EGT2
ENGINEERING TRIPOS PART IIA

Wednesday 28 April 2021 13.30 to 15.10

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 and at the top of each answer sheet.*

STATIONERY REQUIREMENTS
Write on single-sided paper.

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
You are allowed access to the electronic version of the Engineering Data Books.

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

The time taken for scanning/uploading answers is 15 minutes.

Your script is to be uploaded as a single consolidated pdf containing all answers.

1 A climate scientist would like to characterise the variability of daily temperature measurements. They take a set of N scalar temperature measurements $\{x_n\}_{n=1}^N$ which have been centred so that they have zero mean. They model the temperature measurements as independent draws from a zero mean Gaussian with unknown variance σ^2, so that $p(x_n|\sigma^2) = \mathcal{N}(x_n;\ 0,\ \sigma^2)$. They place a prior over the unknown variance

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

The prior is a valid probability density over the variance with parameters α and β (which are positive scalars) and $Z(\alpha, \beta)$ is the normalising constant.

(a) Compute the posterior distribution over the variance parameter $p(\sigma^2|\{x_n\}_{n=1}^N)$ taking care to leave your answer in a simple form. Provide an intuitive interpretation for the parameters of the prior: α and β. [30%]

(b) The climate scientist would now like to compute a point estimate for the unknown variance parameter using the same model described in the previous question.

(i) Define the *maximum a posteriori* (MAP) estimate and the *maximum likelihood* estimate of the unobserved parameter σ^2 in terms of probability distributions. Comment on the similarities and differences between the definitions of the two estimators. [20%]

(ii) Compute the MAP estimate of the parameter σ^2. [20%]

(iii) When will the MAP estimate of σ^2 be identical to the maximum likelihood estimate? [15%]

(iv) How might the climate scientist quantify the uncertainty in the estimate of the parameter σ^2? [15%]

Here, and later in the exam, we have used the following notation to indicate univariate Gaussian distributions:

$$\mathcal{N}(z;\ \mu,\ \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right).$$

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 observation noise is Gaussian, with mean 0, but it has a variance that depends on the input $p(\epsilon_n) = \mathcal{N}(\epsilon_n;\ 0,\ 1 + x_n^4)$. A standard Gaussian prior is placed on the slope parameter so $p(m) = \mathcal{N}(m;\ 0,\ 1)$.

The slope m must be learned from a training dataset $\{x_n, y_n\}_{n=1}^N$ in a Bayesian way.

(a) Compute the posterior distribution over m after seeing N data points, $\{x_n, y_n\}_{n=1}^N$, that is $p(m|\{x_n, y_n\}_{n=1}^N)$. [30%]

(b) Compute the posterior distribution over m for the following datasets:

(i) A dataset comprising $N = 1$ data point $x_1 = 0,\ y_1 = 100$

(ii) A dataset comprising $N = 2$ data points $x_1 = -1,\ y_1 = 3$ and $x_2 = 1,\ y_2 = -3$

Provide intuitive explanations for these results. [30%]

(c) You are allowed to select an input location x at which you will be provided with an output y. Which locations are most informative about the parameter m? Explain your reasoning. [40%]

3 A physicist measures radioactive decay events in a detector. A source is located at $x = 0$ and the distance that decay events take place from the source is measured by the detector and denoted x_n. (The detector can be assumed to be infinitely large for the purposes of this question i.e. there is no upper limit on the size of x_n which can be measured.)

The source emits two types of radioactive particle (denoted $s_n = 0$ and $s_n = 1$). The probability of emitting particle type $s_n = 1$ is $p(s_n = 1|\rho) = \rho$. The decay events from each type of particle are given by exponential distributions with decay constants that depend on the particle type, denoted λ_0 and λ_1, that is

$$p(x_n|s_n = k, \lambda_0, \lambda_1) = \frac{1}{\lambda_k} \exp(-x_n/\lambda_k) \quad \text{for} \quad k \in \{0, 1\}$$

The physicist would like to use the *EM algorithm* to learn the decay constants (λ_0 and λ_1) and emission probabilities (ρ) from a dataset of N decay measurements $\{x_n\}_{n=1}^N$.

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

(b) Define the *M-step* of the EM algorithm. Calculate this update for the model above, leaving your answer in a form which is suitable for implementation. [50%]

(c) Compute the probability of the decay events given the model parameters, $p(x_n|\rho, \lambda_0, \lambda_1)$. Explain how this quantity relates to the EM algorithm. [20%]

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

$$\mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N) = \sum_{n=1}^{N} \sum_{k=0}^{1} q(s_n = k) \log \frac{p(s_n = k, x_n|\theta)}{q(s_n = k)}$$

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

4 (a) Two sequences $y_{1:T}^{(1)}$ and $y_{1:T}^{(2)}$ are generated from the same bigram model,

$$y_{1:T}^{(1)} = \{A, A, A, A, A, B, B, C, A, A, A, A, A, A, B, B\}$$
$$y_{1:T}^{(2)} = \{B, A, A, A, B, C, A, A, A, A, B, B, B, A, A, B\}.$$

(i) Write down the maximum-likelihood parameters for the bigram model for these data. You do not need to derive the maximum likelihood estimates from first principles. Draw a *state transition diagram* to illustrate your solution. [35%]

(ii) A third sequence from the same model is observed and used as held-out data to evaluate the maximum-likelihood trained model

$$y_{1:T}^{(3)} = \{A, B, A, A, A, A, A, C, A, A, A, A, B, B, A\}.$$

Compute the probability of the observed sequence under the trained model. Describe how the training method could be altered to improve the performance of the trained model on the held-out sequence. [15%]

(b) A parking sensor on a car emits ultra-sonic pulses at regular time intervals $t = 1, 2, 3, \ldots$ and a receiver measures the time it takes for the pulses to travel to a nearby object and be reflected back. Each travel-time, y_t, is related to the distance between the sensor and the object x_t by the speed of sound, c, with a factor of two accounting for the fact that the pulse must travel to the object and back. The sensor is noisy and is well approximated by a Gaussian with variance σ_y^2, that is $p(y_t|x_t) = \mathcal{N}(y_t;\ 2x_t/c,\ \sigma_y^2)$. The distance to the object is assumed to vary slowly over time which is approximated by a Gaussian first order auto-regressive model, $p(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \lambda x_{t-1},\ \sigma^2)$.

(i) What algorithm would be appropriate for estimating the current distance to the object at time t, that is x_t, given a sequence of observed travel-times $y_{1:t}$. Explain your reasoning. [10%]

(ii) The sample rate of the sensor has to be changed. Rather than sampling at each time $t = 1, 2, 3, \ldots$ it now samples at half the rate corresponding to times $t = 1, 3, 5, \ldots$ instead. Convert the original model for the higher sample rate into a new model which is appropriate for the lower sample rate. Explain your reasoning, including how the parameters of the new model relate to those in the old model. [40%]

END OF PAPER

THIS PAGE IS BLANK