ENGINEERING TRIPOS PART IIB

Tuesday 22 April 2008 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

1 A classifier is to be constructed using generative models and Bayes' decision rule. A d-dimensional observation feature-vector is to be used.

(a) Initially the classifier is to be used with $K > 2$ classes.

(i) State Bayes' decision rule and how it may be used with this classifier. [10%]

(ii) Discuss the differences between a classifier constructed using generative models and one using discriminative models. Under what conditions will the classifier using generative models yield a classifier with the minimum probability of error? [20%]

(b) The classes are now grouped together so that only a binary classifier ($K = 2$) is required. Using Bayes' decision rule, the classifier partitions the feature-space into two regions, Ω_1 where the observation is classified as belonging to class ω_1, and Ω_2 where the observation is classified as ω_2.

(i) Give an expression for the probability of error for this classifier in terms of the class-conditional probability distributions for the two classes, $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$, and the priors for the two classes, $P(\omega_1)$ and $P(\omega_2)$. [15%]

(ii) Show that an upper bound on the probability of error, $P(\text{error})$, is given by

$$P(\text{error}) \leq \int \sqrt{p(\mathbf{x}|\omega_1)P(\omega_1)p(\mathbf{x}|\omega_2)P(\omega_2)}d\mathbf{x}$$

Note that for two non-negative numbers a and b, if $a \leq b$ then $a \leq \sqrt{ab}$. [25%]

(iii) Gaussian class-conditional probability distributions are to be used. The mean vector for class ω_1 is μ_1 and for class ω_2 it is μ_2. The covariance matrices for the two classes are the same, Σ. The priors for the two classes are also the same. Find an expression for the probability of error in terms of only μ_1, μ_2, Σ and constant terms. This expression should *not* be a function of $\mathbf{x}$. The following equality may be useful: if $\mathbf{x}$ is a d-dimensional vector then

$$\int \exp\left(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x} + \mu'\Sigma^{-1}\mathbf{x}\right)d\mathbf{x} = (2\pi)^{d/2}|\Sigma|^{1/2}\exp\left(\frac{1}{2}\mu'\Sigma^{-1}\mu\right)$$

[30%]

2 A classifier is to be built for a two-class problem. There are n, d-dimensional, training vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, 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) What form of decision boundary can be obtained with this type of classifier? [10%]

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

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

[10%]

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

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

(ii) Show that the Hessian, $\mathbf{H}$, that may be used to train this classifier can be expressed as

$$\mathbf{H} = -\mathbf{S}'\mathbf{R}\mathbf{S}$$

Find expressions for the matrices $\mathbf{S}$ and $\mathbf{R}$. [25%]

(iii) Give an appropriate formula for finding the model parameters that involves the Hessian. What are the advantages and disadvantages of using this form of expression compared to the one in part (c)(i)? [15%]

(iv) Comment on the form of the Hessian matrix in part (c)(ii) and what it implies for this optimisation problem. [10%]

3 Regression is to be performed using either *basis functions* or 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 the prediction noise, ε, is Gaussian distributed, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

(a) For basis function regression, the regression function is of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i \phi(||\mathbf{x}_i - \mathbf{x}||)$$

(i) Derive an expression for the Maximum Likelihood (ML) estimate of the parameters of the regression function, $w_1, \ldots, w_n$, in terms of the training observations and output values. [30%]

(ii) Hence derive an expression for the distribution of the output y for the observation $\mathbf{x}$ using the ML regression function parameters. [10%]

(b) For Gaussian process regression, the regression function is jointly Gaussian distributed with the training outputs, $\mathbf{y}$. A squared exponential covariance function is to be used to which an additional term is added for the prediction noise ε. The mean function is set to 0.

(i) Give an expression for the squared exponential function between the observation $\mathbf{x}$ and training observation $\mathbf{x}_i$. Is this covariance function stationary? [15%]

(ii) By deriving an expression for the joint distribution of $f(\mathbf{x})$ and the output values $\mathbf{y}$, show that the mean, μ, and variance, σ^2, of the distribution of the output for observation $\mathbf{x}$ has 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) Compare these two forms of regression. You should discuss computational cost, storage and how accurate the regression process is. [15%]

(cont.

