Version JMHL/3

EGT2
ENGINEERING TRIPOS PART IIA

Wednesday 25 April 2018 14.00 to 15.40

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.

1 (a) Explain the concept of *Maximum a Posteriori* (MAP) estimation and describe how it is used to estimate the parameters θ of a probabilistic model from data $\mathscr{D}$. [20%]

(b) Assume that you have a biased coin with probability of heads equal to ρ. The variable x indicates the result of any particular coin flip, with $x = 1$ if it was heads and $x = 0$ if it was tails. The probability of x as a function of ρ is then Bernoulli

$$p(x|\rho) = \rho^x (1-\rho)^{1-x}. \tag{1}$$

(i) You flip the coin 10 times, obtaining 3 heads. What is the likelihood function in terms of ρ given these observations? Find the maximum likelihood estimate of ρ. Justify your answers. [20%]

(ii) Your prior beliefs about ρ are given by

$$p(\rho) = \begin{cases} 2\rho & \text{if} \quad 0 \leq \rho \leq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Compute the MAP estimate of ρ given the observations from Part (b)(i). [20%]

(iii) Let us assume that you now repeatedly toss the coin until you obtain tails for the first time. Let y be the number of times that you flipped the coin until you first obtained tails. Assume that you perform 3 independent measurements of y and obtain the values $y_1 = 3$, $y_2 = 5$ and $y_3 = 4$. Write down the joint probability of y_1, y_2, y_3 and ρ and compute the MAP estimate of ρ given y_1, y_2, y_3. [20%]

(c) Assume a linear regression model in which the output variable y is given by $y = wx + \varepsilon$, where x is the input variable, ε is additive Gaussian noise with zero mean and unit variance: $p(\varepsilon) = \mathscr{N}(\varepsilon|0,1)$. y, x, ε and w are scalars. Let us assume that, given some data $\mathscr{D}$, the posterior distribution for w is

$$p(w|\mathscr{D}) = \mathscr{N}(w|m,v). \tag{3}$$

Given these current beliefs for w, what would be the density of the predictive distribution $p(y_\star|x_\star,\mathscr{D})$ for the output $y_\star$ associated with a new test input $x_\star$? Write the resulting probability density as a function of m and v and justify your answer. [20%]

2 (a) Explain how Bayesian decision theory can be used to select the best possible action a given a reward function $R(a,\theta)$ and a posterior distribution $p(\theta|\mathscr{D})$ for the unknown model parameters θ when conditioning to observed data $\mathscr{D}$. [20%]

(b) You are given a binary classification dataset $\mathscr{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, with output variables $y_n \in \{1,-1\}$ and input features $\mathbf{x}_n \in \mathbb{R}^2$. You consider describing the data using a logistic classification model in which

$$p(y_n|\mathbf{w},\mathbf{x}_n) = \frac{y_n+1}{2}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) + \frac{1-y_n}{2}(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)), \tag{4}$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic function and $\mathbf{w}$ are the model coefficients.

(i) Assume that the maximum likelihood estimate of $\mathbf{w}$ is $\hat{\mathbf{w}}_{\mathrm{MLE}} = (1,1)^{\mathrm{T}}$. Draw the border of the square S formed by the input points $\mathbf{x} = (x_1,x_2)^{\mathrm{T}} \in \mathbb{R}^2$ with $-3 \leq x_1 \leq 3$, $-3 \leq x_2 \leq 3$. Draw in S a line representing the model's decision border, that is, the points $\mathbf{x}_n \in S$ for which $p(y_n = 1|\hat{\mathbf{w}}_{\mathrm{MLE}},\mathbf{x}_n) = 0.5$. [20%]

(ii) Choose randomly 20 input points in S and assume that their output variables are sampled according to Eq. (4) with $\mathbf{w} = \hat{\mathbf{w}}_{\mathrm{MLE}}$. Aim for about 10 input points to be close to the model's decision border. Draw each chosen input point in S as a cross, X, if its output variable took value 1 and as a circle, O, if it took value -1. Include the model's decision border in your drawing. [20%]

