ENGINEERING TRIPOS PART IIB

Tuesday 21 April 2009 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 generative classifier is to be built for a two-class problem. The observation vectors for this task are d-dimensional. The class conditional distributions for the two classes are Gaussian. The parameters for class ω_1 are μ_1 and Σ_1 and those for class ω_2 are μ_2 and Σ_2. The priors for the two classes are $P(\omega_1)$ and $P(\omega_2)$.

(a) State Bayes' decision rule for this task. Under what conditions will this form of generative classifier yield a classifier with the minimum probability of error? [15%]

(b) The covariance matrices for the two classes are constrained to be the same and diagonal, $\Sigma_1 = \Sigma_2 = \Sigma$, and the priors for the two classes are equal.

(i) The model parameters are trained on n training observations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, 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 log-likelihood of the training data can be expressed as

$$\mathscr{L}(\mu_1, \mu_2, \Sigma) = \sum_{i=1}^{n} (y_i \log(\mathscr{N}(\mathbf{x}_i; \mu_1, \Sigma)) + (1 - y_i) \log(\mathscr{N}(\mathbf{x}_i; \mu_2, \Sigma)))$$

What is the maximum-likelihood estimate of the covariance matrix Σ? [20%]

(ii) Show that the posterior probability for class ω_1 given an observation $\mathbf{x}$ can be expressed in the form

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

where $\phi(\mathbf{x})$ is a function of $\mathbf{x}$ that yields a d-dimensional vector. Find expressions for $\mathbf{w}$, $\phi(\mathbf{x})$ and a. [25%]

(c) The means of the two class-conditional distributions are now constrained to be zero, $\mu_1 = \mu_2 = \mathbf{0}$. The covariance matrices for the two classes are allowed to be different, but constrained to be diagonal. Again the priors for the two classes are equal. Show that the posterior probability for class ω_1 can be expressed in the same form as in part (b)(ii) and find expressions for $\mathbf{w}$, $\phi(\mathbf{x})$ and a in this case. [25%]

(d) Compare the decision boundaries that result from the two forms of classifier in parts (b) and (c). [15%]

2 An M-component Gaussian mixture model with diagonal covariance matrices is to be used as the probability distribution for a d-dimensional feature vector. There are N independent training examples, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, to estimate the model parameters. 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 component priors, $c_1, \ldots, c_M$, and component parameters. [10%]

(b) Expectation-Maximisation (EM) is to be used to find the Gaussian component means. The auxiliary function for this problem can be expressed as

$$Q(\theta, \hat{\theta}) = \sum_{i=1}^{N} \sum_{m=1}^{M} P(\omega_m|\mathbf{x}_i, \theta) \log(p(\mathbf{x}_i|\omega_m, \hat{\theta}))$$

where θ is the set of all the model parameters and constant terms have been ignored.

(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) Show that the update formula for the mean of component ω_m is [30%]

$$\hat{\mu}_m = \frac{\sum_{i=1}^{N} P(\omega_m|\mathbf{x}_i, \theta)\mathbf{x}_i}{\sum_{i=1}^{N} P(\omega_m|\mathbf{x}_i, \theta)}$$

(c) A sequential form of update is to be used to estimate the means. The update formula for the estimate of the mean after n training examples, $\hat{\mu}_m^{(n)}$, can be expressed as

$$\hat{\mu}_m^{(n)} = \hat{\mu}_m^{(n-1)} + \eta_m^{(n)}(\mathbf{x}_n - \hat{\mu}_m^{(n-1)})$$

(i) Initially, the set of model parameters, θ, used to compute $P(\omega_m|\mathbf{x}_i, \theta)$ is not sequentially updated. Derive an expression for $\eta_m^{(n)}$ so that after all N training examples have been seen the estimate in part (b)(ii) is obtained. [30%]

(ii) The following approximate form for $\eta_m^{(n)}$ is proposed

$$\eta_m^{(n)} = \frac{P(\omega_m|\mathbf{x}_n, \theta)}{nc_m}$$

Why is this form more suitable when θ is sequentially updated? [15%]

(TURN OVER

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) Find an expression for the marginal likelihood of the training data, $p(\mathbf{y}|\mathbf{X})$. [15%]

(d) The following form of covariance function is to be used

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\sum_{m=1}^{d}\frac{(x_{im} - x_{jm})^2}{\sigma_m^2}\right)$$

where x_{im} is element m of vector $\mathbf{x}_i$. The hyper-parameters of the covariance function, $\sigma_1^2, \ldots, \sigma_d^2$, are to be estimated from the training data.

(i) Briefly discuss how the hyper-parameters can be estimated. [10%]

(ii) Show that as $\sigma_m^2 \to \infty$ the m'th dimension of the observation vector $\mathbf{x}$ does not influence the distribution of the output. [20%]

(iii) Compare Gaussian process regression with this form of covariance function and relevance vector machine regression. [15%]

(cont.

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

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

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 Parzen window is to be used to estimate the class-conditional density for a pattern classification task. The form of the Parzen window density estimate $\tilde{p}(x)$ for the the 1-dimensional vector x is given by

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

where the training data consists of training samples x_1 to x_n. The true distribution, from which the training samples are drawn, is Gaussian distributed with mean μ and variance σ^2. The form of the window function is also Gaussian.

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

(b) Show that the expected value of the Parzen window density estimate is

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

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

$$\int_{-\infty}^{\infty}\mathcal{N}(x;v,\sigma_1^2)\mathcal{N}(v;\mu,\sigma_2^2)dv = \mathcal{N}(x;\mu,\sigma_1^2+\sigma_2^2)$$

[30%]

(c) An approximate form for the true Parzen window density estimate is required.

(i) By using a first-order Taylor series expansion 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 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. What affects how well this approximates the exact Parzen window density estimate? [15%]

(iii) What is the expected value of this approximation? Contrast this expected value with the form given in part (b). [15%]

5 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 an SVM classifier. What is the general form for an inhomogeneous polynomial kernel-function? [15%]

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

$$\Phi(x) = \begin{bmatrix} 1 & \exp(x) & \exp(2x) & \ldots & \exp(Nx) \end{bmatrix}'$$

where x is the point in the input-space.

(i) Compare this form of feature-space with the feature-space associated with an inhomogeneous polynomial kernel. [20%]

(ii) 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

$$k(x_i, x_j) = \frac{b - \exp(a(x_i + x_j))}{b - \exp(x_i + x_j)}$$

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

(d) Express the SVM 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, the number of training samples, m, and N change? [15%]

END OF PAPER