MATH5850M01

This question paper consists
of 7 printed pages, each
of which is identified by the
reference **MATH5850M01**.

Statistical Tables are attached.
Only approved basic scientific
calculators may be used.

Examination for the Module MATH5850M
(January 2005)

MACHINE LEARNING, NEURAL NETWORKS AND STATISTICS

Time allowed: **3 hours**

Do not attempt more than four questions.

All questions carry equal marks.

1. (a) Suppose $p_k(\boldsymbol{x})$ is the probability density function of observations from class k and π_k denotes the proportion of the population belonging to class $k, k = 1, \ldots, K$. If all misclassifications are equally serious, then write down an expression for the classification rule which minimizes the expected error rate.

(b) Suppose a discrete random variable X which has sample space given by $\Omega_X = \{1, 2\}$ with $\mathsf{P}(X = 1) = 0.58$ with observations belonging to one of two classes, with $\pi_A = 0.6\ (= 1 - \pi_B)$. Given the class conditional distribution p_A is given by

	x	
	1	2
$\mathsf{P}_A(X = x)$	0.7	0.3

obtain the distribution p_B. Hence, or otherwise, obtain Bayes' rule for this situation, and show that the expected error rate is 0.34.

(c) Suppose now, more generally, that $L(j, l)$ denotes the loss incurred by making decision l, when the true class is j, $(1 \leq j, l \leq K)$. Describe two real-life applications in which misclassification errors are not equally serious.

(d) For the above discrete random variable X, suppose that the loss matrix is given by

		true class	
		A	B
predicted	A	0	1
class	B	c	0

with $c \geq 0$. Obtain the optimal (i.e. to minimize the expected loss) misclassification rule in this case, as well as the possible values of c for which the misclassification rule is equivalent to that obtained in part (b).

(e) Obtain an expression for the expected loss for general c and verify that substituting $c = 1$ in your answer gives the value obtained in part (b).

2. (a) Given data $(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n$ in which $\boldsymbol{x}_i \in \mathbb{R}^p$ 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 $\boldsymbol{x}_0$, and describe how k might be chosen in practice.

(b) Let $\mathsf{P}(l \mid \boldsymbol{X} = \boldsymbol{x})$ denote the probability that an observation $X = x$ belongs to class l. Explain why the error rate of the 1-nearest neighbour classifier, given $\boldsymbol{X} = \boldsymbol{x}$, is approximately

$$\sum_{l=1}^{K} \mathsf{P}(l \mid \boldsymbol{x})\,(1 - \mathsf{P}(l \mid \boldsymbol{x}))\,.$$

Write an expression for the minimum expected error rate.

(c) Suppose that, at a point $\boldsymbol{x}$, the class which minimizes the expected error rate is l^*. Prove that,

$$\sum_{l=1}^{K} \mathsf{P}(l \mid \boldsymbol{x})\,(1 - \mathsf{P}(l \mid \boldsymbol{x})) \leq 2\,(1 - \mathsf{P}(l^* \mid \boldsymbol{x})) - \frac{K}{K-1}\,(1 - \mathsf{P}(l^* \mid \boldsymbol{x}))^2$$

and explain why this is a useful result.

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

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

(b) The following tree gives the loss as the number of misclassified observations at each node:

```
n= 2000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 2000 780 0 (0.61000000 0.39000000)
   2) V53< 0.0395 1461 314 0 (0.78507871 0.21492129)
     4) V7< 0.065 1327 192 0 (0.85531274 0.14468726)
       8) V52< 0.251 1109  84 0 (0.92425609 0.07574391) *
       9) V52>=0.251 218 108 0 (0.50458716 0.49541284) *
     5) V7>=0.065 134  12 1 (0.08955224 0.91044776)
      10) V27>=0.14 8   0 0 (1.00000000 0.00000000) *
      11) V27< 0.14 126   4 1 (0.03174603 0.96825397) *
   3) V53>=0.0395 539  73 1 (0.13543599 0.86456401)
     6) V25>=0.385 41   5 0 (0.87804878 0.12195122) *
     7) V25< 0.385 498  37 1 (0.07429719 0.92570281) *
```

Plot the tree in the usual way, including the frequency of observations in each of the two classes at each node, as well as the predicted class at each leaf.

Given a new observation in which a subset of the variables is given by ($V1 = 0, V2 = 0.7, V7 = 0, V24 = 10, V25 = 0.1, V26 = 0.2, V27 = 0.1, V52 = 0.3, V53 = 0$), estimate the probability that this observation belongs to each class.

(c) *"Cost-complexity pruning selects a rooted subtree T of the full tree T_0 to minimize*

$$R_\alpha(T) = R(T) + \alpha \, \text{size}(T)$$

for a given α"

In the above statement, what is $R(T)$? Explain what is meant by a *rooted subtree*, and give an interpretation of the formula, which includes a comment on the role of α.

