINEERING TRIPOS PART IIB

Tuesday 20 April 2010 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

MJFG05

1 A classifier is to be constructed for a two class problem. The data for each of the two classes is multivariate Gaussian distributed, with distributions $\mathcal{N}(\mu^{(1)}, \Sigma^{(1)})$ and $\mathcal{N}(\mu^{(2)}, \Sigma^{(2)})$ for classes ω_1 and ω_2 respectively. The values of the mean vectors and covariance matrices for the two classes are

$$\mu^{(1)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}; \quad \mu^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \quad \Sigma^{(1)} = \Sigma^{(2)} = \Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

A large number of training samples for each of the classes is available. The priors for the two classes are fixed and set to be equal.

(a) Derive the equation for any point $\mathbf{x}$ that lies on the optimal decision boundary for this problem based on Bayes' decision rule. [35%]

(b) A generative classifier is trained on the training samples. The class-conditional probability density functions of the classifier are multivariate Gaussian distributions where the covariance matrices are set to the identity matrix. The mean vectors are found using maximum likelihood estimation. Derive an expression for any point $\mathbf{x}$ that lies on the decision boundary for this classifier based on Bayes' decision rule. [15%]

(c) A non-linear transformation is applied to the training samples. A point in the original feature-space, $\mathbf{x}$, is related to the transformed feature-space, $\phi(\mathbf{x})$, by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \quad \phi(\mathbf{x}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_1 x_2 & x_2^2 \end{bmatrix}'$$

(i) The transformed data are used to train a classifier in the feature-space specified by $\phi(\mathbf{x})$. The same form of classifier and training criterion as part (b) are used. The covariance matrices for the two classes are again identity matrices, only the mean vectors are trained. Derive expressions for the mean vectors of the two classes in the transformed feature-space, and for any point, $\mathbf{x}$, in the *original feature-space* that lies on the resulting decision boundary. [30%]

(ii) A linear classifier is to be trained on the data in the transformed feature-space, $\phi(\mathbf{x})$. Find the parameters of the linear classifier that minimises the expected probability of error. [10%]

(iii) Under what conditions is it advantageous to build a linear classifier in the transformed feature-space, $\phi(\mathbf{x})$, rather than the original space, $\mathbf{x}$, if the data from the two classes are known to be multivariate Gaussian distributed? [10%]

2 The exponential family of probability distributions for d-dimensional data may be described by the following equation

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

where α is the vector of parameters associated with the distribution and $\mathbf{f}(\mathbf{x})$ is a function of the data point $\mathbf{x}$ 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 the distribution with parameters $\mu = [\ \mu_1\ \ \ldots\ \ \mu_d\ ]'$, $\mu_i > 0$,

$$p(\mathbf{x}|\mu) = \frac{1}{Z}\prod_{i=1}^{d}\mu_i^{x_i} \quad \text{where } \mathbf{x} = [\ x_1\ \ \ldots\ \ x_d\ ]' \text{ and } x_i \geq 0$$

is a member of the exponential family. Find expressions for α, $\mathbf{f}(\mathbf{x})$ and Z. It may be useful to use the univariate exponential distribution: $p(x) = \lambda\exp(-\lambda x)$. [30%]

(c) Rather than using a single distribution, a mixture of distributions in this family is to be used. This has the form

$$p(\mathbf{x}|\alpha) = \sum_{m=1}^{M} c_m \frac{1}{Z_m}\exp\left(\alpha_m'\mathbf{f}(\mathbf{x})\right)$$

The parameters of this distribution, $\alpha_1,\ldots,\alpha_M$, are to be trained on n independent samples of data, $\mathbf{x}_1,\ldots,\mathbf{x}_n$. 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) Write down an expression for the log-likelihood of the training data using this mixture distribution. [15%]

(ii) The parameters of the model, $\alpha_1,\ldots,\alpha_M$, are to be estimated using Expectation Maximisation (EM). The following form of auxiliary function is to be used

$$Q(\alpha,\hat{\alpha}) = \sum_{m=1}^{M}\sum_{i=1}^{n} P(m|\mathbf{x}_i,\alpha)\log(p(\mathbf{x}_i|m,\hat{\alpha}_m))$$

