Version RET/2

EGT2
ENGINEERING TRIPOS PART IIA

2025

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 $p(x|y = k)$ and 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 $R(\hat{y}, y)$ 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 $\hat{y}$ in the case where the reward is zero if the point estimate is correct $R(y = k, \hat{y} = k) = 0$ and -1 if it is incorrect, $R(y = k, \hat{y} = k') = -1 \; \forall \; k \neq k'$. [15%]

a i) $$p(y=k|x) = \frac{p(x|y=k)p(y=k)}{\sum_{k'=1}^{K} p(x|y=k)p(y=k')}$$

ii) $$\hat{y} = \underset{k'}{\arg\max} \sum_{k=1}^{K} p(y=k|x) R(y=k, \hat{y}=k')$$

iii) $$\hat{y} = \underset{k'}{\arg\max} -\sum_{k\neq k'} p(y=k|x)$$

$\Rightarrow$ pick k' so that the probability of not getting k' is smallest

$\Rightarrow$ pick MAP estimate

(cont.

$$k' = \underset{k}{\arg\max}\, p(y=k|x)$$

(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

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

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%]

b i) $$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N}\left(f_\theta(x_0^{(n)}) - x_t^{(n)}\right)^2 = \mathbb{E}_{p(x_0,x_t)}\left[\left(f_\theta(x_0) - x_t\right)^2\right]$$

$$= \mathbb{E}_{p(x_0)}\mathbb{E}_{p(x_t|x_0)}\left[\left(f_\theta(x_0) - x_t\right)^2\right]$$

$$= \mathbb{E}_{p(x_0)}\left[\left(f_\theta(x_0) - \mu_{t|0}\right)^2 + \underbrace{\sigma^2_{t|0}}_{\text{does not depend on }\theta}\right]$$

where $\mu_{t|0} = \mathbb{E}_{p(x_t|x_0)}(x_t)$ $\quad \sigma^2_{t|0} = \mathbb{E}_{p(x_t|x_0)}\left[\left(x_t - \mu_{t|0}\right)^2\right]$

$\therefore\ f_{\theta_{\text{optimal}}}(x_0) = \mu_{t|0}$ = predictive mean

ii) finite data, sub optimal model $f_\theta(x_0)$ i.e. not expressive enough, imperfect optimisation (local optima, minibatches)

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. [20%]

(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 complex distribution $p(x)$ is non-Gaussian and has mean μ_p, variance σ_p^2, and differential entropy $H(p(x))$ where

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

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%]

a) $\mathrm{KL}(p \| q) \geq 0$ equality iff $p(x) = q(x)$ $\forall x$, measures similarity between p and q

b) $\mathrm{KL}(p \| q) = \mathbb{E}_{p(x)}\left[\log \frac{p(x)}{q(x)}\right] = \mathbb{E}_{p(x)}\left[\log p(x)\right] - \mathbb{E}_{p(x)}\left[\log \mathcal{N}(x; \mu_q, \sigma_q^2)\right]$

$= -H[p(x)] + \frac{1}{2}\log 2\pi\sigma_q^2 + \mathbb{E}_{p(x)}\left[\frac{1}{2\sigma_q^2}(x-\mu_q)^2\right]$

$= -H[p(x)] + \frac{1}{2}\log 2\pi\sigma_q^2 + \frac{1}{2\sigma_q^2}(\mu_p - \mu_q)^2 + \frac{1}{2}\frac{\sigma_p^2}{\sigma_q^2}$

(c) The data scientist has a true distribution $p(x)$ which is a mixture of one dimensional Gaussians. She 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. [20%]

(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

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

c i) $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2)$

ii) Clear from (b) that optimal $\mu_q = \mu_p$

lets find optimal σ_q^2 when $\mu_p = \mu_q$

$$\frac{d}{d\sigma_q^2} KL(p\|q) = \frac{1}{2}\frac{1}{\sigma_q^2} - \frac{1}{2}\frac{1}{\sigma_q^4}\sigma_p^2 \Rightarrow \sigma_q^2 = \sigma_p^2$$

