IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2016

MEng Honours Degree in Electronic and Information Engineering Part IV
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)
MRes in High Performance Embedded and Distributed Systems
MRes 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

PAPER C493

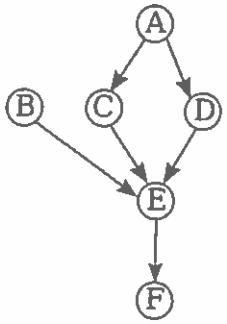
INTELLIGENT DATA AND PROBABILISTIC INFERENCE

Friday 18 March 2016, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 Probability Propagation



In this Bayesian network the variables are all binary, with, for example, variable A having states a_1 and a_2.

The conditional probabilities are given consistently in the form:

$$P(S|Q\&R) = \begin{bmatrix} P(s_1|q_1\&r_1) & P(s_1|q_1\&r_2) & P(s_1|q_2\&r_1) & P(s_1|q_2\&r_2) \\ P(s_2|q_1\&r_1) & P(s_2|q_1\&r_2) & P(s_2|q_2\&r_1) & P(s_2|q_2\&r_2) \end{bmatrix}$$

and are

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

$$P(B) = [0.4, 0.6]$$

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

$$P(F|E) = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix} \qquad P(D|A) = \begin{bmatrix} 0.4 & 0.5 \\ 0.6 & 0.5 \end{bmatrix}$$

$$P(E|B\&C\&D) = \begin{bmatrix} 1 & 0.3 & 0.2 & 1 & 1 & 0.6 & 0 & 0 \\ 0 & 0.7 & 0.8 & 0 & 0 & 0.4 & 1 & 1 \end{bmatrix}$$

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:

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

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

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

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 Using message passing, find the posterior probability of node E ($P'(E)$) in the case where node B is instantiated to b_2 and all other nodes are uninstantiated.

b Node B remains instantiated to state b_2 and node D is instantiated to state d_1. Calculate the probability of node C after probability propagation has finished.

c i) Is the probability that you calculated in part a exact or approximate? Explain your answer.

ii) Is the probability that you calculated in part b exact or approximate? Explain your answer.

d One way to ensure that probability propagation in the network always terminates would be to remove either the arc from A to C or the arc from A to D. Given the conditional probability matrices $P(C|A)$ and $P(D|A)$ are correct, which arc should be removed and why?

e An alternative to deleting an arc would be to use node clustering. Explain how this technique could be applied to the network, and what the advantages and disadvantages of node clustering are compared to arc removal.

The five parts carry equal marks.

2 Dependency Measures and Causal Directions

a The L1 dependency metric is defined as follows:

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

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 The following data set has two variables A and B each with 3 states.

$(a_1, b_1), (a_1, b_2), (a_3, b_3), (a_2, b_2), (a_1, b_1), (a_3, b_3), (a_3, b_1), (a_3, b_1)$

Find the dependency between A and B using the L1 dependency metric.

c A spanning tree found for a four variable data set has the following configuration:

The joint probability table for the triple $A - B - C$ is:

	$a_1 \& c_1$	$a_1 \& c_2$	$a_2 \& c_1$	$a_2 \& c_2$
b_1	0.25	0.1	0	0.2
b_1	0	0.2	0.25	0

Use the method of marginal indepence to determine whether any causal directions can be found for the network.

d What problems may occur in practice in determining arc directions using marginal independence?

e An alternative dependency measure, normally used in preference to the L1 metric is the Kullback Leibler divergence.

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

i) What are the properties of the Kullback Leibler that make it suitable for measuring dependency?

The five parts carry equal marks.

3 PCA and Parameter Estimation

a Consider a data set $\boldsymbol{X} \in \mathbb{R}^3$ with mean 0 and covariance matrix $\boldsymbol{C}$, and a data point $\boldsymbol{x}$, where

$$\boldsymbol{C} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \qquad \boldsymbol{x} = \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix} \in \mathbb{R}^3.$$

i) Project $\boldsymbol{x}$ onto the 1-dimensional principal subspace that minimizes the average reconstruction error.

ii) Compute the projection error, i.e., the distance between $\boldsymbol{x}$ and its projection onto the principal subspace.

b Consider the linear regression setting

$$y = \boldsymbol{\theta}^\top \boldsymbol{x} + \epsilon$$

where $\boldsymbol{x} \in \mathbb{R}^D$, $y \in \mathbb{R}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. Gaussian noise. Moreover, $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector. A (training) data set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ is given.

