EGT2
ENGINEERING TRIPOS PART IIA

Monday 11 May 2026   14.00 to 15.40

Module 3F8

INFERENCE

Answer not more than ***three*** *questions.*

All questions carry the same number of marks.

The ***approximate*** *percentage of marks allocated to each part of a question is indicated in the right margin.*

Write your candidate number ***not*** *your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
Engineering Data Book.

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 A scientist uses two sensors to measure the location y of an object. The audio sensor output is equal to the object's location plus audio sensor noise, $a = y + \eta_a$ where the noise is zero-mean $\mathbb{E}[\eta_a] = 0$ and has variance $\mathbb{E}[\eta_a^2] = \sigma_a^2$. The visual sensor output is equal to the object's location plus visual sensor noise $v = y + \eta_v$ where the noise is zero-mean $\mathbb{E}[\eta_v] = 0$ and has variance $\mathbb{E}[\eta_v^2] = \sigma_v^2$. The audio noise and visual noise are independent from one another. The scientist wants to use these two pieces of information, a and v, to estimate the location of the object. They propose using an estimator $\hat{y} = \alpha a + \beta v$ where α and β are scalar valued weights.

(a) (i) By considering the expected value of the estimator $\hat{y}$, propose a sensible setting for β in terms of α . [15%]

(ii) By considering the second moment of the estimator $\hat{y}$, find the value of α which results in the lowest variance estimate for y. [20%]

(iii) What happens to your solution as $\sigma_a^2 \to \infty$ and as $\sigma_a^2 \to 0$? Explain this behaviour. [15%]

(b) (i) Explain how *maximum likelihood estimation* can be used to infer an unknown quantity y from known observations x. [10%]

(ii) Relate the estimator derived in part (a) (ii) to maximum likelihood estimation in a probabilistic model. Explain any assumptions that you needed to make. [25%]

(iii) Explain how the *Bayesian* approach to inferring y differs from maximum likelihood estimation? Are there situations where Bayesian inference for y relates to the maximum-likelihood estimate? (You do not need to derive a Bayesian estimate to answer this question.) [15%]

2 A *regression problem* comprises scalar inputs x_n and scalar outputs y_n which are linearly related $y_n = mx_n + \epsilon_n$. The inputs x_n take non-negative values. The observation noise ϵ_n is Gaussian, with mean 0, but it has a variance that depends on the input $p(\epsilon_n) = \mathcal{N}(\epsilon_n; 0, \sigma^2(x_n))$ where $\sigma^2(x_n) = \exp(x_n)$. A Gaussian prior with mean μ_m and variance σ_m^2 is placed on the slope parameter so $p(m) = \mathcal{N}(m; \mu_m, \sigma_m^2)$.

(a) Roughly sketch what a typical sample from this model might look like when $\mu_m = 3$, $\sigma_m^2 = 0.01$ and the inputs are drawn i.i.d. from a uniform distribution between (0,+3). A sample includes a single draw of m and a set of input-output pairs. Label your sketch. [20%]

(b) (i) Compute the expected value of the output given an input $\mathbb{E}_{p(y_n|x_n,\mu_m,\sigma_m^2)}[y_n]$ by marginalising over m. [10%]

(ii) Similarly, compute the expected value of the product of two outputs, y_n and $y_{n'}$, given the corresponding inputs $\mathbb{E}_{p(y_n,y_{n'}|x_n,x_{n'},\mu_m,\sigma_m^2)}[y_n \ y_{n'}]$ by marginalising over m. [20%]

(iii) Explain how the results in parts (b)(i) and (b)(ii) lead to an alternative way of viewing the regression model. [10%]

(c) (i) The slope m must now be learned from a training dataset $\{x_n, y_n\}_{n=1}^N$ in a Bayesian way. Compute the posterior distribution over m after seeing N data points, that is $p(m|\{x_n, y_n\}_{n=1}^N, \mu_m, \sigma_m^2)$. [20%]

(ii) You are allowed to select an input location x at which you will be provided with an output y. Which location is most informative about the parameter m? Explain your reasoning. [20%]

The formula for the probability density of a multivariate Gaussian distribution over a variable $\mathbf{x}$ of mean μ and covariance Σ is given by

$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right).$$

3 A machine learner uses a *latent variable model* for a dataset comprising N real-valued scalar variables $\{x_n\}_{n=1}^N$. The model comprises binary latent variables $s_n \in \{0,1\}$. The prior distribution over the latent variable is uniform $p(s_n = 1) = \frac{1}{2}$. The observed data are generated from a zero-mean Gaussian distribution whose variance depends on the latent variable $p(x_n|s_n, \sigma_0^2, \sigma_1^2) = \mathcal{N}(x_n; 0, (1-s_n)\sigma_0^2 + s_n\sigma_1^2)$. The machine learner would like to use the *EM algorithm* to fit the model to the dataset.

