EGT2
ENGINEERING TRIPOS PART IIA

Wednesday 24 April 2024 2 to 3.40

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) An engineer has obtained a noisy measurement y of the potential energy in a spring which has spring constant k and extension x, that is $y = \frac{1}{2}kx^2 + \epsilon$. The measurement noise ϵ is drawn from a standard Gaussian distribution $p(\epsilon) = \mathcal{N}(\epsilon; 0, 1)$. *A priori* the spring extension is assumed to follow a Gaussian distribution, that is $p(x) = \mathcal{N}(x; 0, \sigma^2)$.

(i) Derive the *maximum a posteriori* (MAP) estimate for the extension x, starting from the definition of the MAP estimate. [40%]

(ii) What happens to the *maximum a posteriori* (MAP) estimate for the extension x as $\sigma^2 \to \infty$? Explain this behaviour and how it relates to the *maximum likelihood* estimate of x. [10%]

(b) Two noisy depth sensors measure the distance to an object an unknown distance d metres away. The depth is assumed, *a priori*, to be distributed according to a distribution $p(d)$. The depth sensors return two noisy measurements of the depth, y_1 and y_2, whose conditional distribution is denoted $p\,(y_1, y_2|d)$.

(i) Show how Bayes' rule can be used to compute the posterior distribution over the depth given the two noisy measurements, $p(d|y_1, y_2)$. [10%]

(ii) The depth can be now be assumed, *a priori*, to be distributed according to a standard Gaussian distribution $p(d) = \mathcal{N}(d; 0, 1)$ and the noisy measurements can be assumed to be distributed according to a multivariate Gaussian distribution centred on the unknown distance, that is

$$p\,(y_1, y_2|d) = \mathcal{N}\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}; \begin{bmatrix} d \\ d \end{bmatrix}, \ \Sigma\right) \quad \text{where} \quad \Sigma^{-1} = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$$

Compute the posterior distribution $p(d|y_1, y_2)$ for this case. [30%]

(iii) Compare the posterior uncertainty when $\alpha = 0$ and $1 > \alpha > 0$. Explain what is happening. [10%]

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

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

2 (a) A data scientist is fitting a linear regression model to a data set comprising N scalar outputs y_n and scalar inputs x_n. The data scientist uses a simple model in which each output is produced by multiplying the input by a scalar weight w and adding independent Gaussian noise with mean 0 and variance σ^2, that is

$$y_n = wx_n + \epsilon_n \quad \text{where} \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

(i) Write down the likelihood of the parameters w and σ^2. [10%]

(ii) Find the maximum likelihood setting for w. [20%]

(iii) The data scientist would like to apply the model to a setting where data are continuously arriving. For privacy reasons, the data scientist is not allowed to store old data indefinitely. However, it is permitted to store quantities that are averaged over old data points. Denoting a new data point as $\{x_{N+1}, y_{N+1}\}$ and the old data as $\{x_n, y_n\}_{n=1}^N$, derive a sequential algorithm which returns the maximum likelihood setting for w through incrementally updated averaged quantities. [20%]

(b) The data scientist would like to extend the regression model described in part (a). In addition to estimating the weights using maximum-likelihood estimation, they would like to also infer the noise variance σ^2 using *maximum a posteriori* estimation with a prior

$$p(\sigma^2|\alpha, \beta) = \frac{1}{Z(\alpha, \beta)} \frac{1}{(\sigma^2)^\alpha} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

where the scalar parameters of the prior distribution are $\alpha > 1$ and $\beta > 0$. $Z(\alpha, \beta)$ is a normalising constant.

