IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2015

MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
MSc in Computing Science
MSc in Computing Science (Specialist)
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

INTELLIGENT DATA AND PROBABILISTIC INFERENCE

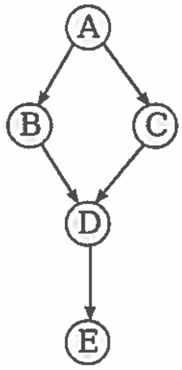
Thursday 26 March 2015, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

Bayesian Network Definition

The following Bayesian network definition is referred to in questions 1, 2 and 3. The variables are all binary, with, for example, variable A having states a_1 and a_2.



$$P(A) = [0.6 \quad 0.4]$$

$$P(B|A) = \begin{bmatrix} 0.4 & 0 \\ 0.6 & 1 \end{bmatrix}$$

$$P(C|A) = \begin{bmatrix} 0.5 & 0.3 \\ 0.5 & 0.7 \end{bmatrix}$$

$$P(D|B\&C) = \begin{bmatrix} 0.2 & 1 & 0.3 & 0.1 \\ 0.8 & 0 & 0.7 & 0.9 \end{bmatrix}$$

$$P(E|D) = \begin{bmatrix} 1 & 0.3 \\ 0 & 0.7 \end{bmatrix}$$

The conditional probabilities are given consistently in the form:

$$P(D|B\&C) = \begin{bmatrix} P(d_1|b_1\&c_1) & P(d_1|b_1\&c_2) & P(d_1|b_2\&c_1) & P(d_1|b_2\&c_2) \\ P(d_2|b_1\&c_1) & P(d_2|b_1\&c_2) & P(d_2|b_2\&c_1) & P(d_2|b_2\&c_2) \end{bmatrix}$$

The equations for propagating probabilities in Bayesian networks are:

The λ message from child C to parents A and B is given by:

$$\lambda_C(a_i) = \sum_{j=1}^{m} \pi_C(b_j) \sum_{k=1}^{n} P(c_k|a_i\&b_j)\lambda(c_k)$$

In the case where we have a single parent (A) this reduces to:

$$\lambda_C(a_i) = \sum_{k=1}^{n} P(c_k|a_i)\lambda(c_k)$$

and for the case of the single parent we can use the simpler matrix form:

$$\lambda_\mathbf{C}(\mathbf{A}) = \lambda(\mathbf{C})\mathbf{P}(\mathbf{C}|\mathbf{A})$$

The matrix form for multiple parents relates to the joint states of the parents.

$$\lambda_\mathbf{C}(\mathbf{A}\&\mathbf{B}) = \lambda(\mathbf{C})\mathbf{P}(\mathbf{C}|\mathbf{A}\&\mathbf{B})$$

It is necessary to separate the λ evidence for the individual parents with a scalar equation of the form:

$$\lambda_C(a_i) = \Sigma_j \pi_C(b_j)\lambda_C(a_i\&b_j)$$

The π evidence to child node C from two parents A and B is given by:

$$\pi(c_k) = \sum_{i=1}^{l} \sum_{j=1}^{m} P(c_k|a_i\&b_j)\pi_C(a_i)\pi_C(b_j)$$

This can be written in matrix form as follows:

$$\pi(\mathbf{C}) = \mathbf{P}(\mathbf{C}|\mathbf{A}\&\mathbf{B})\pi_{\mathbf{C}}(\mathbf{A}\&\mathbf{B})$$

where

$$\pi_C(a_i\&b_j) = \pi_C(a_i)\pi_C(b_j)$$

The single parent matrix equation is:

$$\pi(\mathbf{C}) = \mathbf{P}(\mathbf{C}|\mathbf{A})\pi_{\mathbf{C}}(\mathbf{A})$$

1 **Probability Propagation.**

For the Bayesian network defined above:

a Calculate the probability distribution over variable D after initialisation, but before any nodes are instantiated.

b Is the probability distribution calculated in part a exact or approximate? Explain your answer.

c Variable E is instantiated to state e_1. Describe what messages are passed as a result.

d In addition to variable E, variable C is now instantiated to state c_2. Calculate the probability distribution over node B.

e If cutset conditioning is used in propagating probabilities, which individual nodes could be used as a cut set?

Which node, if any, is the best choice of cutset?

The five parts carry equal marks.

2 **Joining Nodes.**

a It is suggested that the Bayesian network defined above could be modified to remove the loop by joining nodes B and C. Draw the clustered network (with B and C joined), and calculate any new conditional probability matrices that are required, assuming $P(B\&C|A) = P(B|A)P(C|A)$.

b Assuming that D is instantiated to d_1 and A is instantiated to a_2, calculate the probability distribution over variable B using the clustered network.

c Probability propagation is to be computed using a join tree. Find a suitable join tree by drawing the moral graph and finding its cliques. Explain how your join tree satisfies the running intersection property.

d For the join tree in which variable E belongs to the bottom node, calculate the λ message that the bottom node will send to its parent if varible E is instantiated to e_2.

e What are the advantages and disadvantages of using the join tree method compared to simple node clustering?

The five parts carry equal marks.

3 Monte Carlo Markov Chain Methods

a. Explain what is meant by the Markov blanket of a node in a Bayesian network. For the network defined above what are the Markov blankets of nodes A and C?

b. Given that, in the Bayesian network defined above, only node E is instantiated explain how a Monte Carlo Markov Chain (MCMC) method, based on Gibbs sampling, could be used to find the probability distributions over the uninstantiated nodes (A, B, C and D).

c. Gibbs sampling is said to be:

   i) Ergodic

   ii) Balanced

   Explain what is meant by the terms, and why the MCMC method that you described in part b ensures that these properties hold.

d. The Metropolis-Hastings algorithm provides a criterion for adding a new sample to a sample chain. It uses a probability of acceptance defined as:

   $p_t = min(\alpha, 1)$

   where: $\alpha = P(X^{t+1})Q(X^t|X^{t+1})/P(X^t)Q(X^{t+1}|X^t)$.

   $Q(X|Y)$ is called the proposal density and is the probability of sampling state X given that the network is in state Y. With reference to your answer to part b explain how the proposal density $Q([a_1\ b_1\ c_2\ d_1]|[a_2\ b_1\ c_2\ d_1])$ could be computed.

e. Compare and contrast the use of the MCMC methods with the join tree algorithm for making inferences in Bayesian networks. Consider cases where the variables are all highly dependent, and there are therefore many arcs in the network, and those where there is little dependency between the variables, and therefore there are few arcs in the network.

The five parts carry equal marks.

4 **Linear Discriminant Analysis.**

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

Class 1	(2,1)	(2,3)	(2,3)	(1,2)
Class 2	(1,1)	(-1,1)	(-1,-1)	(1,-1)

a Find the two individual class means and, using all the data, the grand mean and the pooled scatter matrix S_p. (The scatter matrix is the un-normalised co-variance matrix.)

b Calculate the between class scatter 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 $(L - \lambda I)\phi = 0$ to find the eigenvectors ϕ.

d i) Make an accurate sketch of the data plotted on the xy-Cartesian plane showing the most discriminant LDA feature direction.

ii) On the same sketch draw the approximate direction of the first principal component of the whole data set, and the approximate position of the SVM separating hyperplane for the two classes. Label your sketch clearly.

e What assumptions does the LDA method make about the individual class covariances? Explain briefly the circumstances in which this could cause errors.

The five parts carry equal marks.