Take the number of misclassified observations as a measure of impurity for the tree. Using the pruning algorithm based on cost-complexity we take $R(T_t)$ as the number misclassified by a tree with root node at t and $R(t)$ the number misclassified at node t. By calculating

$$g(t) = \frac{R(t) - R(T_t)}{\text{size}(T_t) - \text{size}(t)}$$

for each internal node, obtain the optimal pruned trees of size 5, 4 and 3, as well as the corresponding values of $\alpha_j, j = 5, 4, 3$, which denote the values of α at which the solutions to $R_\alpha(T)$ change.

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

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

where there are N observations at the node, and n are misclassified.

Explain how the above formula is used, and the reasoning behind this method.

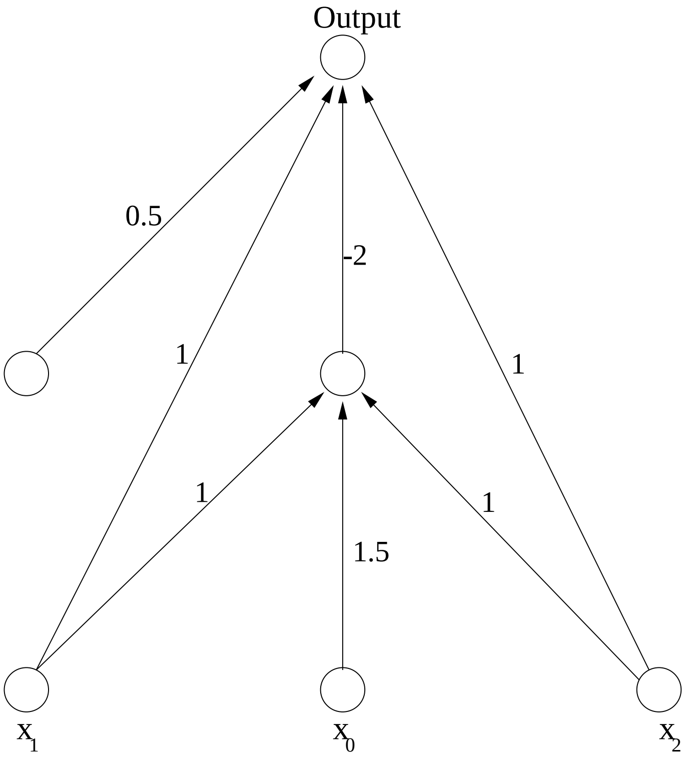
Using $\alpha = 0.25$ determine whether the following split would be pruned using the above method.

	Class		
	A	B	C
cond 1	6	2	0
cond 2	0	0	2

5. (a) Given a set of points $(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n$ in which $\boldsymbol{x}_i \in \mathbb{R}^p$ 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*.

(b) State the XOR problem.

The figure below shows a neural network, involving a single hidden neuron and some *skip layer* connections. Show that this network solves the XOR problem for a suitably defined "activation function" at the central node.



(c) Given data $x_i \in \mathbb{R}^p$ and corresponding $Y_i \in \{1, \dots, K\}$ for $i = 1, \dots, n$ 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.

Binomial Distribution Function for n = 8

This table gives $\mathsf{P}(X \leq r)$ for $X \sim \text{Bin}(8, p)$

For $p \geq .5$ you may use the result

$$P(X \leq r) = 1 - P(Y \leq n - r - 1) \qquad \text{with} \quad Y \sim \text{Bin}(n, 1 - p)$$

				r				
p	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0.01	0.9227	0.9973	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
0.03	0.7837	0.9777	0.9987	0.9999	1.0000	1.0000	1.0000	1.0000
0.05	0.6634	0.9428	0.9942	0.9996	1.0000	1.0000	1.0000	1.0000
0.07	0.5596	0.8965	0.9853	0.9987	0.9999	1.0000	1.0000	1.0000
0.09	0.4703	0.8423	0.9711	0.9966	0.9997	1.0000	1.0000	1.0000
0.11	0.3937	0.7829	0.9513	0.9929	0.9993	1.0000	1.0000	1.0000
0.13	0.3282	0.7206	0.9257	0.9871	0.9985	0.9999	1.0000	1.0000
0.15	0.2725	0.6572	0.8948	0.9786	0.9971	0.9998	1.0000	1.0000
0.17	0.2252	0.5943	0.8588	0.9672	0.9950	0.9995	1.0000	1.0000
0.19	0.1853	0.5330	0.8185	0.9524	0.9917	0.9991	0.9999	1.0000
0.21	0.1517	0.4743	0.7745	0.9341	0.9871	0.9984	0.9999	1.0000
0.23	0.1236	0.4189	0.7276	0.9120	0.9809	0.9973	0.9998	1.0000
0.25	0.1001	0.3671	0.6785	0.8862	0.9727	0.9958	0.9996	1.0000
0.27	0.0806	0.3193	0.6282	0.8567	0.9623	0.9936	0.9994	1.0000
0.29	0.0646	0.2756	0.5772	0.8237	0.9495	0.9906	0.9990	0.9999
0.31	0.0514	0.2360	0.5264	0.7874	0.9339	0.9866	0.9984	0.9999
0.33	0.0406	0.2006	0.4764	0.7481	0.9154	0.9813	0.9976	0.9999
0.35	0.0319	0.1691	0.4278	0.7064	0.8939	0.9747	0.9964	0.9998
0.37	0.0248	0.1414	0.3811	0.6626	0.8693	0.9664	0.9949	0.9996
0.39	0.0192	0.1172	0.3366	0.6172	0.8414	0.9561	0.9928	0.9995
0.41	0.0147	0.0963	0.2948	0.5708	0.8105	0.9437	0.9900	0.9992
0.43	0.0111	0.0784	0.2560	0.5238	0.7765	0.9289	0.9864	0.9988
0.45	0.0084	0.0632	0.2201	0.4770	0.7396	0.9115	0.9819	0.9983
0.47	0.0062	0.0504	0.1875	0.4306	0.7001	0.8914	0.9761	0.9976
0.49	0.0046	0.0398	0.1581	0.3854	0.6584	0.8682	0.9690	0.9967