(iii) You are worried that many y_n in $\mathscr{D}$ might not be well explained by the previous model. To address this, you consider a new robust model in which y_n is sampled uniformly from $\{1,-1\}$ with probability ε and according to Eq. (4) with probability $1-\varepsilon$. The parameters of the new robust model are $\mathbf{w}$ and ε. Write down the likelihood function for such model, that is, write down $p(y_n|\mathbf{w},\varepsilon,\mathbf{x}_n)$. [20%]

(iv) Consider the robust model from Part (b)(iii). Let us introduce $z_n \in \{0,1\}$ as a variable that indicates whether y_n was sampled uniformly from $\{1,-1\}$ (when $z_n = 1$) or according to Eq. (4) (when $z_n = 0$). Give an expression for $p(z_n = 1|y_n,\mathbf{w},\mathbf{x}_n,\varepsilon)$. When do you expect this probability to be high? Explain your answer. [20%]

3 (a) Describe how the expectation maximization (EM) algorithm can be used to obtain the maximum likelihood estimate of the parameters θ of a probabilistic model $p(x,z|\theta)$ from a dataset $\mathscr{D} = \{x_1,\ldots,x_N\}$ containing only samples of x and no corresponding samples from the latent unobserved variable z. That is, $\mathscr{D}$ is generated by sampling $(x_i,z_i) \sim p(x,z|\theta)$, $i = 1,\ldots,N$, and then throwing away $z_1,\ldots,z_N$ and keeping only $x_1,\ldots,x_N$. [20%]

(b) Two coins A and B have probability of landing heads ρ_A and ρ_B, respectively. A data scientist chooses one of the two coins uniformly at random, tosses it 10 times and counts the resulting number of heads. This process is repeated 5 times. Let $\mathbf{x} = (x_1,\ldots,x_5)^{\mathrm{T}}$, where $x_1,\ldots,x_5 \in \{0,1,2,\ldots,10\}$ denote the number of heads obtained during each series of tosses. Similarly, let $\mathbf{z} = (z_1,\ldots,z_5)^{\mathrm{T}}$, where $z_1,\ldots,z_5 \in \{0,1\}$ denote which coin was tossed in each series: $z_i = 1$ if coin A was used in the i-th series and $z_i = 0$ otherwise.

Note that the probability of obtaining heads x times in N independent tosses of a coin with probability of heads ρ is

$$p(x|\rho,N) = \frac{N!}{x!(N-x)!}\rho^x(1-\rho)^{N-x}, \quad \text{for } x = 0,\ldots,N. \tag{5}$$

Assume that the data scientist tells you $\mathbf{x}$ but not $\mathbf{z}$ and you want to estimate ρ_A and ρ_B using the EM algorithm. The free-energy is given by

$$\mathscr{F}(\rho_A,\rho_B,q_1,\ldots,q_5) = \sum_{i=1}^{5}\sum_{z_i=0}^{1}\left[q_i(z_i)\log\frac{p(x_i,z_i|\rho_A,\rho_B)}{q_i(z_i)}\right], \tag{6}$$

where $q_i(z_i) = p_i^{z_i}(1-p_i)^{(1-z_i)}$ and each p_i is a variational parameter indicating the probability that the i-th series of coin tosses was done using coin A.

(i) Compute the form of the free-energy in terms of $p_1,\ldots,p_5$ and ρ_A and ρ_B. [30%]

(ii) Using the answer to Part (b)(i), compute the M-step update equations for ρ_A and ρ_B in terms of $p_1,\ldots,p_5$ and $x_1,\ldots,x_5$. [25%]

(iii) Using the answer to Part (b)(i), compute the E-step update equations for $p_1,\ldots,p_5$ as a function of ρ_A, ρ_B and and $x_1,\ldots,x_5$. [25%]

4 (a) Describe the problems that Monte Carlo methods aim to solve in machine learning. How do they generally solve these problems? What are the advantages and disadvantages of Monte Carlo methods? [20%]

(b) Assume that the random variable $x \in [-1,1]$ follows a probability distribution whose density $p(x)$ satisfies

$$p(x) \propto \begin{cases} \mathcal{N}(x|0,1) & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where the symbol $\propto$ means *"proportional to"* and $\mathcal{N}(x|0,1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$. Assume that you have access to a computer routine that samples the random variable y uniformly in $[-1,1]$ and note that $\int_{-\infty}^{-1} \mathcal{N}(x|0,1)\,dx = 0.1586553$.

