IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2013

MSci 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
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

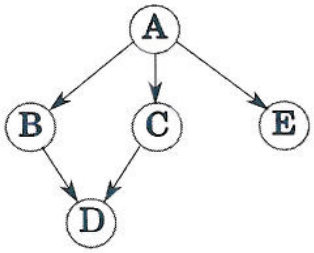
Thursday 16 May 2013, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 Probability Propagation.

The following Bayesian network has been proposed to model air pollution at the junction of Cromwell Road and Earls Court Road. The variables are all binary, with, for example, variable A having states a_1 and a_2.



The conditional probabilities are given in the form

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

and are

$$P(A) = [0.1, 0.9]$$

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

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

$$P(D|B\&C) = \begin{bmatrix} 0.6 & 0.4 & 0 & 0.1 \\ 0.4 & 0.6 & 1 & 0.9 \end{bmatrix}$$

$$P(E|A) = \begin{bmatrix} 0.2 & 0.3 \\ 0.8 & 0.7 \end{bmatrix}$$

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:

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

a Node E is instantiated to state e_1 and all other nodes remain uninstantiated. Calculate the π evidence for node D when propagation terminates.

b Node D is now instantiated to state d_2. Calculate the first λ message that it will send to node B.

c The method of cutset conditioning could be used to calculate exact probabilities for the instantiation described in part b. Given that node C is taken to be the cutset, and that the two states of C are equiprobable, calculate the λ evidence for A when propagation finishes.

d Is the evidence found in part a approximate or exact? Explain your answer.

The four parts carry, respectively, 20%, 20%, 40%, and 20% of the marks.

2 Dependency Measures and Causal Directions

a The Kullback-Leibler divergence is defined as follows:

$$Dep(A, B) = \sum_{A \times B} P(a_i \& b_j) log_2((P(a_i \& b_j)/(P(a_i)P(b_j)))$$

where the sum is taken over all the states of variables A and B.

Explain how and why it is used as a dependency measure between two discrete variables in a data set.

b A possible alternative to using the Kullback-Leibler divergence is to use the correlation between two variables.

$$C(A, B) = \Sigma_{AB}/\sqrt{\sigma_A \sigma_B}$$

i) Explain briefly how correlation can be used in practice to estimate the dependency of two discrete variables.

ii) Explain how the correlation method differs from the Kullback-Leibler divergence method.

c Use the following data on discrete variables A and B to find the dependency between A and B using the Kullback-Leibler divergence.

A	a_1	a_1	a_1	a_2	a_2	a_3	a_3	a_3
B	b_1	b_2	b_3	b_2	b_1	b_3	b_1	b_2

d A spanning tree algorithm is used on a four variable data set A, B, C and D and produces the following undirected network.

In order to determine the causal directions the joint probability table for the three variables B, C and D is calculated from the data and it is:

B	C	D	$P(B\&C\&D)$
b_1	c_1	d_1	0.2
b_1	c_1	d_2	0.1
b_1	c_2	d_1	0.05
b_1	c_2	d_2	0.15
b_2	c_1	d_1	0.05
b_2	c_1	d_2	0.1
b_2	c_2	d_1	0.2
b_2	c_2	d_2	0.15

What can be deduced about the causal directions from this table? Explain your answer.

e List the possible sources of error in determining causal direction using marginal independence.

The five parts carry equal marks.

3 a A factor graph is a bipartite graph in which one type of node represents one or more of a set of random variables $(X_1, X_2, X_3, ...X_n)$, and the other type of nodes, called factors $(f_1, f_2, ...f_m)$, represent relations between the variables. In probabilistic inference the factors are so called because they belong to a factorisation of the joint probability distribution:

$$P(X_1, X_2, X_3, ...X_n) = \prod_{j=1}^{m} f_j(S_j)$$

The Bayesian network defined in question 1 can be transformed into a join tree which has the following graphical model:

For both the original Bayesian network and the join tree:

i) List the factors of the joint probability probability distribution.

ii) Draw the corresponding factor graph.

b Monte Carlo Markov Chain MCMC methods can be used to make inferences from Bayesian Networks in cases where propagation will not terminate. Explain how the method would work for the network given at the start of question 1 in the case where node D only is instantiated.

c The Metropolis algorithm provides a way of deciding whether to accept a newly drawn sample and add it to a sample chain in an MCMC process. In its simplest form it computes a probability of acceptance:

$$P_{acc} = \min[P(X^{n+1})/P(X^n), 1]$$

where X^n and X^{n+1} are the previously accepted and newly drawn sample. Sample X^{n+1} is accepted with probability P_{acc}

Explain what effect this acceptance criterion has on the MCMC sample chain.

d The Mahalanobis distance between two two-dimensional points $(x_1, y_1)(x_2, y_2)$ is defined by the formula:

$$\sqrt{((x_2 - x_1, y_2 - y_1)\Sigma^{-1}(x_2 - x_1, y_2 - y_1)^T)}$$

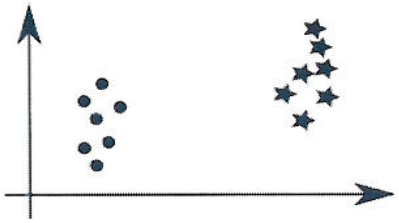
where Σ is a class covariance matrix.

Explain, with a suitable diagram, why the Mahalanobis distance might be used in preference to the Euclidian distance in classification problems.

The four parts carry equal marks.

4 Linear and non-linear classifiers.

a With reference to the two class example shown below, explain the difference between the Principal Component Analysis (PCA) transformation and the Linear Discriminant Analysis (LDA) transformation.



Without calculating numeric values, copy the figure above and show the approximate directions of the LDA and PCA axes.

b Support vector machines are a radically different classification method. Explain how a support vector machine works for a linearly separable pair of classes. To illustrate your answer copy the figure of part a above and show the approximate position of the separating hyperplane and the positions of the support vectors.

c In cases where the classes are not linearly separable, support vector machines (and other classifiers) make use of a higher dimension space. In order to do this it is necessary to find a space whose dimension is greater than $n - 1$ where n is the total number of data points.

i) With the aid of suitable diagrams show how, in all 2D two class problems with three or less points, the classes can be separated by a linear classifier.

ii) Give an example of a 2D two class problem with four points in which the classes cannot be separated linearly.

d Support vector machines work by replacing the dot product in the linear class boundary: $\mathbf{w} \cdot \mathbf{x} + b = 0$, by a kernel function which has the same effect as separating the classes in a higher dimension space: $K(\mathbf{w}, \mathbf{x}) + b = 0$. $\mathbf{x}$ are the variables and $\mathbf{w}$ is the parameter vector to be estimated. One quadratic kernel is defined as:

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v})^2$$

Given that u and v are two dimensional vectors find a mapping function ϕ such that $(\mathbf{u} \cdot \mathbf{v})^2 = \phi(\mathbf{u}) \cdot \phi(\mathbf{v})$ that has the effect of lifting a data point from two to three dimensions.

e What are the advantages and disadvantages of non-linear support vector machines when compared to the use of the Linear Discriminant Analysis (LDA) method.

The five parts carry equal marks.