(i) Derive the MAP estimate for σ^2 . [35%]

(ii) Provide an interpretation for the prior parameters α and β that will help the data scientist select suitable values for these parameters. [15%]

3 A data scientist is fitting a clustering model to a data set comprising N data points $\{\mathbf{x}_n\}_{n=1}^{N}$. Each data point is a D dimensional vector comprising binary values i.e. each element $x_{n,d} \in \{0,1\}$. In the generative model, each cluster is assumed to be equally probable *a priori*, that is the latent cluster membership variable, $s_n \in \{1,2,\ldots,K\}$, is drawn from a uniform categorical distribution $p(s_n = k) = 1/K$ where K is the total number of clusters. Each element of each observed data point is then generated from a Bernoulli distribution with a parameter determined by the cluster membership $p(x_{n,d} = 1|s_n = k, \pi_{k,d}) = \pi_{k,d}$. The data scientist would like to use the *EM algorithm* to fit the model to the data set.

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

(b) The data scientist would like to use a hard E-step in which each data point is assigned to just the most probable cluster. Discuss whether this hard E-Step has an interpretation in terms of distance minimisation similar to the assignment step of the k-means algorithm. [25%]

(c) Define the *M-step* of the EM algorithm. Calculate the *M-step* update for the model described above when using a soft E-Step, leaving your answer in a form which is suitable for implementation. [30%]

(d) If the E-Step assignment is instead a hard assignment to the most probable cluster, describe what form the M-Step takes. [15%]

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

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

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

$$\mathrm{KL}(q(s_n)||p(s_n|\mathbf{x}_n,\theta)) = \sum_{k=1}^{K} q(s_n = k) \log \frac{q(s_n = k)}{p(s_n = k|\mathbf{x}_n,\theta)}$$

4 The scalar variables $x_{1:T} = \{x_1, x_2, \ldots, x_T\}$ are distributed according to a Gaussian AR(1) process.

(a) Define the Gaussian AR(1) process mathematically. [10%]

(b) Let $z_t = \sum_{t'=1}^{t} x_{t'}$ so that $z_1 = x_1$, $z_2 = x_1 + x_2$, $z_3 = x_1 + x_2 + x_3$ and so on. Show that $z_{1:T} = \{z_1, z_2, \ldots, z_T\}$ follows a Gaussian AR process and find the parameters for this process. [40%]

(c) Define a Linear Gaussian State Space Model (LGSSM) with observed variables $\mathbf{y}_t$ and latent variables $\mathbf{s}_t$. Your answer should clearly specify the parameters of the model. [10%]

(d) At each time-step, noisy observations are made of the variables x_t and z_t defined above so that

$$\mathbf{y}_t = \begin{bmatrix} x_t \\ z_t \end{bmatrix} + \eta_t$$

The measurement noise η_t is drawn from a two dimensional multivariate Gaussian distribution with zero mean and identity covariance $\eta_t \sim \mathcal{N}(\mathbf{0}, I)$.

Explain how to rewrite this model as a LGSSM with a first order Markov hidden state. [30%]

(e) Suggest an algorithm that could be used to infer the most likely trajectory $z^*_{1:T}$ given the observations $\mathbf{y}_{1:T}$ where

$$z^*_{1:T} = \arg\max_{z_{1:T}} p(z_{1:T}|\mathbf{y}_{1:T})$$

Explain your reasoning. [10%]

END OF PAPER

THIS PAGE IS BLANK

Selected short numerical answers

1a) if $y \geq \frac{1}{k\sigma^2}$ then $x_{\text{MAP}} = \pm\sqrt{\frac{2y}{k} - \frac{2}{k^2\sigma^2}}$ else $x_{\text{MAP}} = 0$
1b) ii) $\sigma^2_{d|y_1,y_2} = \frac{1}{3+2\alpha}$ and $\mu_{d|y_1,y_2} = \frac{1}{3+2\alpha}(1+\alpha)(y_1+y_2)$
2b) i) $\sigma^2_{\text{MAP}} = \frac{\beta+\frac{1}{2}\sum_{n=1}^{N}(y_n - w_{\text{MAP}}x_n)^2}{\alpha+N/2}$
4b) $z_t = (1+\lambda)z_{t-1} - \lambda z_{t-2} + \sigma\epsilon_t$