What statistics must be extracted from the training data to allow the model parameters to be estimated? [25%]

(iii) Discuss the maximisation of this auxiliary function. [20%]

3 Regression is to be performed using a Gaussian process. There are n, d-dimensional, training observations, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, with associated output values $\mathbf{y} = [y_1, \ldots, y_n]'$. The outputs are related to the observations by $y_i = f(\mathbf{x}_i) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The regression function, $f(\mathbf{x})$, is jointly Gaussian distributed with the training outputs. The mean function is set to 0. The covariance function between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is $k(\mathbf{x}_i, \mathbf{x}_j)$. An additional term is added to this covariance function for the prediction noise, ε.

(a) What is the advantage of using Gaussian process regression over basis function regression? [10%]

(b) By finding an expression for the joint distribution of $f(\mathbf{x})$ and the training data output values $\mathbf{y}$, show that the mean, μ, and variance, σ^2, of the distribution of the output for observation $\mathbf{x}$ have the form

$$\begin{aligned} \mu &= \mathbf{d}'\mathbf{E}^{-1}\mathbf{y} \\ \sigma^2 &= c - \mathbf{d}'\mathbf{E}^{-1}\mathbf{d} + \sigma_\varepsilon^2 \end{aligned}$$

Find expressions for the scalar c, vector $\mathbf{d}$ and matrix $\mathbf{E}$. [30%]

(c) Rather than using all n training examples, only the first $n-1$ are used. The variance of the output for observation $\mathbf{x}$ using all n examples is denoted as $\mathrm{var}_n(f(\mathbf{x}))$.

(i) By expressing the variance using all n examples as

$$\mathrm{var}_n(f(\mathbf{x})) = \mathrm{var}_{n-1}(f(\mathbf{x})) - b\,(k(\mathbf{x}, \mathbf{x}_n) - a)^2$$

or otherwise, show that

$$\mathrm{var}_n(f(\mathbf{x})) \leq \mathrm{var}_{n-1}(f(\mathbf{x}))$$

where $\mathrm{var}_{n-1}(f(\mathbf{x}))$ is the prediction using the first $n-1$ points. [50%]

(ii) Discuss the implication of this result. [10%]

The following matrix equality may be useful for this question. For a symmetric matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}' & c \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{b}\left(c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}\right)^{-1}\mathbf{b}'\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{b}\left(c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}\right)^{-1} \\ -\left(c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}\right)^{-1}\mathbf{b}'\mathbf{A}^{-1} & \left(c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}\right)^{-1} \end{bmatrix}$$

The following equality may be useful for this question. If $\mathbf{a}$ and $\mathbf{b}$ are jointly Gaussian,

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right)$$

then

$$\mathbf{a}|\mathbf{b} \sim \mathcal{N}(\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{b} - \mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$$

4 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 $\mathbf{w}^{(k)\prime}$ 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%]

5 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

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

(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%]

(iii) To improve the training of the model parameters the Hessian matrix is to be used. How is the Hessian matrix defined, and how can it be used to improve the training of the classifier? [10%]

(b) A regularisation term is added to the log-probability. The parameters are now estimated based on maximising the following function

$$\mathscr{F}(\mathbf{b}) = \mathscr{L}(\mathbf{b}) - \lambda \mathbf{b}'\mathbf{b}$$

where λ is a fixed scalar value.

(i) Discuss why this form of expression may yield an estimate of $\mathbf{b}$ that generalises better. How does the value of λ influence the estimate of $\mathbf{b}$? [15%]

(ii) Derive an expression for the derivative of $\mathscr{F}(\mathbf{b})$ with respect to $\mathbf{b}$. [15%]

(iii) How does the inclusion of the regularisation term change the Hessian? Discuss how the value of λ can be used to ensure that the Hessian is appropriate in general for finding the model parameters even when d becomes large. [15%]

END OF PAPER