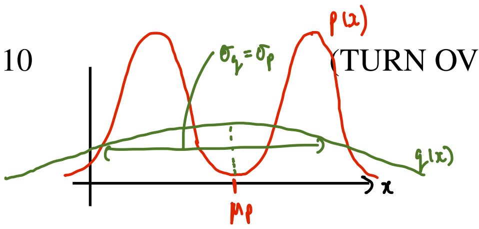
i.e. the KL matches the first & second moments of q to p

$$\mu_q = \mu_p = \int x\left(\sum_k \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2)\right)dx = \sum_k \pi_k \mu_k$$

$$\sigma_q^2 = \sigma_p^2 = \int (x-\mu_p)^2 \left(\sum_k \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2)\right)dx$$

$$= \sum_k \pi_k (\sigma_k^2 + \mu_k^2) - \mu_p^2$$

iii) Consider when p has two separated modes



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. [20%]

(b) Calculate the *E-step* update for the model above, leaving your answer in a form which is suitable for implementation. [15%]

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

a) E-Step : $q^{(new)}(s_n) = \underset{q(s_n)}{\arg\max}\ \mathcal{F}(\theta, \{q(s_n)\}_{n=1}^{N}) \Rightarrow q(s_n=1) = p(s_n=1|y_n,\theta)$

b)
$$q(s_n=1) = \frac{p(y_n|s_n=1,\theta)\,p(s_n=1|\theta)}{p(y_n|s_n=1,\theta)\,p(s_n=1|\theta) + p(y_n|s_n=0,\theta)\,p(s_n=0|\theta)} = \frac{N(y_n; m_1 x_n, 1)\times 1/2}{\frac{1}{2}N(y_n; m_1 x_n, 1) + \frac{1}{2}N(y_n; m_0 x_n, 1)}$$

$$= \frac{1}{1 + \frac{N(y_n; m_0 x_n, 1)}{N(y_n; m_1 x_n, 1)}} = \frac{1}{1 + e^{-x_n y_n (m_1 - m_0) - \frac{1}{2}x_n^2 (m_0^2 - m_1^2)}}$$

c) = logistic regression with basis functions $[x_n y_n,\ x_n^2]$ & weights $[m_1 - m_0,\ \frac{m_0^2 - m_1^2}{2}]$

(d) Define the *M-step* of the EM algorithm. [20%]

(e) Calculate the *M-step* update for the model described in the previous question, leaving your answer in a form which is suitable for implementation. [15%]

(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%]

d) $$\theta^{(new)} = \underset{\theta}{\text{argmax}} \; \mathcal{F}\left(\theta, \{q(s_n)\}_{n=1}^{N}\right)$$

$$\{m_0^{(new)}, m_1^{(new)}\} = \underset{\{m_0, m_1\}}{\text{arg max}} \sum_{n=1}^{N} \sum_{k=0}^{1} q(s=k) \log p(y_n | s_n = k, m_k)$$

e) $$\frac{\partial \mathcal{F}}{\partial m_0} = \frac{\partial}{\partial m_0} \sum_{n=1}^{N} \Big\{ q(s_n=0)\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}(y_n - m_0 x_n)^2\right)$$
$$+ q(s_n=1)\left[-\frac{1}{2}\log 2\pi - \frac{1}{2}(y_n - m_1 x_n)^2\right]\Big\}$$

$$= \sum_{n=1}^{N} q(s_n=0)\left[y_n - m_0 x_n\right] x_n = 0$$

$$\Rightarrow m_0 = \sum_{n=1}^{N} q(s_n=0)\, y_n x_n \Big/ \sum_{n=1}^{N} q(s_n=0)\, x_n^2$$

