UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2006

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

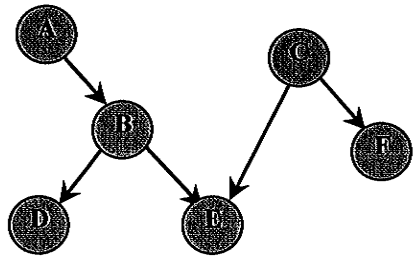
Monday 24 April 2006, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1. 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.



$$P(A) = [P(a_0), P(a_1)] = [0.25, 0.75], \quad P(C) = [0.5, 0.5]$$

$$P(B|A) = \begin{bmatrix} P(b_0|a_0) & P(b_0|a_1) \\ P(b_1|a_0) & P(b_1|a_1) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.9 \\ 0.2 & 0.1 \end{bmatrix}$$

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

$$P(E|B\&C) = \begin{bmatrix} P(e_0|b_0\&c_0) & P(e_0|b_1\&c_0) & P(e_0|b_0\&c_1) & P(e_0|b_1\&c_1) \\ P(e_1|b_0\&c_0) & P(e_1|b_1\&c_0) & P(e_1|b_0\&c_1) & P(e_1|b_1\&c_1) \end{bmatrix}$$

$$= \begin{bmatrix} 0.2 & 0.6 & 0.5 & 0.9 \\ 0.8 & 0.4 & 0.5 & 0.1 \end{bmatrix}$$

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 d_0 and F is instantiated to f_1 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 e_0 and C to state c_1, 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.

Pearl's Operating Equations for Probability propagation

Operating Equation 1: λ message

The λ message from C to A is given by

For one parent only

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

For two parents

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

Operating Equation 2: The π Message

If C is a child of A, the π message from A to C is given by:

$$\pi_c(a_j) = \begin{array}{ll} 1 & \text{if } A \text{ is instantiated for } a_j \\ 0 & \text{if } A \text{ is instantiated but not for } a_j \\ P'(a_j)/\lambda_c(a_j) & \text{if } A \text{ is not instantiated} \end{array}$$

Operating Equation 3: The λ evidence

If C is a node with n children $D_1, D_2, \ldots D_n$, then the λ evidence for C is:

$$\lambda(c_j) = \begin{cases} 1 & \text{if } C \text{ is instantiated for } c_j \\ 0 & \text{if } C \text{ is instantiated but not for } c_j \\ \prod_i \lambda_{Di}(c_j) & \text{if } C \text{ is not instantiated} \end{cases}$$

Operating Equation 4: The π evidence

If C is a child of two parents A and B the π evidence for C is given by:

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

Operating Equation 5: the posterior probability

If C is a variable the (posterior) probability of C based on the evidence received is written as:

$P'(c_i) = \alpha\, \lambda(c_i)\, \pi(c_i)$

where α is chosen to make $\Sigma\, P'(c_i) = 1$

2. The Minimum Description Length Metric

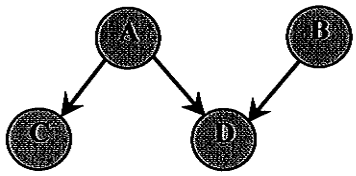
The minimum description length metric is used for selecting the best among competing inference networks. It has the following equation:

$$\text{MDL}(B|D) = |B|\,(\log_2 N)/2 - \log_2(P(D|B))$$

where B is a Bayesian network, D a data set, $|B|$ is the number of parameters in the network and N the number of samples in the data set.

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

A	B	C	D
a_0	b_0	c_1	d_1
a_0	b_0	c_0	d_1
a_1	b_1	c_0	d_0
a_1	b_0	c_1	d_0
a_1	b_1	c_0	d_1
a_1	b_1	c_0	d_1
a_2	b_1	c_0	d_1
a_2	b_0	c_1	d_1



(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 $P(D|B)$?

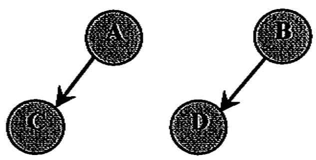
c. Calculate the number of parameters of the network of part (a), designated $|B|$, in the MDL formula.

d. Explain why the number, N, of items in the data set 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 each entry in the data set of part (a) were duplicated, so that the data set contained twice as many entries, what would the MDL measure become? Comment briefly on your answer.

g. A competing network is shown below. Given the data set above would the MDL measure choose the network in this part over that in part a?



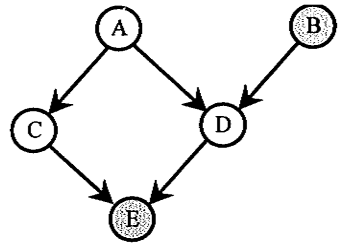
The seven parts carry, respectively, 25%, 10%, 10%, 15%, 10%, 10% and 20% of the marks.

3. Sampling and Re-Sampling

a. In a Bayesian network, the Markov blanket of node A is the set of nodes that separate A from all the other nodes of the network. Given the probability distributions over the states of each node of the Markov blanket of A, the probability distribution of node A can be determined without reference to any other node of network. Formally, the Markov blanket of a node A of a Bayesian network is defined as a set containing the following nodes:
(i) The parents of A
(ii) The children of A
(iii) The other parents of the children of A

Explain why the other parents of the children of A are a necessary part of the Markov blanket of A.

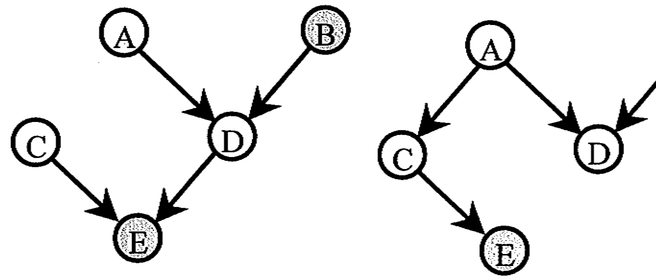
b. In the following network the nodes B and E are instantiated after initialisation. Pearl's propagation algorithm cannot be used since there is a loop in the network.



Instead it is proposed that the probabilities of nodes A, C and D will be calculated using a Monte Carlo Markov chain based on Gibbs sampling. Describe briefly how this algorithm works.

c. An alternative method of propagating the probabilities is to use cutset conditioning. Briefly describe how the cutset conditioning algorithm could be applied to the network of part b.

d. Yet another solution to propagating probability in the network of part b is to delete an arc. Given that the network was created from a large data set D and that the dependencies of AC and DE are both quite low there are two feasible networks that could be used:



Explain how the "leave one out" re-sampling method could be used with the data set D in order to determine which is the more accurate network.

e. Explain what is meant by a bootstrap data set and why they are used.

The five parts carry equal marks.

4. Classification

A classifier is to be designed to separate two classes for which the following data points are known:

Class 1: (2,1), (1,2), (2,4), (2,3)
Class 2: (3,6), (4,5), (5,7), (6,8)

a. Find the two individual class means and, using all the data, the grand mean and the pooled covariance matrix S_p.

b. Calculate the "between class" covariance matrix S_b from the two class means found in part a above.

c. Find the direction of the most discriminant LDA feature by finding the principal eigenvector of the matrix $L = S_p^{-1} S_b$. Hint: First calculate the eigenvalues using $\det(L-\lambda I)=0$, then solve $(L-\lambda I)\,\boldsymbol{e} = 0$ to find the eigenvectors.

d. Make an accurate sketch of the data plotted on the xy-Cartesian plane. On your sketch show:
   (i) The most discriminant LDA feature direction
   (ii) The support vectors that would separate the classes using a support vector machine

The four parts carry equal marks.