EGT2
ENGINEERING TRIPOS PART IIA

Monday 12 May 2025 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) A data scientist has a dataset in which each data point belongs to one of K classes denoted $y \in \{1,2,\ldots,K\}$. She models observed data x using a probability density which depends on the latent class variable $p(x|y = k)$. Denote the prior probability of each class as $p(y = k)$.

(i) Use *Bayes' rule* to compute the posterior distribution over the latent class variable given the observed data, that is $p(y = k|x)$. [10%]

(ii) The data scientist must return a point estimate of the class y to their client. The client provides the data scientist with a reward function $R(y, \hat{y})$ that indicates their satisfaction with a point estimate $\hat{y}$ when the true state of the variable is y. Explain how to use *Bayesian Decision Theory* to determine the optimal point estimate, $\hat{y}$. [20%]

(iii) Compute the optimal point estimate $\hat{y}$ when the reward is

$$R(y = k, \hat{y} = k') = \begin{cases} 0 & \text{when } k' = k \\ -1 & \text{when } k' \neq k \end{cases}$$

i.e. the reward is zero if the point estimate is correct and -1 if it is incorrect. [15%]

(b) Consider a forecasting model with parameters θ that takes in an initial state x_0 as input and forecasts the state t steps into the future $f_\theta(x_0) \approx x_t$.
The forecasting model has been trained to minimise the average square error on a dataset of initial and final state pairs $\left\{x_0^{(n)},\ x_t^{(n)}\right\}_{n=1}^{N}$, that is

$$\arg\min_{\theta} \ \frac{1}{N}\sum_{n=1}^{N}\left(f_\theta(x_0^{(n)}) - x_t^{(n)}\right)^2 \quad \text{where} \ \{x_0^{(n)},\ x_t^{(n)}\} \sim p(x_0, x_t).$$

Here $p(x_0, x_t)$ is the underlying joint distribution of the data points.

(i) What function $f_\theta(x_0)$ minimises the training loss in the limit of large data $N \to \infty$? Justify your answer mathematically. [40%]

(ii) In practice, what factors could prevent this optimal solution being reached? [15%]

2 A data scientist would like to summarise a complex distribution $p(x)$ with a simpler distribution $q(x)$ by minimising the *Kullback-Leibler (KL) divergence*

$$\mathrm{KL}(p(x)||q(x)) = \int p(x) \log \frac{p(x)}{q(x)}\, \mathrm{d}x.$$

(a) Name three key properties of the KL divergence. [15%]

(b) Consider the case where x is a real valued scalar and the approximating distribution is a univariate Gaussian distribution $q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2)$. The complicated distribution $p(x)$ is non-Gaussian and has mean μ_p, variance σ_p^2, and differential entropy $H(p(x))$ where

$$\mu_p = \int x\, p(x)\, \mathrm{d}x, \;\; \sigma_p^2 = \int (x - \mu_p)^2\, p(x)\, \mathrm{d}x, \;\; \text{and} \;\; H(p(x)) = -\int p(x) \log p(x)\, \mathrm{d}x.$$

Compute the KL divergence between $p(x)$ and $q(x)$ in terms of the means (μ_p and μ_q) and variances (σ_p^2 and σ_q^2) of the two distributions, and the entropy $H(p(x))$. [30%]

(c) The data scientist has a true distribution $p(x)$ which is a mixture of one dimensional Gaussians. He would like to use the univariate Gaussian distribution $q(x)$ to approximate the mixture.

(i) Define the mixture of Gaussians model for $p(x)$ mathematically. [15%]

(ii) Using your answer to part (b), compute the optimal form for the parameters of the approximation, μ_q and σ_q^2, in terms of the parameters of the mixture model. [25%]

(iii) Is this version of the KL divergence mode-seeking or mode-covering? Justify your answer with an example. [15%]

The formula for the probability density of a Gaussian distribution over a scalar variable x with mean μ and variance σ^2 is given by

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

3 A regression problem involves inputs x_n and outputs y_n. Both inputs and outputs are scalar and real valued. A machine learner models the regression problem using a latent variable model. A binary latent variable s_n controls the slope of the linear trend for each data point: if $s_n = 1$ then $y_n = m_1 x_n + \eta_n$ and if $s_n = 0$ then $y_n = m_0 x_n + \eta_n$. The observation noise is drawn from a standard Gaussian distribution, $\eta_n \sim \mathcal{N}(0, 1)$. The prior distribution over the latent variable is uniform $p(s_n = 1) = \frac{1}{2}$. The machine learner would like to use the *EM algorithm* to fit the model to a dataset $\{x_n, y_n\}_{n=1}^{N}$.

(a) Define the *E-step* of the EM algorithm. [15%]

(b) Calculate the *E-step* update for the model above, leaving your answer in a form which is suitable for implementation. [20%]

(c) Relate the *E-step* for this model to logistic regression. [10%]

(d) Define the *M-step* of the EM algorithm. [15%]

(e) Calculate the *M-step* update for the model described above, leaving your answer in a form which is suitable for implementation. [20%]

(f) The machine learner would like to use a gradient based update for the *M-step* instead. Write down such an update. Can the new EM algorithm be made convergent in this case? [20%]

For reference the variational free-energy for a model with inputs x_n and outputs y_n with parameters θ and categorical latent variables s_n is given by

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

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

4 A *Hidden Markov Model* (HMM) has a binary latent state x_t and a scalar real valued observed state y_t.

The latent state has initial distribution $p(x_1 = 1) = \frac{1}{3}$ and the transition distribution has $p(x_t = 1|x_{t-1} = 1) = \frac{3}{4}$ and $p(x_t = 0|x_{t-1} = 0) = \frac{1}{3}$ for $t = 2 \ldots T$. The emission distributions are given by Laplace's distribution with a mean that depends on the hidden state: $p(y_t|x_t = 0) = \frac{1}{2}\exp(-|y_t + 1|)$ and $p(y_t|x_t = 1) = \frac{1}{2}\ \exp(-|y_t - 1|)$.

(a) Compute the marginal joint density over the first two observed variables $p(y_1, y_2)$ and relate this density to mixture models. [15%]

(b) Sketch a contour plot of $p(y_1, y_2)$ as a function of y_1 and y_2, labelling key aspects. [20%]

(c) The first observed state takes a value $y_1 = 1$. Compute the predictive distribution for the next latent state $p(x_2|y_1 = 1)$. [25%]

(d) How would $p(x_2|y_1 = 1)$ be used in the *forward algorithm*? [10%]

(e) Compute the predictive distribution for any future latent state $p(x_t|y_1 = 1)$ for $t \geq 2$. [15%]

(f) As $t \to \infty$ to what value does $p(x_t = 1|y_1 = 1)$ converge? Explain your reasoning. [15%]

END OF PAPER

THIS PAGE IS BLANK

Selected short relevant and numerical answers

1a) iii) $\hat{y} = \arg\max_k p(y = k|x)$ i.e. the MAP estimate
1b) i) $f_\theta(x_0) = \mathbb{E}_{p(x_t|x_0)}(x_t)$

4 c) $p(x_2 = 0|y_1 = 1) \approx 0.27$
4 f) $p(x_\infty = 0|y_1 = 1) = 3/11$