MATH485001

This question paper consists
of 4 printed pages, each
of which is identified by the
reference **MATH485001**.

New Cambridge Elementary
Statistical Tables are provided.
Only approved basic scientific
calculators may be used.

Examination for the Module MATH4850
(May/June 2003)

MACHINE LEARNING, NEURAL NETWORKS AND STATISTICS

Time allowed: **3 hours**

Do not attempt more than four questions.

All questions carry equal marks.

1. Let π_l denote the proportion of a population which belongs to class l . Let $G(X)$ denote the class label of a random element of the population, with random variable, say X, such that $P(G = l) = \pi_l$, and let $\hat{G}(X \mid X = x)$ denote the predicted class of an observation x.

(a) Give an expression for the expected error rate with a brief justification.

If $L(k, l)$ denotes the loss incurred by making decision $\hat{G} = l$ when the true class is $G = k$ obtain an equivalent expression for the expected loss.

Verify that if all misclassifications are equally serious and if no cost is associated with the correct classification, then the expected loss simplifies to the expected error rate.

(b) Show that the classification rule which minimizes the expected error rate is given by

$$\arg\max P(G(X) = l \mid X = x).$$

(c) Suppose there are two groups in which $\pi_1 = \pi_2 = 1/2, L(1, 2) = 3L(2, 1)$ and that the two groups have class conditional distributions given by

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-2)^2}{2}\right\}$$

$$f_2(x) = \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{(x+2)^2}{8}\right\}.$$

Derive the posterior probability that an observation $X = x$ belongs to group 1 and obtain Bayes' rule to minimize the expected loss when the loss matrix has elements $L(1, 1) = L(2, 2) = 0, L(1, 2) = 1, L(2, 1) = 3$.

2. (a) Given data $(x_i, Y_i), i = 1, \ldots, n$ in which $x_i \in \mathbb{R}^p$ is a feature vector and $Y_i \in \{1, \ldots, K\}$ is a corresponding class label define the k-nearest neighbour classifier of an observation x_0, and briefly indicate how k might be chosen in practice.

(b) Let $p_l(x)$ denote the probability that an observation $X = x$ belongs to class l, and suppose that, at a point x, the true class is l^*. Explain why the error rate of the 1-nearest neighbour classifier, given $X = x$, is approximately

$$\sum_{l=1}^{K} p_l(x) \left(1 - p_l(x)\right).$$

Obtain an expression for the Bayes' error.

(c) Prove that

$$\sum_{l=1}^{K} p_l(x) \left(1 - p_l(x)\right) \le 2 \left(1 - p_{l^*}(x)\right) - \frac{K}{K-1} \left(1 - p_{l^*}(x)\right)^2$$

and explain why this is a useful result.

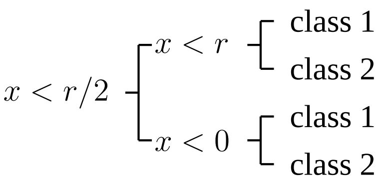
(d) Outline the 1-nearest neighbour **condense** algorithm and k-nearest neighbour **multiedit** algorithm and briefly explain how they work and when they could be used.

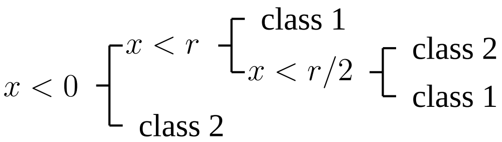
3. (a) Describe the process that obtains a decision tree using *specific-to-general* rules. Define any terminology or notation that you use.

(b) Suppose that data are drawn from two populations with equal priors. The first has a standard exponential distribution ($f_1(x) = e^{-x}$, for $x \geq 0$, and 0 otherwise) and the second has density

$$f_2(x) = \begin{cases} e^{-(r-x)} & \text{if } -\infty < x \leq r \\ 0 & \text{otherwise} \end{cases} \qquad \text{with } r \geq 0$$

Consider the two trees shown below – the lower (left) branch is taken if the condition is true.





(i) Plot the allocation regions, and work out the expected misclassification rate (in terms of r) for each tree. Determine the value of r which maximizes this rate. Show that the maximum error rate is $1/4$.

(ii) Given that $r = 4$, calculate the expected number of splits for each tree and state which tree is therefore more computationally efficient.

(iii) Calculate the value of r for which the two trees are equivalent in efficiency.

4. (a) Two measures which can be used as a splitting criterion for a decision tree are deviance, and the χ^2 test statistic.

Give formulae which can be used for calculating these quantities, and show that they are approximately equal.

Obtain both of these measures for the following split:

	Class		
	A	B	C
cond 1	50	10	0
cond 2	10	10	20

(b) (i) Define what is meant by pre-pruning and post-pruning of classification trees. What are their advantages and disadvantages?

(ii) Quinlan's pessimistic pruning method at a node estimates $\hat{p}$ (given α), the probability of misclassification at a node, using the relationship:

$$P(X \leq n) = \sum_{r=0}^{n} \binom{N}{r} \hat{p}^r (1-\hat{p})^{N-r} = \alpha$$

where there are N observations at the node, and n are misclassified.
Explain the reasoning behind this method.

(iii) Using $\alpha = 0.25$ determine whether the following split would be pruned using the above method.

	Class		
	A	B	C
cond 1	8	2	0
cond 2	0	0	2

5. (a) Given a set of points $(x_i, Y_i), i = 1, \ldots, n$ in which $x_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$ denotes the class label define what is meant by saying that the observations belonging to the two classes are *linearly separable*.

If we define the convex hull of the points belonging to class $j = -1, 1$ by

$$H_j = \{x : x = \sum_{i:Y_i=j} \alpha_i x_i, \text{ for some } \alpha \text{ such that } \alpha_i \geq 0, \sum_{i:Y_i=j} \alpha_i = 1\}$$

then show that the convex hulls of class -1 and class 1 do not intersect if and only if the points are linearly separable.

(b) Given data $x_i \in \mathbb{R}^p$ and corresponding $Y_i \in \{1, \ldots, K\}$ for $i = 1, \ldots, n$ describe fully a Radial basis function network, defining any notation you use.

State the parameters that need to be estimated, and give procedures for learning this type of network.

END