UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2002

MEng Honours Degree in Electrical Engineering Part IV
MEng Honours Degree in Information Systems Engineering Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C493=I4.48=E4.41

INTELLIGENT DATA AND PROBABILISTIC INFERENCE

Monday 29 April 2002, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 a i) Describe briefly the following algorithms for supervised learning and unsupervised learning:

1) Decision tree learning

2) K-means method.

ii) Compare and contrast the concepts of model-based classification methods and instance-based learning classification methods.

b The table below details the results of a clinical trials study on the effectiveness of the new Hay Fever treatment. In this table the columns "Gender" and "Dose" are the attributes of each trial, and the column "Effectiveness" is measured outcome. Your task is to use this table as training examples for a decision tree classification module that can predict the effectiveness of the treatment on unseen cases.

Patient ID	Gender	Dose	Effective
1	M	High	Yes
2	M	Low	Yes
3	M	Medium	No
4	M	Low	Yes
5	M	High	No
6	M	Medium	Yes
7	F	High	Yes
8	F	Low	No
9	F	Low	No
10	F	High	No

i) Taking this data set as an example, explain briefly what is meant by the entropy of a data set with respect a target class, and explain what is meant by information gain.

ii) Construct a decision tree from this training data set.

The two parts carry, respectively, 25% (12.5% each subparts), 75% (25%, 50%, each subparts) of the marks.

2 The goal of Association Rule discovery is to find rules that have the form: if item-A then item-B with support: S% and confidence: C%

a.

i) Explain what is meant by support and confidence in the above example

ii) Describe how using thresholds defining minimum support and confidence values can be used to improve the efficiency of algorithms for finding association rules in large data sets.

iii) Given the data set described below find all rules between single items that have a support > 55%. For each rule report both the support and confidence of the rule.

```
Basket1:  (cheese, egg)
Basket2:  (butter, egg)
Basket3:  (bread, butter, cheese, egg)
Basket4:  (butter, cheese, egg)
Basket5:  (bread, butter, cheese)
Basket6:  (bread, butter, cheese, egg)
Basket7:  (butter, cheese, egg)
Basket8:  (bread, butter, cheese)
Basket9:  (cheese, egg)
Basket10: (egg)
```

b.

i) Explain how hierarchical clustering algorithms work

ii) Based on a Euclidean distance similarity measure, calculate the similarity matrix between the observations.

Observation ID	X	Y
A	0	1
B	0	5
C	1	1
D	0	6

iii) Based on the distance matrix use hierarchical clustering to generate a cluster tree using single linkage method

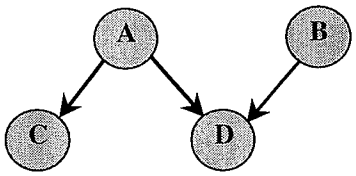
The two parts carry, respectively, 50% (10%, 10%, 30% each subparts), 50% (10%, 20%, 20% each subparts) of the marks.

3. The Minimum Description Length Metric

The minimum description length metric is used for selecting the best among competing networks, for making inferences about a set of variables described by a data set. It has the following equation:

$$MDL(B|D) = |B| (\log_2 N)/2 - \log_2(P(D|B)$$

a. Given the following network and data set, calculate the accuracy ($\log_2(P(D|B))$ of the network.



A	B	C	D
a0	b0	c1	d1
a0	b0	c0	d1
a1	b1	c0	d0
a1	b0	c1	d0
a1	b1	c0	d1
a1	b1	c0	d1
a2	b1	c0	d1
a2	b0	c1	d1

(NB A has three states the other nodes have two)

b. Explain briefly why it is computationally desirable to use the log likelihood, rather than just the probability of the data given the network P(D|B)?

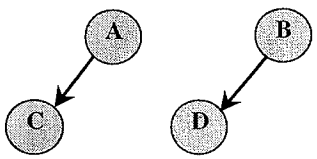
c. Calculate is the number of parameters of the network of part (a), designated |B|, in the MDL formula.

d. Explain why the number of items in the data set N is used in the computation of the size of a network.

e. From your answers to part (a) and part (c) calculate the MDL measure for the network and data set of part (a).

f. If every entry in the data set of part (a) were duplicated, so that it contained twice as many entries, what would the MDL measure become? Comment briefly on your answer.

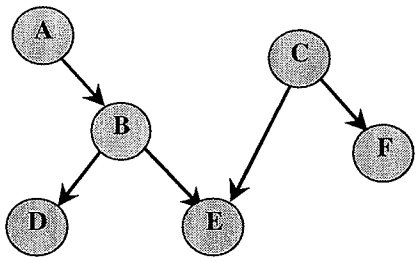
g. A competing network is shown below. Would the MDL measure choose the network in this part over that in part a?



The seven parts carry, respectively, 25%, 10%, 10%, 15% ,10%, 10% and 20% of the marks.

4. Probability propagation.

The following questions relate to propagating probabilities in the following network in which every variable has just two states. Pearl's operating equations are given at the end of the question.



P(A) = [P(a0), P(a1)] = [0.25,0.75], P(C) = [0.5,0.5]

$$P(B|A) = \begin{matrix} P(b0|a0) & P(b0|a1) \\ P(b1|a0) & P(b1|a1) \end{matrix} = \begin{matrix} 0.8 & 0.9 \\ 0.2 & 0.1 \end{matrix}$$

$$P(D|B) = \begin{matrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{matrix} \quad P(F|C) = \begin{matrix} 0.9 & 0.8 \\ 0.1 & 0.2 \end{matrix}$$

$$P(E|B\&C) = \begin{matrix} P(e0|b0\&c0) & P(e0|b1\&c0) & P(e0|b0\&c1) & P(e0|b1\&c1) \\ P(e1|b0\&c0) & P(e1|b1\&c0) & P(e1|b0\&c1) & P(e1|b1\&c1) \end{matrix} = \begin{matrix} 0.2 & 0.6 & 0.5 & 0.9 \\ 0.8 & 0.4 & 0.5 & 0.1 \end{matrix}$$

a. Following the propagation of π evidence during initialisation what is the initial posterior probability distribution over variable E?

b. If D is instantiated to d0 and F is instantiated to f1 compute the posterior probability distribution over A.

c. If, in addition to the instantiations given in part (b), node E is now instantiated to state e0 and C to state c1, what is the new probability distribution over variable A?

d. It is discovered that for one state of variable B the variables D and E show a strong correlation. Explain briefly why this is undesirable for making inferences about the probability distribution over A.

e. Discuss two ways in which you might alter the network structure to take account of the correlation noted in part (d). Mention the merits and demerits of each.

The five parts carry equal marks.