UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2007

MEng Honours Degree in Electrical Engineering Part IV
MSc in Computing for Industry
MEng Honours Degree in Information Systems Engineering Part IV
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=I4.48=E4.41

INTELLIGENT DATA AND PROBABILISTIC INFERENCE

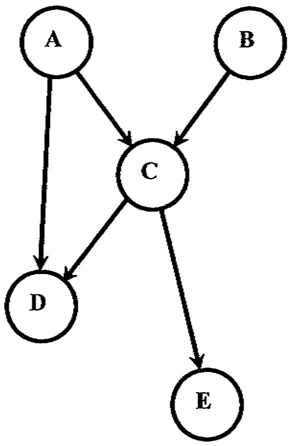
Tuesday 1 May 2007, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1. Probability Propagation

A Bayesian network has the following structure, prior probabilities and conditional probability matrices:



P(A)	a_1	a_2
	0.5	0.5

P(B)	b_1	b_2
	0.25	0.75

P(C|A&B)		a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2
	c_1	0.5	0	1	0.25
	c_2	0.5	1	0	0.75

P(E|C)		c_1	c_2
	e_1	0.25	0.75
	e_2	0.75	0.25

P(D|A&C)		a_1 c_1	a_1 c_2	a_2 c_1	a_2 c_2
	d_1	0	1	0.25	0.5
	d_2	1	0	0.75	0.5

a Find the joint probability of the network for the data point {a1,b1,c2,d1,e1}

b The network is to be used for inference. Describe what happens during the initialisation stage, and calculate the π evidence for variable D after initialisation.

c The following three instantiations are made: B=b1, C=c2, D=d1. Calculate the *posterior* probabilities of variables A and E.

d Given that node D alone is instantiated, explain briefly why probability propagation using Perl's algorithm is not possible.

e If the method of cutset conditioning is to be used to allow probability propagation which nodes of the graph could be used as a cutset? Which node is the best choice for a cutset?

f What does the Markov blanket of a node in a Bayesian network represent? Which nodes belong to the Markov blanket of B.

Pearl's operating equations for probability propagation are provided on a supplementary sheet.

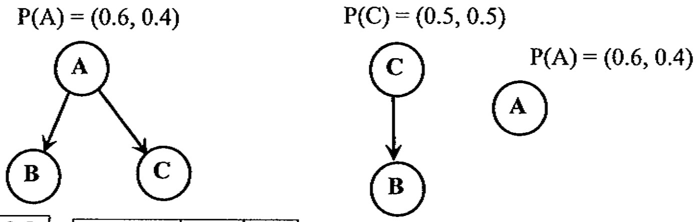
The six parts carry, respectively, 15%,20%, 20%15%,15%,15% of the marks.

2 Model Accuracy

A data set over three binary variables, A,B and C, consists of the following eight data points:

a_1 b_1 c_1
a_2 b_1 c_1
a_2 b_1 c_1
a_2 b_1 c_2
a_2 b_1 c_1
a_1 b_1 c_1
a_2 b_1 c_1
a_2 b_1 c_2

Two Bayesian networks are proposed to represent the data set.



P(B|A) =	1	0.5
	0	0.5

P(C|A) =	1	0.5
	0	0.5

P(B|C) =	1	1
	0	0

The prior probabilities and the link matrices for each network are given above.

a Compute the joint probability distribution of the variables (ie the probability distribution over the eight possible states of the variables) for the data set.

b For each of the two networks compute the joint probability distribution over all possible states of the variables.

c By computing the Euclidian distance between the joint probability distributions computed from the data in part a, and the joint probability distribution calculated from each network in part b, determine which network is the most accurate.

d Compute the model accuracy for each network using the log likelihood of the data given the network, $\log_2(P(D|B))$, as a measure. This is the accuracy measure used to compute the MDL score.

