IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2010

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
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

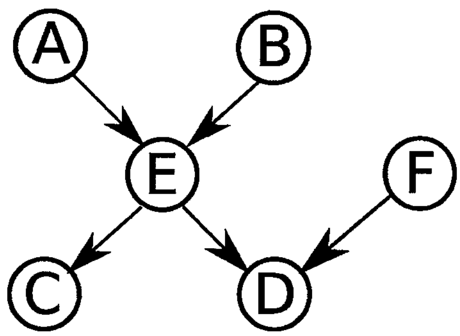
Friday 30 April 2010, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 Probability Propagation

The following questions relate to propagating probabilities in the the network below in which every variable has just two states.



The prior probabilities, denoted, for example, $P(A) = [a_1, a_2]$ are:

$$P(A) = [0.3, 0.7]\ ,\ P(B) = [0.2, 0.8]\ ,\ P(F) = [0.9, 0.1]$$

The conditional probabilities are given in the form

$$P(E|A\&B) = \begin{bmatrix} P(e_1|a_1\&b_1) & P(e_1|a_1\&b_2) & P(e_1|a_2\&b_1) & P(e_1|a_2\&b_2) \\ P(e_2|a_1\&b_1) & P(e_2|a_1\&b_2) & P(e_2|a_2\&b_1) & P(e_2|a_2\&b_2) \end{bmatrix}$$

and are:

$$P(E|A\&B) = \begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.7 \\ 0.9 & 0.7 & 0.6 & 0.3 \end{bmatrix} \quad P(D|E\&F) = \begin{bmatrix} 0.5 & 0.3 & 0.5 & 0.2 \\ 0.5 & 0.7 & 0.5 & 0.8 \end{bmatrix}$$

$$P(C|E) = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}$$

Pearl's operating equations for probability propagation are given at the end of the paper.

a. Following the propagation of π evidence during initialisation what is the initial posterior probability distribution over variable C?

b. If C is instantiated to c_1 and F is instantiated to f_2 compute the posterior probability distribution over A.

c. If, in addition to the instantiations given in part (b), node D is now instantiated to state d_1, what is the new probability distribution over variable E?

d. It is discovered that for one state of variable E the variables C and D show a strong correlation. Explain briefly why this is undesirable for making inferences about the probability distribution over nodes A and B.

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.

2 The MDL Accuracy Network

The MDL measure is to be used to determine whether the arc ED in the Bayesian network shown in question 1 can be eliminated. The data set on which this decision is to be based contains the following eight points:

a_1	b_2	c_2	d_1	e_2	f_2
a_2	b_1	c_1	d_1	e_1	f_2
a_2	b_1	c_2	d_1	e_1	f_2
a_1	b_1	c_1	d_1	e_2	f_1
a_2	b_2	c_2	d_2	e_1	f_1
a_1	b_1	c_2	d_2	e_2	f_1
a_2	b_2	c_1	d_2	e_1	f_2
a_1	b_2	c_2	d_2	e_2	f_2

a Calculate the MDL size of the model given in question 1, and the MDL size of the model with the arc ED removed.

b Using the matrix $P(D|E\&F)$, and the assumption that D is independent from E calculate the conditional probability matrix $P(D|F)$ in the network with arc ED removed.

c Calculate the model accuracy ($P(Network|Data)$) for the original network shown in question 1.

d Calculate the model accuracy for the network with arc ED removed, noting that the only change in the calculation is that $P(D|E\&F)$ has been replaced by $P(D|F)$. Hence deduce which network yields the best MDL score.

e If the data set were quadrupled in size to 32 points, but its probability distribution remained the same, what would be the new MDL scores for the two networks.

The five parts carry equal marks.

3 Mixture Models

A mixture model is one in which a probability distribution is represented as a weighted sum of Gaussian distributions. It is written as:

$$p(x|\theta) = \sum_{j=1}^{M} \alpha_j p_j(x|\theta_j)$$

where there are M probability distributions in the mixture and α_i are the mixing weights that have the property $\sum_{j=1}^{M} \alpha_j = 1$, and the symbol θ_j denotes the vector of unknown parameters that defines each individual distribution. Given that we wish to model a data set with N points, our objective is find the mixture parameters that will maximise the log likelihood of the data.

a Explain, with the aid of a suitable diagram what is meant by overfitting of a mixture model.

b Write down an expression for the log likelihood of the mixture defined above. Explain why the log liklihood is preferred to the likelihood.

c The most usual approach to estimating mixture models is to use the Expectation-Maximisation (EM) algorithm. This algorithm begins by selecting random values for the unknown distributions $(\alpha_j, \mu_j, \Sigma_j)$ and then iteratively updates them until convergence.

Explain carefully what is calculated during the E (Expectation) step of the algorithm.

d Explain carefully what is calculated during the M step of the algorithm.

The four parts carry equal marks.

4 Pattern classification using PCA and LDA

Let an $N \times n$ data matrix D be composed of N observations (rows) with n variables (columns).

a Explain how a mean centered data matrix U can be computed from D, and how the $n \times n$ variable co-variance matrix Σ can be computed from U.

b Suppose that Σ_p is a pooled covariance matrix made up for a small sample size problem in which there are g classes and in total N observations of the n variables. A PCA projection matrix Φ is computed by finding the eigenvectors of Σ_p. It is written as $\Phi = [\phi_1, \phi_2, ..\phi_m]$ where the ϕ_i elements are the eignevectors in column format. Note that, for each of the sample groups, Σ_i is computed from N_i observations and so Σ_i can have at most $N_i - 1$ non-zero real eigenvalues.

What is the maximum number of non-zero real eigenvalues that Φ can have, and what is the corresponding dimension of Φ? Assume that all N observations are linearly independent.

c Suppose that a PCA projection matrix has its maximum rank. Suppose that U_p is the projection of the mean centered data matrix U into the PCA space. What is the matrix equation of the projection, and what is the dimension of matrix U_p?

d After performing the PCA projection, an LDA projection L is formed by computing the eigenvectors of $\Sigma_w^{-1}\Sigma_b$, where Σ_w is the within class (pooled) covariance matrix and Σ_b is the between class covariance matrix of the projected data U_p. What is the dimension of L?

e The most discriminant features matrix U_f is found by projecting U_P into the LDA space. What is the dimension of U_f.

f The original data can be reconstructed by projecting U_f back to the original n-dimensional space, forming a reconstructed data matrix R. Find the matrix equation for R.

g Are there any differences between the original data U and the reconstructed data set R?

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

Pearl's Operating equations for probability propagation

Equation 1: The λ message

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

and for the case of the single parent there is a simpler matrix form:

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

Equation 2: The π Message from A to C is given by:

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

Equation 3: The λ evidence of node C with n children $D_1, D_2, ..D_n$:

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

Equation 4: The π evidence of node C with two parents A and B:

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

This can be written in matrix form using as follows:

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

where

$$\pi_C(a_j\&b_k) = \pi_C(a_j)\pi_C(b_k)$$

Equation 5: The posterior probability of variable C:

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

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