i) Determine the maximum likelihood estimate $\boldsymbol{\theta}_{\text{ML}}$ by optimizing the log-likelihood.

ii) What is "overfitting" and why is it a problem?

iii) Describe how MAP estimation differs from maximum likelihood estimation and why it may be more robust to overfitting than maximum likelihood.

iv) What prior would you place on the parameter vector $\boldsymbol{\theta}$ if you wanted to marginalize them out? Justify your answer.

v) Draw the directed graphical model (including all deterministic parameters) for Bayesian linear regression, i.e., a linear regression setting with a (suitable) prior on $\boldsymbol{\theta}$.
Hint: Pay attention to parameters and random variables.

The two parts carry, respectively, 30%, and 70% of the marks.

4 **Probabilistic Inference**

a i) Briefly describe what is meant by detailed balance in the design of MCMC samplers. (1–2 sentences)

ii) Why might detailed balance be a desirable property? (1–2 sentences.)

iii) Consider the distribution shown in Fig. 1. Discuss whether standard Gibbs sampling for this distribution would sample correctly from this distribution.

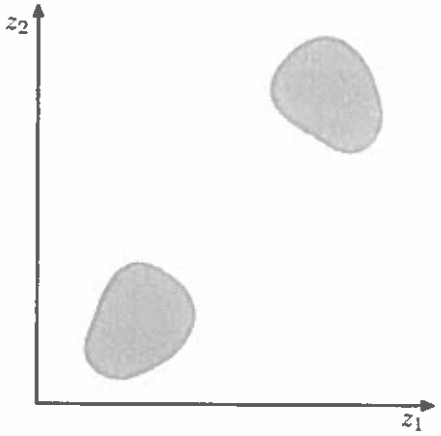


Fig. 1: A probability distribution over two variables z_1, z_2 that is uniform over the shaded regions and that is zero everywhere else.

b

You may use:

- Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

$$= \mathcal{N}(x|\mu,\tau^{-1}) = \frac{1}{\sqrt{2\pi\tau^{-1}}} \exp \frac{-(x-\mu)^2\tau}{2}$$

with $\tau^{-1} = \sigma^2$.

- Gamma distribution

$$\text{Gamma}(\tau|a,b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{b\tau}$$

i) Consider a model $p(\boldsymbol{X}, \boldsymbol{Z})$ and a factorized variational posterior $q(\boldsymbol{Z}) = \prod_{i=1}^{M} q(Z_i)$.

State the equation for computing the optimal factors $q^*(Z_i)$ that minimize $KL(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}))$

ii) You are given a model with likelihood

$$p(\boldsymbol{X}|\mu, \tau) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \tau^{-1})$$

and priors

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}),$$
$$p(\tau) = \text{Gamma}(\tau|a_0, b_0).$$

on the mean μ and the precision (inverse variance) τ.

A) Draw the corresponding directed graphical model, including all deterministic parameters.

B) We wish to find a variational posterior $q(\mu, \tau) = q(\mu)q(\tau)$. Let the optimal factors be $q^*(\mu)$ and $q^*(\tau)$. Show that $q^*(\mu)$ satisfies

$$\log q^*(\mu) = -\tfrac{1}{2}\left(\lambda_0(\mu - \mu_0)^2 + \sum_{n=1}^{N}(x_n - \mu)^2\right)\mathrm{E}[\tau] + \text{const.}$$

C) Describe briefly (1–2 sentences) why it follows that $q^*(\mu)$ is Gaussian.

D) The mean μ_N and precision λ_N (inverse variance) of $q^*(\mu)$ are given by

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N},$$
$$\lambda_N = (\lambda_0 + N)\mathrm{E}[\tau],$$

respectively, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Find the limit of μ_N, λ_N as $N \to \infty$ and explain how the variational posterior relates to the maximum likelihood estimate of μ.

c Consider two random variables x, y with joint distribution $p(x, y)$. Show that

$$\mathrm{V}_x[x] = \mathrm{E}_y\left[\mathrm{V}_x[x|y]\right] + \mathrm{V}_y\left[\mathrm{E}_x[x|y]\right].$$

where $\mathrm{V}_x[x|y]$ is the variance of x under $p(x|y)$.

Hint: You can use

$$\mathrm{E}_x[x] = \mathrm{E}_y\left[\mathrm{E}_x[x|y]\right],$$

which we have shown in Tutorial 6.

The three parts carry, respectively, 30%, 50%, and 20% of the marks.