Version RET/3

EGT2
ENGINEERING TRIPOS PART IIA

Wednesday 27 April 2022 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 group of K hospitals each have access to their own private observed data x_k where $k = 1, \ldots, K$. Each hospital has individually used the same parametric model $p(x_k|\theta)$ with the same prior over parameters $p(\theta)$ to infer the parameters θ using Bayes' rule, forming local posteriors $p(\theta|x_1)$, $p(\theta|x_2)$, $\ldots$, $p(\theta|x_K)$.

(i) Explain how Bayes' rule has been used to compute the local posteriors. [10%]

(ii) The hospitals are not allowed to share their data directly due to privacy rules, but they are permitted to share their posteriors. They would therefore like to compute the global posterior $p(\theta|x_1, x_2 \ldots x_K)$ in terms of the local posteriors. Derive an equation for this purpose. You do not need to explicitly compute the normalising constant in your solution. [25%]

(iii) One of the hospitals has collected new test data. The hospital uses the global posterior to make predictions and compares these predictions to those obtained by only using the local posterior. The predictions obtained using the local posterior are found to be superior. Is this expected? Explain what might be happening. [20%]

(b) A diligent member of the population has been carrying out twice weekly lateral flow tests for COVID-19. They do not have symptoms, but receive a positive result from a lateral flow test. A follow-up PCR test is negative.

For the lateral flow test the probability of a negative test result (LFT = 0) given that the individual does not have COVID-19 ($C = 0$) is $P(\text{LFT} = 0|C = 0) = 0.999$. The probability of a positive lateral flow test result given the individual does have COVID-19 is $P(\text{LFT} = 1|C = 1) = 0.85$.

For the PCR test the probability of a negative test result (PCR = 0) given that the individual does not have COVID-19 is $P(\text{PCR} = 0|C = 0) = 1$. The probability of a positive PCR test result given the individual does have COVID-19 is $P(\text{PCR} = 1|C = 1) = 0.95$.

The COVID-19 prevalence in asymptomatic individuals is estimated to be 1%.

Compute the posterior probability that the asymptomatic individual has COVID-19 given the test results, that is $P(C = 1|\text{LFT} = 1, \text{PCR} = 0)$. Explain your reasoning and any assumptions you make. [45%]

2 (a) An online education platform assesses students using an exam question. Each student's answers are marked and given an integer score. The highest score possible is ten and the minimum score is zero. The platform would like to predict the score a student will get on the question based on a set of features they have obtained which describes each student. The score on the question is denoted $y \in \{0, 1, 2, 3, \ldots, 10\}$. The features, a vector of length D, are denoted $\mathbf{x}$. The platform has collected a training set comprising the features and scores from N students $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$ and would like to use this to train a model and apply it to new students.

The platform is considering using a softmax multi-class classification model to perform the prediction so that $p(y = k|\mathbf{x}) \propto \exp(\mathbf{w}_k^\top \mathbf{x})$. Here the model parameters are a set of D dimensional weights $\{\mathbf{w}_k\}_{k=0}^{10}$.

They are also considering using a Gaussian regression model $p(y|\mathbf{x}) = \mathcal{N}(y; \mathbf{w}^\top \mathbf{x}, \sigma_y^2)$. Here the parameters are a single D dimensional weight $\mathbf{w}$ and a variance parameter σ_y^2.

(i) Compare and contrast these two modelling choices, listing strengths and weaknesses. [25%]

(ii) Design your own model for the data making sure that you explain your design choices. [25%]

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

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

(b) The KL divergence between two probability densities $q(x)$ and $p(x)$ is defined as

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

Here log denotes natural logarithm i.e. log to base e.

Consider two exponential densities over a non-negative scalar x. The first has mean λ, that is $q(x) = \frac{1}{\lambda} \exp(-x/\lambda)$. The second has mean equal to 1, that is $p(x) = \exp(-x)$.

(i) Compute the KL divergence between the densities, $\mathrm{KL}(q(x)||p(x))$. [25%]

(ii) Plot the KL divergence as a function of λ and label salient aspects of the plot. [25%]

3 A supermarket chain is looking to use machine learning for automation. Specifically, in a store, customers select a number of oranges to put into a bag and then weigh the bag on weighing scales. The total weight of the bag is recorded by the supermarket, not how many oranges it contains. The supermarket charges the customer per orange and so it would like to develop a system that automatically infers the number of oranges in customer c's bag, denoted N_c, from the measured weight x_c.