Binomial Distribution Function for n = 10

This table gives $\mathsf{P}(X \leq r)$ for $X \sim \text{Bin}(10, p)$

For $p \geq .5$ you may use the result

$$P(X \leq r) = 1 - P(Y \leq n - r - 1) \qquad \text{with} \quad Y \sim \text{Bin}(n, 1-p)$$

					r					
p	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.01	0.9044	0.9957	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.03	0.7374	0.9655	0.9972	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.05	0.5987	0.9139	0.9885	0.9990	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
0.07	0.4840	0.8483	0.9717	0.9964	0.9997	1.0000	1.0000	1.0000	1.0000	1.0000
0.09	0.3894	0.7746	0.9460	0.9912	0.9990	0.9999	1.0000	1.0000	1.0000	1.0000
0.11	0.3118	0.6972	0.9116	0.9822	0.9975	0.9997	1.0000	1.0000	1.0000	1.0000
0.13	0.2484	0.6196	0.8692	0.9687	0.9947	0.9994	0.9999	1.0000	1.0000	1.0000
0.15	0.1969	0.5443	0.8202	0.9500	0.9901	0.9986	0.9999	1.0000	1.0000	1.0000
0.17	0.1552	0.4730	0.7659	0.9259	0.9832	0.9973	0.9997	1.0000	1.0000	1.0000
0.19	0.1216	0.4068	0.7078	0.8961	0.9734	0.9951	0.9994	0.9999	1.0000	1.0000
0.21	0.0947	0.3464	0.6474	0.8609	0.9601	0.9918	0.9988	0.9999	1.0000	1.0000
0.23	0.0733	0.2921	0.5863	0.8206	0.9431	0.9870	0.9979	0.9998	1.0000	1.0000
0.25	0.0563	0.2440	0.5256	0.7759	0.9219	0.9803	0.9965	0.9996	1.0000	1.0000
0.27	0.0430	0.2019	0.4665	0.7274	0.8963	0.9713	0.9944	0.9993	0.9999	1.0000
0.29	0.0326	0.1655	0.4099	0.6761	0.8663	0.9596	0.9913	0.9988	0.9999	1.0000
0.31	0.0245	0.1344	0.3566	0.6228	0.8321	0.9449	0.9871	0.9980	0.9998	1.0000
0.33	0.0182	0.1080	0.3070	0.5684	0.7936	0.9268	0.9815	0.9968	0.9997	1.0000
0.35	0.0135	0.0860	0.2616	0.5138	0.7515	0.9051	0.9740	0.9952	0.9995	1.0000
0.37	0.0098	0.0677	0.2206	0.4600	0.7061	0.8795	0.9644	0.9929	0.9991	1.0000
0.39	0.0071	0.0527	0.1840	0.4077	0.6580	0.8500	0.9523	0.9897	0.9986	0.9999
0.41	0.0051	0.0406	0.1517	0.3575	0.6078	0.8166	0.9374	0.9854	0.9979	0.9999
0.43	0.0036	0.0309	0.1236	0.3102	0.5564	0.7793	0.9194	0.9798	0.9969	0.9998
0.45	0.0025	0.0233	0.0996	0.2660	0.5044	0.7384	0.8980	0.9726	0.9955	0.9997
0.47	0.0017	0.0173	0.0791	0.2255	0.4526	0.6943	0.8729	0.9634	0.9935	0.9995
0.49	0.0012	0.0126	0.0621	0.1888	0.4018	0.6474	0.8440	0.9520	0.9909	0.9992

END