(i) Compute the normalization constant of $p(x)$ and indicate the value of $p(x)$ at its maximizer: the x, $x \in [-1,1]$, that produces the largest $p(x)$. [10%]

(ii) Describe how to use the rejection sampling method to generate samples of x from samples of y. [20%]

(iii) What is the acceptance probability of rejection sampling in this problem? [10%]

(c) A sequence $\{Y_t\}_{t=1}^T$ is formed by elements $Y_t \in \{A,B\}$. Let us assume that there is a hidden binary sequence $\{X_t\}_{t=1}^T$ that controls the generation of each entry Y_t. Each X_t takes values in $\{0,1\}$ and follows the following transition matrix:

$$\begin{bmatrix} p(X_t=0|X_{t-1}=0) & p(X_t=0|X_{t-1}=1) \\ p(X_t=1|X_{t-1}=0) & p(X_t=1|X_{t-1}=1) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}. \tag{8}$$

The emission probabilities for Y_t as a function of X_t are

$$\begin{bmatrix} p(Y_t=A|X_t=0) & p(Y_t=A|X_t=1) \\ p(Y_t=B|X_t=0) & p(Y_t=B|X_t=1) \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}. \tag{9}$$

and the initial state probabilities are $p(X_1=0)=0.5$ and $p(X_1=1)=0.5$.

(i) Describe how to compute the likelihood $p(Y_{1:T})$ efficiently using recursion. That is, how can you efficiently sum $p(Y_{1:T},X_{1:T})$ over all $X_{1:T}$ using recursion? [20%]

(ii) Using your response to Part (c)(i), compute the probability given by the model to the sequence AB. Show your derivations and, to simplify them, use the fact that

$$p(Y_1=A,X_2) = \sum_{X_1} p(X_2|X_1)p(Y_1=A,X_1) = X_2 0.19 + 0.31(1-X_2). \tag{10}$$

[20%]

END OF PAPER

THIS PAGE IS BLANK

Analytic answers:

1b)
i) $3/10$
ii) $4/11$
iii) $10/13$

1c) $\mathcal{N}(y_\star | x_\star m, x_\star^2 v + 1)$

2b)
iii) $p(y_n|\mathbf{w},\varepsilon,\mathbf{x}_n) = \varepsilon\frac{1}{2} + (1-\varepsilon)p(y_n|\mathbf{w},\mathbf{x}_n)\varepsilon\frac{1}{2} + (1-\varepsilon)\left[\frac{1+y_n}{2}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) + \frac{1-y_n}{2}(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n))\right]$.
iv) $p(z_n = 1|y_n,\mathbf{w},\varepsilon,\mathbf{x}_n) = \frac{\varepsilon\frac{1}{2}}{\varepsilon\frac{1}{2}+(1-\varepsilon)p(y_n|\mathbf{w},\mathbf{x}_n)}$

3b)
i)

$$\mathscr{F}(\rho_A,\rho_B,q_1,\ldots,q_5) = \sum_{i=1}^{5} p_i\left[x_i\log\rho_A + (10-x_i)\log(1-\rho_A) - \log p_i\right] + (1-p_i)\left[x_i\log\rho_B + (10-x_i)\log(1-\rho_B) - \log(1-p_i)\right] + \text{const}$$

ii)
$\rho_A = \frac{\sum_{i=1}^5 p_i x_i}{10\sum_{i=1}^5 p_i}$
$\rho_B = \frac{\sum_{i=1}^5 (1-p_i)x_i}{10\sum_{i=1}^5 (1-p_i)}$

iii) $p_i = \sigma\left(x_i\log\frac{\rho_A}{\rho_B} + (10-x_i)\log\frac{(1-\rho_A)}{(1-\rho_B)}\right)$, where $\sigma(x) = 1/(1+\exp(-x))$

4b)
i) $Z = 0.6826894$ and $p(x_{\text{maximizer}}) = 0.5843686$
iii) $p = 0.855$.

4c)
ii) 0.226.