UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2001

MSci Honours Degree in Mathematics and Computer Science 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
This paper is also taken for the relevant examinations for the Associateship of the Royal College of Science

PAPER C493

INTELLIGENT DATA AND PROBABILISTIC INFERENCE

Tuesday 15 May 2001, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

1 a Describe briefly the concepts:

i) Classification in Predicative Modeling

ii) Association Rules in Data Mining

iii) Entropy

iv) Distance-based Clustering and Density-based Clustering

b

i) Given the following collection of symbolic data :

{red, yellow, red, green, yellow, green, red, red, blue, blue}

What is the entropy of this data set?

Given the following data set:

Instance	Blood Pressure	Weather	Mood (class value)
1	High	Raining	Bad
2	High	Sunny	Normal
3	Low	Cloudy	Bad
4	Medium	Sunny	Normal
5	Low	Sunny	Good
6	High	Sunny	Good
7	Medium	Cloudy	Good
8	High	Cloudy	Bad
9	Low	Raining	Normal
10	Low	Cloudy	Normal

ii) To build up a decision tree with Mood as the class value, which attribute should be chosen as the root node and why?

iii) Construct a decision tree and derive two rules from the tree.

iv) Discuss the ways of testing the accuracy of the tree and the ways of using the tree for prediction.

The two parts carry, respectively, 20% (5% each subpart), 80% (20% each subpart) of the marks.

2 a i) Describe the principle of the naïve Bayes classification algorithm.

ii) Describe the K-nearest neighbours (KNN) algorithm for classification.

b Given the following table,

Transaction ID	Items
T1	Bread, Butter, Eggs
T2	Butter, Eggs, Milk
T3	Butter
T4	Bread, Butter

i) List all the itemsets with a support no less than 40%.

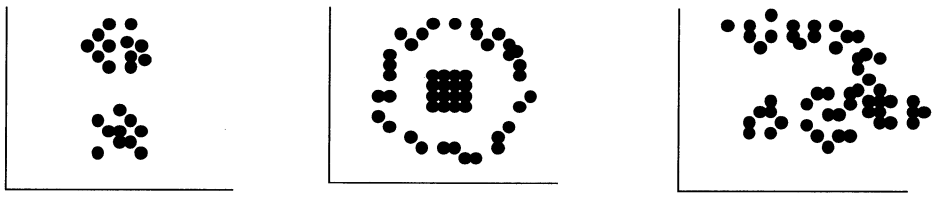
ii) Compute all the association rules with the confidence no less than 60%.

iii) Let X, Y, Z be three itemsets. After deriving the association rules X=>Y and Y =>Z , can we conclude that X=> Z and why?

c

i) Compare the advantages and disadvantages of both the decision tree approach and neural networks approach for classification.

ii) Describe the K-means clustering algorithm. Given the three data sets (two dimensional data) in the following shape, discuss the performance of K-means on each data set.

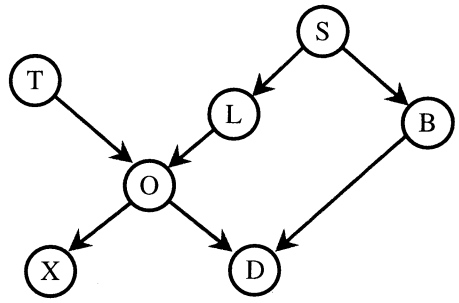


a b c

The three parts carry, respectively, 25% (10% first subpart and 15% the second subpart), 45% (15% each subparts), 30% (15% each subparts) of the marks.