(a) Sketch the marginal distribution over the observed data $p(x_n|\sigma_0^2, \sigma_1^2)$. [10%]

(b) Define the *E-step* of the EM algorithm. [10%]

(c) Calculate the *E-step* update for the model above, leaving your answer in a form which is suitable for implementation. [30%]

(d) What happens to your solution when $\sigma_0^2 = \sigma_1^2$? Explain this behaviour. [10%]

(e) Define the *M-step* of the EM algorithm. [10%]

(f) Calculate the *M-step* update for the model leaving your answer in a form which is suitable for implementation. [30%]

For reference, the variational free-energy for a model of observed data $\{x_n\}_{n=1}^N$, with parameters θ, and categorical latent variables $\{s_n\}_{n=1}^N$ is given by

$$\begin{aligned}\mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N) &= \sum_{n=1}^N \sum_{k=1}^K q(s_n = k) \log \frac{p(s_n = k, x_n|\theta)}{q(s_n = k)} \\ &= \sum_{n=1}^N \left[\log p(x_n|\theta) - \mathrm{KL}(q(s_n)||p(s_n|x_n, \theta))\right]\end{aligned}$$

where $q(s_n)$ is an arbitrary distribution over the categorical variable s_n.

4 (a) (i) Define a *Hidden Markov Model* (HMM) for real-valued observed data $\{y_t\}_{t=1}^T$ and discrete latent variables $\{x_t\}_{t=1}^T$. Your solution should define the *initial state probabilities*, *transition probabilities*, and the *emission distribution* of the HMM. [15%]

(ii) A machine learner has developed a non-standard HMM comprising two hidden state variables at each time step, $x_t^{(1)}$ and $x_t^{(2)}$. Each of these variables can take one of two states and they are generated by the following *bigram* models,

$$p(x_1^{(1)} = 1) = 1/2, \;\; p(x_t^{(1)} = 1|x_{t-1}^{(1)} = 1) = p(x_t^{(1)} = 2|x_{t-1}^{(1)} = 2) = 0.9$$
$$p(x_1^{(2)} = 1) = 3/4, \;\; p(x_t^{(2)} = 1|x_{t-1}^{(2)} = 1) = p(x_t^{(2)} = 2|x_{t-1}^{(2)} = 2) = 0.8$$

The observed variable y_t is a continuous scalar and is modelled using a Gaussian distribution whose mean and variance depend on the two hidden states

$$p(y_t|x_t^{(1)}, x_t^{(2)}) = \mathcal{N}\left(y_t; \mu(x_t^{(1)}, x_t^{(2)}), \sigma^2(x_t^{(1)}, x_t^{(2)})\right).$$

In this way the mean and the variances are stored in 2×2 matrices and if, say, $x_t^{(1)} = 2$ and $x_t^{(2)} = 1$ then the mean of the Gaussian is $\mu(2,1)$ and the variance is $\sigma^2(2,1)$.

Write the model above in terms of an equivalent standard HMM. Explain your reasoning. [35%]

(b) Consider a joint distribution over seven variables $x_1, x_2, \ldots, x_7$ each of which is discrete and can take one of D different values.

(i) Write an expression for computing the marginal distribution $p(x_3)$. How costly is it to compute $p(x_3)$ in the general case? [15%]

(ii) Now consider a joint distribution which factorises according to

$$p(x_{1:7}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_4, x_6)p(x_4|x_7)p(x_5|x_3)p(x_6|x_7)p(x_7).$$

Write down an efficient way of computing the marginal $p(x_3)$ in this case. [25%]

(iii) How costly is it to compute $p(x_3)$ using your solution in part (b)? How does it relate to *dynamic programming*? [10%]

END OF PAPER

THIS PAGE IS BLANK

Selected short numerical answers

2 c) ii) $x = 2$
3 d) $q_n(s_n = 1) = \frac{1}{2}$

4 a) ii) initial state probabilities $\pi = \left[\frac{3}{8}, \frac{1}{8}, \frac{3}{8}, \frac{1}{8}\right]^\top$

transition state probabilities $T = \begin{bmatrix} 0.72 & 0.18 & 0.08 & 0.02 \\ 0.18 & 0.72 & 0.02 & 0.08 \\ 0.08 & 0.02 & 0.72 & 0.18 \\ 0.02 & 0.08 & 0.18 & 0.72 \end{bmatrix}$