Similarly $$m_1 = \sum_{n=1}^{N} q(s_n=1)\, y_n x_n \Big/ \sum_{n=1}^{N} q(s_n=1)\, x_n^2$$

f)

```
$\mu_{old} = \mu$    % set old setting of parameters to current value of parameters
$\mathcal{F}_{old} = \mathcal{F}(\mu_{old})$    % evaluate current free energy
while $\mathcal{F}_{new} > \mathcal{F}_{old}$    % continue loop until free-energy decreases
    $\mu_{new} = \mu_{old} + \eta \frac{d}{d\mu} \sum_n \sum_k q(s_n = k) \log p(y_n | s_n = k, \mu_k)$
        % perform gradient ascent step on free energy
    $\mathcal{F}_{new} = \mathcal{F}_{new}(\mu_{new})$    % evaluate new free energy
endwhile

$\mu = \mu_{old}$    % update parameters to setting with highest free-energy
```

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 $p(y_t|x_t = 1) = \frac{1}{2}\ \exp(-|y_t - 1|)$.

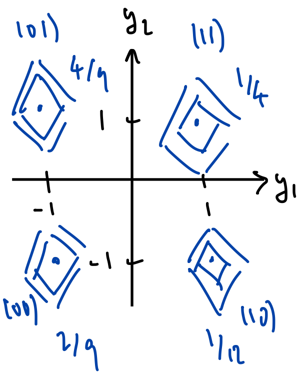
(a) Compute the marginal joint density over the first two observed variables $p(y_1, y_2)$ and relate this density to mixture models. [10%]

(b) Sketch a contour plot of $p(y_1, y_2)$ as a function of y_1 and y_2, labelling key aspects. [25%]

a) $$p(y_1, y_2) = \sum_{x_1, x_2} p(x_1)\, p(x_2|x_1)\, p(y_1|x_1)\, p(y_2|x_2)$$

$$= p(x_1=0)\, p(x_2=0|x_1=0)\, p(y_1|x_1=0) p(y_2|x_2=0)$$
$$+$$
$$p(x_1=0)\, p(x_2=1|x_1=0)\, p(y_1|x_1=0)\, p(y_2|x_2=1)$$
$$+$$
$$p(x_1=1)\, p(x_2=0|x_1=1)\, p(y_1|x_1=1)\, p(y_2|x_2=0)$$
$$+$$
$$p(x_1=1)\, p(x_2=1|x_1=1)\, p(y_1|x_1=1)\, p(y_2|x_2=1)$$

b)



$$= \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2}\, e^{-|y_1+1| - |y_2+1|} + \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{2}\, e^{-|y_1+1| - |y_2-1|}$$
$$+ \frac{1}{3} \cdot \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}\, e^{-|y_1-1| - |y_2+1|} + \frac{1}{3} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{2} e^{-|y_1-1| - |y_2-1|}$$

$$= \underbrace{\frac{2}{9}}_{\pi_{00}} \cdot \frac{1}{4}\, e^{-|y_1+1| - |y_2+1|} + \underbrace{\frac{4}{9}}_{\pi_{01}} \cdot \frac{1}{4}\, e^{-|y_1+1| - |y_2-1|} + \underbrace{\frac{1}{12}}_{\pi_{10}} \cdot \frac{1}{4}\, e^{-|y_1-1| - |y_2+1|}$$

$\mu_{00} = (-1,-1)$ $\mu_{01} = (-1,1)$ $\mu_{10} = (1,-1)$

(cont.

$$+ \underbrace{\frac{1}{4}}_{\pi_{11}} \cdot \frac{1}{4}\, e^{-|y_1-1| - |y_2-1|}$$

$\mu_{11} = (1,1)$

mixture of bivariate Laplace distributions with means & mixing proportions as above

(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 $p(x_t|y_1 = 1)$ converge to? Explain your reasoning. [15%]

END OF PAPER

c) $p(x_1=0|y_1=1) = p(y_1=1|x_1=0)p(x_1=0) \,/ p(y_1)$

$$= \frac{\frac{1}{2}e^{-2} \times 2/3}{1/2\, e^{-2} \times 2/3 + 1/2\, e^{0} \times 1/3} = \frac{1}{1+\frac{1}{2}e^2}$$

$$p(x_1=1|y_1=1) = 1 - p(x_1=0|y_1=1) = \frac{\frac{1}{2}e^2}{1+\frac{1}{2}e^2}$$