3. Probability Propagation

The following network is used for reasoning about patients with suspected lung disease



B	Bronchitis
D	Dyspnea
L	Lung Cancer
O	Reduced Lung Capacity
S	Smoker
T	Tuberculosis
X	Positive XRay

All the nodes are binary, and the prior and conditional probabilities are as follows:

$$P(O|T\&L) = \begin{bmatrix} P(O1|T1\&L1) & P(O1|T1\&L2) & P(O1|T2\&L1) & P(O1|T2\&L2) \\ P(O2|T1\&L1) & P(O2|T1\&L2) & P(O2|T2\&L1) & P(O2|T2\&L2) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P(D|O\&B) = \begin{bmatrix} P(D1|O1\&B1) & P(D1|O1\&B2) & P(D1|O2\&B1) & P(D1|O2\&B2) \\ P(D2|O1\&B1) & P(D2|O1\&B2) & P(D2|O2\&B1) & P(D2|O2\&B2) \end{bmatrix} = \begin{bmatrix} 0.9 & 0.8 & 0.7 & 0.1 \\ 0.1 & 0.2 & 0.3 & 0.9 \end{bmatrix}$$

$$P(L|S) = \begin{bmatrix} P(L1|S1) & P(L1|S2) \\ P(L2|S1) & P(L2|S2) \end{bmatrix} = \begin{bmatrix} 0.2 & 0.1 \\ 0.8 & 0.9 \end{bmatrix}$$

$P(T) = (0.1,0.9)$

$$P(B|S) = \begin{bmatrix} P(B1|S1) & P(B1|S2) \\ P(B2|S1) & P(B2|S2) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.1 \\ 0.9 & 0.9 \end{bmatrix}$$

$P(S) = (0.3,0.7)$

$$P(X|O) = \begin{bmatrix} P(X1|O1) & P(X1|O2) \\ P(X2|O1) & P(X2|O2) \end{bmatrix} = \begin{bmatrix} 0.4 & 0.1 \\ 0.6 & 0.9 \end{bmatrix}$$

Given that the evidence propagated between nodes is defined by the following equations:

For one parent only

$$\lambda_c(a_k) = \sum_{j=1}^{m} P(c_j \mid a_k)\, \lambda(c_j)$$

For two parents

$$\lambda_c(a_k) = \sum_{i=1}^{n} \pi_C(b_i) \sum_{j=1}^{m} P(c_j \mid a_k \& b_i)\, \lambda(c_j)$$

$$\pi(c_i) = \sum_{j=1}^{n} \sum_{k=1}^{m} P(c_i \mid a_j \& b_k)\, \pi_C(a_j)\, \pi_C(b_k)$$

a. Calculate the π evidence for the nodes L and O before any measurements are made.

b. A new patient arrives and has an X-Ray taken. Fortunately for him it is negative (state X_2). Given just this evidence calculate the probability of his suffering from Lung cancer (i.e. the probability distribution over L).

c. He is now examined and found to be suffering from Dyspnea (D is in state D1). Explain why it is no longer possible to compute a probability of his suffering from lung cancer.

d. He now admits to being a smoker (S is in state s1). Calculate the probability of his suffering from Lung Cancer.

e. Explain briefly the advantages and disadvantages of using a join tree for calculating probabilities rather than simple λ and π messages.

4. The maximum Weighted Spanning Tree

A data warehouse contains the following vast data set connecting three variables A B and C:

A	B	C
a1	b1	c1
a1	b1	c2
a1	b1	c2
a1	b2	c1
a2	b2	c2
a2	b2	c1
a2	b2	c2
a2	b2	c1

a. Construct co-occurrence matrices for the three possible pairings AB, BC and AC:

	a1	a2
b1		
b2		

etc

b. From the co-occurrences construct the joint probability table for each pair, and the marginalisations, using the following format:

	a1	a2	P(B)
b1			
b2			
P(A)			

etc.

c. Calculate the L1 metric for each possible pair of nodes

$$Dep(A,B) = \sum_{AXB} |P(ai\&bj) - P(ai)P(bj)|$$

d. Given that A is the root node, construct the tree.

e. Calculate the prior probability distribution P(A).

f. Calculate the two conditional probability matrices for the tree found in part 2d.

g. The tree is being used in the case where it is not possible to measure node B. The state of C is found to be c2. Estimate the probabilities of nodes A and B.