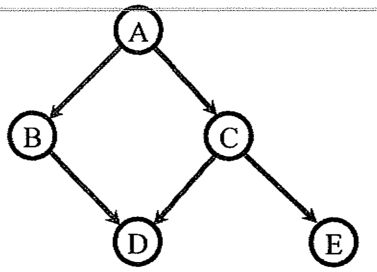
e If the data set was increased to 16 points by duplicating each of the original 8 points, explain how the model accuracy computed in part d, and the Euclidian distances computed in part c would change.

f Suggest why the log likelihood measure of model accuracy (computed in part d) is used in preference to a measure of distributional similarity, for example the Euclidean distribution computed in part c.

The six parts carry, respectively, 15%, 20%, 15%, 15%, 15%, 20% of the marks.

3 Probability propagation in Join Trees

The following network relating to metastatic cancer has been proposed:.



P(A)	0.5	0.5

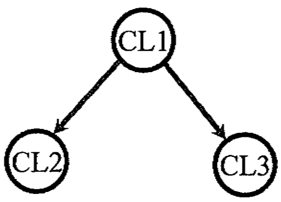
P(B|A)	a_1	a_2
b_1	1	0.5
b_2	0	0.5

P(C|A)	a_1	a_2
c_1	0	0.5
c_2	1	0.5

P(D|B&C)	$b_1 c_1$	$b_1 c_2$	$b_2 c_1$	$b_2 c_2$
d_1	0.5	0.75	0	0.5
d_2	0.5	0.25	1	0.5

P(E|C)	c_1	c_2
e_1		
e_2	0.5	0.5

In order to make inferences it was converted to the following join tree form.



Variable	Clique	Clique Variables W	Clique Function ψ
E	3	C,E	P(E|C)
D	2	B,C,D	P(D|B&C)
C	1	A,B,C	P(A)P(C|A)P(B|A)
B	1	A,B,C	
A	1	A,B,C	

a Explain what is meant by the running intersection property of the three cliques and show that it is satisfied in the above join tree.

b The clique potential function is stored as a table of values for each of the possible states of the variables of the clique. Calculate the potential function for clique no 1.

c In an particular case D is instantiated to state d_2. For this single instantiation fill up a table defining the cliques using the following format:

Clique	R	S	States
1			
2			
3			

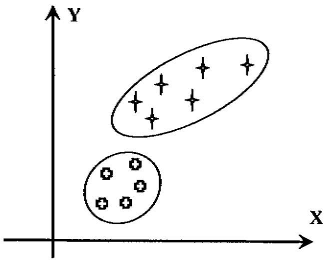
d Given that the λ message from a clique to its parent is defined as: $\Sigma_R \psi(Wi)$ calculate the λ message from clique 2 to clique 1 following the instantiation of D

e Using the λ message calculated in part d update the potential function of clique 1 which you calculated in part b.

The five parts carry equal marks.

4 The most discriminant feature method

a With reference to the two class example shown below, explain the difference between the PCA transformation and the LDA transformation. Without calculating numeric values, sketch the same figure showing the approximate directions of the LDA and PCA axes.



b Let an N x n mean adjusted data matrix U be composed of N observations with n variables. Assume that U is defined with n columns and N rows, there are g different classes ($g \geq 1$), and all observations are linearly independent. The covariance matrix for the data is estimated using the pooled method:

$$\Sigma_p = \frac{1}{N-g} \sum_{i=1}^{g} (N_i - 1)\Sigma_i$$
$$= \frac{(N_1 - 1)\Sigma_1 + (N_2 - 1)\Sigma_2 + \cdots + (N_g - 1)\Sigma_g}{N-g}$$

Let P be the corresponding PCA projection matrix. What is the maximum number eigenvectors with non-zero eigenvalues that P can have? What is the corresponding dimension of P?

c The most expressive features matrix U_p is found by projecting the training set U on P. What dimension does U_p have (assuming the PCA projection is maximum)?

d Following the PCA transformation an LDA is performed on the transformed points.What is the maximum dimension of the linear discriminant analysis transformation L assuming that there are g classes in the data?

e The most discriminant features matrix U_f is found by projecting U_p on L. Define the final dimension of the data matrix (U_f) in the PCA+LDA space.

The five parts carry equal marks.