Version MJFG/3

EGT3
ENGINEERING TRIPOS PART IIB

Tuesday 22 April 2014 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.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

1 A classifier is to be constructed using generative models and Bayes' decision rule for a binary classification problem. A d-dimensional observation feature-vector is used. The true data value for class ω_1 is $\mathbf{a}$, and the true data value for class ω_2 is $\mathbf{b}$. Because of the way that the values of the data are measured, the data has multivariate Gaussian distributed noise added onto the true sample values. The covariance matrix for the noise, Σ, is known to be an identity matrix and the mean of the noise is zero. The priors for the two classes are known to be equal. For each measurement the noise is independent of all other noise samples.

(a) State Bayes' decision rule for generative classifiers for binary classification tasks. [10%]

(b) For this task derive an expression for a point $\mathbf{x}$ that lies on the decision boundary that yields the minimum probability of error. [20%]

(c) Show that the minimum probability of error for this classifier, P_{e}, can be expressed in the form:

$$P_{\mathsf{e}} = \int_{-\infty}^{c} \mathcal{N}(z;0,1)dz$$

What is the value of c? [35%]

(d) To improve the accuracy of the classifier, measurements for each test sample are repeated five times. The noise for each of these measurements is independent.

(i) If the values from each of the five measurements are averaged, derive an expression for the minimum probability of error in this case. [20%]

(ii) An alternative approach is proposed, where each of the five repeated test samples is classified and the most common classification outcome used as the result. Comment on the probability of error for this scheme compared to averaging the values. [15%]

2 An M-component mixture model is to be used as the probability distribution for a 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) A second binary variable, z, is now observed, so that there are n observation pairs $\{x_1, z_1\}, \ldots, \{x_n, z_n\}$. The joint distribution for x and z for component m is given by

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

α is thus shared over all components. EM is again to be used to estimate an M-component mixture model.

(i) Discuss how $P(\omega_m|x_i, z_i, \lambda, \alpha)$ changes compared to $P(\omega_m|x_i, \lambda)$ in part (b). [15%]

(ii) Derive the update formula for α. [25%]

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

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

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

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

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

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? [25%]

(b) In order to improve the performance of the optimisation scheme, a second-order approximation is used. This has the form

$$E(\mathbf{b}) \approx E(\mathbf{b}^{(\tau)}) + (\mathbf{b}-\mathbf{b}^{(\tau)})'\mathbf{f} + \frac{1}{2}(\mathbf{b}-\mathbf{b}^{(\tau)})'\mathbf{A}(\mathbf{b}-\mathbf{b}^{(\tau)})$$

where $E(\mathbf{b})$ is the *error* for parameter $\mathbf{b}$, and $\mathbf{b}^{(\tau)}$ is the estimate of $\mathbf{b}$ at iteration τ.

(i) Derive an expression for the value of $\mathbf{b}$ that will minimise this second-order approximation. [20%]

(ii) Discuss how this quadratic form can be used to obtain the estimate of $\mathbf{b}$ in part (a). [10%]

(iii) By considering a second-order Taylor series expansion about the point $\mathbf{b}^{(\tau)}$ find expressions for $\mathbf{f}$ and $\mathbf{A}$ in terms of $\mathbf{b}^{(\tau)}$. [30%]

4 A product of experts system is to be used for speech synthesis. For a particular segment of speech data, the experts are Gaussian, with the form

$$\begin{aligned} p(x_t) &= \mathcal{N}(x_t;\mu_1,1) \\ p(x_t - x_{t-1}) &= \mathcal{N}(x_t - x_{t-1};\mu_2,1) \end{aligned}$$

The data sequence to be generated, $\mathbf{x}$, is known to be of length 3, $\mathbf{x} = [x_1, x_2, x_3]'$. The data sequence is also known to start in silence, which has a value of 0, thus $x_0 = 0$. In order to use the experts to generate the observation sequence a transformation $\mathbf{A}$ is introduced that doubles the length of the vector $\mathbf{x}$, so that each dimension of the transformed vector $\mathbf{Ax}$ can be related to one of the two experts. The transformed data, $\mathbf{Ax}$, is Gaussian distributed, so

$$p(\mathbf{x}) = \frac{1}{Z}\mathcal{N}(\mathbf{Ax};\mu,\Sigma) = \mathcal{N}(\mathbf{x};\overline{\mu},\overline{\Sigma})$$

where Z is the appropriate normalisation term to ensure a valid PDF.

(a) Derive a suitable form for the matrix $\mathbf{A}$. How will the form of this matrix vary as the length of the sequence changes? [20%]

(b) What expression must be satisfied by Z? [10%]

(c) Find expressions for μ and Σ in terms of the parameters of the experts. [20%]

(d) By using the following expression (or otherwise)

$$\exp\left(-\frac{1}{2}(\mathbf{Ax}-\mu)'(\mathbf{Ax}-\mu)\right) = \exp\left(-\frac{1}{2}\left(\mathbf{x}'\mathbf{A}'\mathbf{Ax} - 2\mu'\mathbf{Ax} + \mu'\mu\right)\right)$$

find an expression for the mean vector, $\overline{\mu}$, and covariance matrix, $\overline{\Sigma}$, of the distribution of $\mathbf{x}$ in terms of $\mathbf{A}$ and the parameters of the experts. Hence show that the normalisation term Z has the form

$$Z = \frac{|\overline{\Sigma}|^{1/2}}{(2\pi)^{3/2}}\exp\left(-\frac{1}{2}\mu'\mu + \frac{1}{2}\overline{\mu}'\overline{\Sigma}^{-1}\overline{\mu}\right)$$

[35%]

(e) Two approaches can be adopted for maximum likelihood training of the experts. The first is to train each expert separately. The second is to the train all experts together by maximising the likelihood of the product of experts. Contrast these two training approaches. [15%]

5 A Parzen window is to be used to estimate the class-conditional density for a pattern classification task.

(a) Contrast the use of a Parzen window density estimate with using a single multivariate Gaussian distribution as the class-conditional density. You should comment on memory requirements, computational cost and factors that will affect the performance. [25%]

(b) The form of the Parzen window density estimate $\tilde{p}(\mathbf{x})$ for the the d-dimensional vector $\mathbf{x}$ is given by

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

where the training data consists of training samples $\mathbf{x}_1$ to $\mathbf{x}_n$.

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

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

(c) For a particular application the data is one dimensional, $d = 1$, and the form of the window function is a Gaussian.

(i) By using a first order Taylor series expansion based 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 only functions of the training data. What are the values of b_0, b_1 and b_2? [30%]

(ii) Discuss how the use of this approximation affects the memory requirements and computational speed of using the Parzen window density estimate. [15%]

END OF PAPER