Individual oranges are assumed to have a weight w distributed according to a Gaussian distribution with mean μ and unit variance $p(w|\mu) = \mathcal{N}(w; \mu, 1)$. A bag may contain up to a maximum of 10 oranges. A uniform *a priori* distribution over the number of oranges in a bag is appropriate. The weighing scales can be assumed to be noiseless and the weight of the bag is so small that it can be neglected.

(a) Compute the probability of recording a measured weight x_c given the number of oranges in the bag N_c and their average weight μ, that is $p(x_c|N_c, \mu)$. [15%]

(b) Compute the posterior distribution over the number of oranges in a bag N_c given the weight of the bag x_c, that is $p(N_c|x_c, \mu)$. [15%]

(c) The supermarket would like to learn the average weight of an orange from measurements obtained from C customers $\{x_c\}_{c=1}^{C}$ using the EM algorithm.

(i) Explain how the EM algorithm can be used to learn the average weight of an orange and what the E-step and M-steps involve. [20%]

(ii) Compute the M-step for learning the mean parameter μ. [50%]

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

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

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

4 A *Hidden Markov Model* (HMM) has a discrete hidden state variable x_t which can take one of two values. The initial state distribution and transition probabilities for the hidden state variable are given by,

$$\begin{bmatrix} p(x_1 = 1) \\ p(x_1 = 2) \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}, \quad \begin{bmatrix} p(x_t = 1|x_{t-1} = 1) & p(x_t = 1|x_{t-1} = 2) \\ p(x_t = 2|x_{t-1} = 1) & p(x_t = 2|x_{t-1} = 2) \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}.$$

The observed state variable y_t is also discrete and can take one of three values

$$\begin{bmatrix} p(y_t = 1|x_t = 1) & p(y_t = 1|x_t = 2) \\ p(y_t = 2|x_t = 1) & p(y_t = 2|x_t = 2) \\ p(y_t = 3|x_t = 1) & p(y_t = 3|x_t = 2) \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix}.$$

(a) Write down a sequence of $T = 20$ hidden state and observed state variables $\{x_t, y_t\}_{t=1}^{T}$ that might plausibly have been generated from this model. Annotate your sequence explaining how it relates to the parameters of the model. [25%]

(b) Compute $p(y_t|y_{t-1})$ in terms of $p(x_{t-1}|y_{t-1})$ for this HMM. [30%]

(c) Using your answer to part (b), or otherwise, show mathematically that this HMM is equivalent to a first order Markov Model over the observed state, i.e. a bigram model. Write down the parameters of this equivalent bigram model. [30%]

(d) Is it generally true that an HMM with discrete hidden and observed states can be written in terms of an equivalent bigram model? Explain your reasoning. [15%]

END OF PAPER

THIS PAGE IS BLANK

Numerical answers

1 (b) $p(\mathrm{C} = 1|\mathrm{LFT} = 1, \mathrm{PCR} = 0) \approx 0.3$

4 (b)

$$\begin{bmatrix} p(y_t = 1|y_{t-1} = 1) & p(y_t = 1|y_{t-1} = 2) & p(y_t = 1|y_{t-1} = 3) \\ p(y_t = 2|y_{t-1} = 1) & p(y_t = 2|y_{t-1} = 2) & p(y_t = 1|y_{t-1} = 3) \\ p(y_t = 3|y_{t-1} = 1) & p(y_t = 3|y_{t-1} = 2) & p(y_t = 1|y_{t-1} = 3) \end{bmatrix} = \begin{bmatrix} 0.45 & 0.45 & 0.05 \\ 0.45 & 0.45 & 0.05 \\ 0.1 & 0.1 & 0.9 \end{bmatrix}$$

4 (c) Transition matrix as given in 4b above. Initial state distribution:

$$\begin{bmatrix} p(y_1 = 1) \\ p(y_1 = 2) \\ p(y_1 = 3) \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.5 \end{bmatrix}$$