Version MJFG/4

EGT3
ENGINEERING TRIPOS PART IIB

Monday 20 April 2015 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

10 minutes reading time is allowed for this paper.

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 for a K class problem 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 equal to Σ. The priors for the two classes are also equal. Find an expression for the bound on the probability of error in terms of only μ_1, μ_2, Σ and constant terms. This expression should *not* be a function of $\mathbf{x}$. [30%]
The following equality may be useful: if $\mathbf{x}$ is a d-dimensional vector then

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

2 (a) Discuss the differences between using mixture of experts and product of experts for combining information from multiple experts. You should assume each expert has the form $p(\mathbf{x}|\theta_m)$ for an observation $\mathbf{x}$. You should include the forms of the two approaches in your discussion. [20%]

(b) Two Gaussian experts with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 are combined within a product of experts framework. Show that the resulting distribution is Gaussian with a mean vector, μ, and covariance matrix, Σ, which are given by

$$\mu = \Sigma\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right);\ \ \Sigma = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$$

[25%]

(c) A set of sensors are designed to measure a 3-dimensional signal, $\mathbf{x} = [\ x_1 \quad x_2 \quad x_3\ ]'$. The sensors yield the following information:

$$\begin{aligned} p(x_1) &= \mathcal{N}(x_1;\mu_a,1) \\ p(x_2-x_1) &= \mathcal{N}(x_2-x_1;0,1) \\ p(x_2+x_3) &= \mathcal{N}(x_2+x_3;\mu_c,1) \end{aligned}$$

The information from the sensors is to be combined in a product of experts framework. In order to combine the experts a transformation, $\mathbf{A}$, is introduced so that each dimension of the transformed vector $\mathbf{Ax}$ can be related to one of the three 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.

(i) Derive a suitable form for the matrix $\mathbf{A}$. [15%]

(ii) Show that the mean and covariance matrix of the signal $\mathbf{x}$ can be expressed in the form

$$\overline{\Sigma} = (\mathbf{A}'\mathbf{A})^{-1};\ \ \overline{\mu} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{v}$$

Derive expressions for $\mathbf{A}$ and $\mathbf{v}$. What is the value of the inverse covariance matrix, $\overline{\Sigma}^{-1}$? [30%]

(iii) Comment on the form of the inverse covariance matrix, and what it implies about the data. [10%]

3 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 the 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 $\hat{\theta}$ the parameters to be estimated. Constant terms have been ignored in this expression.

(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

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

[30%]

(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%]

4 A multi-layer perceptron (MLP) is to be used for a multi-class classification problem. There are K classes, $\omega_1, \ldots, \omega_K$. The input to the network is a d-dimensional feature vector and there are n examples to train the MLP, $\mathbf{x}_1, \ldots, \mathbf{x}_n$. For each of the training samples there is a target vector, $\mathbf{t}_1, \ldots, \mathbf{t}_n$, which uses a *1-out-of-K* coding for the class of each of the training examples.

(a) What needs to be considered when designing the neural network? Your answer should include a discussion of the the number of MLP parameters. [20%]

(b) The MLP parameters are to be trained using cross-entropy. This cost function has the form

$$E = -\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\log(b_{ij})$$

(i) For this expression describe what the variables I, J, a_{ij} and b_{ij} represent. [10%]

(ii) The activation function for the output layer is a *softmax* function. Write down the form of this activation function and an expression for the derivative of E with respect to the parameters of the output-layer. [25%]

(c) The Hessian matrix is to be used in the optimisation process for the model parameters in (b)(ii).

(i) How is the Hessian matrix defined, and how can it be used to improve the training of the MLP? [15%]

(ii) Derive an expression for the elements of the Hessian matrix for the parameters of the output layer. [20%]

(iii) Discuss the computational cost of using the Hessian in optimising the output-layer weights. [10%]

5 A Support Vector Machine (SVM) classifier is to be built for a two class problem. There are a total of m, d-dimensional, training samples $\mathbf{x}_1$ to $\mathbf{x}_m$ with associated labels y_1 to y_m where $y_i \in \{-1, 1\}$.

(a) The decision boundary for SVMs with a linear kernel has the form

$$\mathbf{w}'\mathbf{x} + b = 0$$

How does this expression change if a general kernel function $k(\mathbf{x}_i, \mathbf{x})$ is to be used? Why are SVMs suited to classification with kernel functions? [20%]

(b) A polynomial kernel can be written in the following forms

$$k(\mathbf{x}_i, \mathbf{x}) = \left(\mathbf{x}_i'\mathbf{x} + a\right)^c = \sum_{i=0}^{c} \begin{pmatrix} c \\ i \end{pmatrix} a^{c-i} (\mathbf{x}_i'\mathbf{x})^i$$

(i) How do the parameters a and c impact the form of the decision boundary? [15%]

(ii) Show that if a kernel can be expressed as the sum of kernels then the decision boundary can also be expressed as a sum of kernels. [10%]

(iii) Show that if $d = 2$, $c = 3$ and $a = 0$, then the feature-space associated with the polynomial kernel, $\Phi(\mathbf{x})$, may be written in the form

$$\Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^3 & \sqrt{3}x_1^2 x_2 & \sqrt{3}x_1 x_2^2 & x_2^3 \end{bmatrix}'$$

[20%]

(iv) Show that the dimensionality of the feature-space for the polynomial kernel when $d = 2$ and $a = 1$ is

$$\begin{pmatrix} 2+c \\ c \end{pmatrix}$$

How does this expression change as d increases? [35%]

END OF PAPER