So
$$\begin{bmatrix} p(x_2=0|y_1=1) \\ p(x_2=1|y_1=1) \end{bmatrix} = \begin{pmatrix} 1/3 & 1/4 \\ 2/3 & 3/4 \end{pmatrix} \begin{pmatrix} 1 \\ 1/2\, e^2 \end{pmatrix} \frac{1}{1+1/2\, e^2}$$

$$= \begin{pmatrix} 1/3 + 1/8\, e^2 \\ 2/3 + 3/8\, e^2 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{1+\frac{1}{2}e^2} \end{pmatrix} \simeq \begin{pmatrix} 0.27 \\ 0.73 \end{pmatrix}$$

Forwards:

d) $p(x_2|y_1=1, y_2) \propto \overset{\text{part (c) used here}}{\overset{\downarrow}{p(x_2|y_1=1)}} \times \overset{\text{combine with likelihood}}{\overset{\downarrow}{p(y_2|x_2)}}$

e)
$$\begin{bmatrix} p(x_t=0|y_1=0) \\ p(x_t=1|y_1=0) \end{bmatrix} = \begin{pmatrix} 1/3 & 1/4 \\ 2/3 & 3/4 \end{pmatrix}^{t-1} \begin{pmatrix} 1 \\ 1/2\, e^2 \end{pmatrix} \frac{1}{1+\frac{1}{2}e^2}$$

PTO

Stationary distribution

(g) $$\begin{pmatrix} 1/3 & 1/4 \\ 2/3 & 3/4 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$ eigenvector with eigenvalue = 1

$$\Rightarrow \quad 1/3\, a + 1/4\, b = a \quad \Rightarrow \quad b = 8/3\, a$$

$$\Rightarrow \begin{bmatrix} p(x_\infty = 0 \mid y_1 = 1) \\ p(x_\infty = 1 \mid y_1 = 1) \end{bmatrix} = \frac{1}{11}\begin{pmatrix} 3 \\ 8 \end{pmatrix}$$

THIS PAGE IS BLANK

ASSESSOR'S REPORT
ENGINEERING TRIPOS PART IIA 2024/25
ASSESSOR'S REPORT
MODULE 3F8: INFERENCE

The examination was taken by 126 candidates in total. The raw marks (from those who had taken IB) had an average of 64.4% and standard deviation 14.0% with the top at 95% and bottom at 33%.

Q1 Fundamental Inference Concepts

91 attempts, Ave. raw mark 11.8/20, St.Dev. 2.9, Maximum 20, Minimum 7.

Quite a popular question, but the hardest on the exam. Part (a) was generally well-answered. However, many candidates struggled with part (b) where the key was to write the objective as an expectation over two variables and then use the product rule, or results from the decision theory part of the course, to split this into an average over a conditional mean.

Q2 The KL divergence

117 attempts, Ave. raw mark 13.7/20, St.Dev. 3.5, Maximum 20, Minimum 6.

Generally very well answered. All parts were very well done, except for computing the variance of a mixture of Gaussians which many people wrongly assumed is given by a sum of the variances of the individual components weighted by the mixing proportions — consider two very thin Gaussian components separated by a large distance: the mixture's variance will be large in this case.

Q3 The EM Algorithm

114 attempts, Ave. raw mark 13.1/20, St.Dev. 2.9, Maximum 19, Minimum 6.

This question is on a challenging topic, but it was well answered. In part (f) no one realised that a convergent gradient based E-step can be made by evaluating the free-energy function after each update, and, if the free-energy does not improve, rejecting the update, reducing the step size, and trying again.

Q4 Discrete Hidden Markov Models

57 attempts, Ave. raw mark 12.5/20, Stan. Dev. 4.3, Maximum 20, Minimum 5.

This question was conceptually quite simple, but involved taking care with working and many candidates made small mistakes (these were not harshly penalised though). I was surprised that many candidates could not produce a reasonable sketch of the mixture distribution in part (b). Many produced 1D plots, rather than contour plots, and many did not realise that the four mixture components are at [+/-1,+/-1].

R. E. Turner (Principal Assessor)
26/5/2025