The following equality for vectors may be useful for this question. If **a** and **b** are jointly Gaussian,

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

then

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

4 An M-component mixture model is to be used as the probability distribution for a 1-dimensional binary value x. Each of the component distributions has the same form, a Bernoulli distribution. Thus for component m, ω_m, the distribution may be written as

$$p(x|\omega_m, \lambda_m) = \lambda_m^x (1-\lambda_m)^{(1-x)}$$

There are n independent training examples, $x_1, \ldots, x_n$, to estimate the model parameters. The component priors, $c_1, \ldots, c_M$, are known and fixed. The parameters of the model are to be estimated using Maximum Likelihood (ML) estimation.

(a) Find an expression for the log-likelihood of the training data in terms of the model parameters, $\lambda_1, \ldots, \lambda_M$. [15%]

(b) Expectation-Maximisation (EM) is to be used. The auxiliary function for this task may be written as (ignoring terms not involving $\lambda_1, \ldots, \lambda_M$)

$$Q(\lambda, \hat{\lambda}) = K + \sum_{i=1}^{n} \sum_{m=1}^{M} P(\omega_m|x_i, \lambda) \log(p(x_i|\omega_m, \hat{\lambda}_m))$$

where λ is the set of all model parameters, $\lambda_1, \ldots, \lambda_M$.

(i) Describe how EM is used to estimate the model parameters and the part played by the auxiliary function. Why is EM often used for mixture models? [15%]

(ii) Derive the update formula for finding the parameter estimates. [30%]

(c) The exponential family is an important class of probability distributions. Rather than using a mixture of Bernoulli distributions, a mixture of members of the exponential family is to be used. For component m the distribution is now

$$p(x|\omega_m, \alpha_m) = \frac{1}{Z_m} \exp(\alpha_m' \mathbf{f}(x))$$

where Z_m is a normalisation term, $\mathbf{f}(x)$ and α_m are vectors, possibly one-dimensional.

(i) Show that the Bernoulli distribution is a member of the exponential family and find expressions for Z_m and α_m. [20%]

(ii) Derive an expression for the auxiliary function in terms of the parameters of the exponential distributions, $\alpha_1, \ldots, \alpha_M$, and the associated normalisation terms, $Z_1, \ldots, Z_M$. Comment on using EM with members of the exponential family. [20%]

5 Speaker Verification is to be performed using Support Vector Machines (SVMs). A d-dimensional feature vector extracted every 10 milli-seconds is used to parameterise the speech data. A separate SVM is trained for each speaker.

(a) The system is to use a *Fisher kernel* based on an M-component Gaussian mixture model (GMM) that is Maximum A-Posteriori (MAP) adapted to the enrolment data of the speaker. For this system the following form of the Fisher kernel is used

$$k(\mathbf{O}^{(m)}, \mathbf{O}^{(n)}) = \left[\nabla_\mu \log(p(\mathbf{O}^{(m)}|\theta))\right]' \left[\nabla_\mu \log(p(\mathbf{O}^{(n)}|\theta))\right]$$

where $\mathbf{O}^{(m)}$ and $\mathbf{O}^{(n)}$ are two different sequences and $\log(p(\mathbf{O}^{(m)}|\theta)$ is the log-likelihood of sequence $\mathbf{O}^{(m)}$ using the MAP-adapted GMM. ∇_μ indicates that derivatives with respect to all the component mean vectors are used.

(i) Briefly describe how an SVM with the Fisher kernel can be used in a speaker verification task. [20%]

(ii) By writing the expression for the log-likelihood of observation sequence $\mathbf{O}^{(m)}$, derive an expression for the feature-space associated with the Fisher kernel in terms of the component means and variances. What is the dimensionality of the feature-space? [30%]

(b) The Fisher kernel is replaced by a *sequence kernel* which has the form

$$k(\mathbf{O}^{(m)}, \mathbf{O}^{(n)}) = \sum_{i=1}^{T^{(m)}} \sum_{j=1}^{T^{(n)}} k^{\mathrm{s}}(\mathbf{o}_i^{(m)}, \mathbf{o}_j^{(n)})$$

where $\mathbf{o}_i^{(m)}$ is the i^{th} vector in the sequence $\mathbf{O}^{(m)}$. $k^{\mathrm{s}}(.,.)$ is either a linear kernel, or a Gaussian kernel with width σ.

(i) Under what conditions will the sequence kernel with a linear kernel yield the same classifier as a Fisher kernel? [25%]

(ii) Give the form of the Gaussian kernel. Compare the sequence kernel using this Gaussian kernel and the Fisher kernel for speaker verification. You should discuss the computational cost and strengths and weaknesses of the two forms. [25%]

END OF PAPER