CRIB

Version RET/3

EGT2
ENGINEERING TRIPOS PART IIA

Thursday 25 April 2019 14.00 to 15.40

Module 3F8

INFERENCE

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** number 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
Information 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 what is *maximum likelihood estimation* and how it is used to estimate parameters in a probabilistic model from data. [20%]

(b) A regression problem comprises scalar inputs x_n and scalar outputs y_n which are linearly related: $y_n = mx_n + c + \varepsilon_n$. The noise ε_n is Gaussian with mean 0 but with variance that depends quadratically on the input: $p(\varepsilon_n) = \mathcal{N}(\varepsilon_n; 0, \alpha x_n^2)$. Due to physical constraints, the inputs lie in the region $1 < x_n < 100$.

The offset c and noise parameter α are known, but the slope m must be learned from a training dataset $\{y_n, x_n\}_{n=1}^N$.

(i) Compute the log-likelihood of m. [20%]

(ii) Compute the maximum likelihood estimate of m. [40%]

(iii) You are allowed to select a new input location x at which you will be provided with a corresponding output y and the new pair $\{y, x\}$ will be added to the training data. Which value of x will be most informative about the parameter m? Explain your reasoning. [20%]

Q1

a) $\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log p(y|\theta)$

- θ_{MLE}: maximum likelihood estimate of θ
- $\underset{\theta}{\operatorname{argmax}}$: θ
- $\log p(y|\theta)$: model
- y: data
- θ: parameters

b) i) $$\mathcal{L}(m) = \log \prod_n p(y_n | x_n, m, c, \alpha)$$
$$= -\frac{1}{2} \sum_n \left[\log(\alpha x_n^2) + \frac{(mx_n + c - y_n)^2}{\alpha x_n^2} \right]$$

ii) $$\left. \frac{d\mathcal{L}(m)}{dm} \right|_{m=m_{MLE}} = -\sum_n \frac{1}{\alpha x_n^2} x_n (m_{MLE} x_n + c - y_n) = 0$$

$$\Rightarrow N m_{MLE} + \sum_n \frac{(c - y_n)}{x_n} = 0$$

$$\therefore m_{MLE} = \frac{1}{N} \sum_n \frac{y_n - c}{x_n}$$

iii) Two effects in opposite directions

- select smallest x_n allowed ($x_n = 1$) since noise is smallest there
- select largest x_n allowed since signal is largest

$$y_n = \underbrace{mx_n + c}_{\text{signal}} + \underbrace{\sqrt{\alpha}\, x_n n_n}_{\text{noise}} \qquad n_n \sim N(0,1)$$

These two effects cancel

$$m_{MLE} = \frac{1}{N} \sum_n \frac{(y_n - c)}{x_n} = \frac{1}{N} \sum_n \frac{(mx_n + c + \alpha x_n n_n - c)}{x_n}$$

(sub in true value for this)

$$= \frac{1}{N} \sum_n (m + \alpha n_n)$$

(m: true slope; n_n: Gaussian noise)

$\Rightarrow$ terms in sum statistically identical $\therefore$ doesn't matter how we set x_n.

2 A retailer has access to the collar size and waist size measurements of N customers. The retailer wants to use this information to estimate the relative size of each customer in order to recommend further products. The mean of the data is removed and the two measurements for each customer are stacked into a vector $\mathbf{y}_n = [y_{1,n}, y_{2,n}]^\top$. The full data set is denoted $\{\mathbf{y}_n\}_{n=1}^N$.

The retailer models the pair of measurements for each customer in terms of a scalar latent relative-size variable s_n

$$\mathbf{y}_n = \begin{bmatrix} y_{1,n} \\ y_{2,n} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} s_n + \sigma_y \begin{bmatrix} \varepsilon_{1,n} \\ \varepsilon_{2,n} \end{bmatrix} = \mathbf{w} s_n + \sigma_y \boldsymbol{\varepsilon}_n.$$

Here the weights $\mathbf{w} = [w_1, w_2]^\top$ capture the dependence of the two measurements on relative-size and $\boldsymbol{\varepsilon}_n = [\varepsilon_{1,n}, \varepsilon_{2,n}]^\top$ is independent and identically distributed Gaussian noise with mean zero and identity covariance $p(\boldsymbol{\varepsilon}_n) = \mathcal{N}(\varepsilon_n; \mathbf{0}, \mathbf{I})$. Relative-size is assumed to follow a standard Gaussian distribution $p(s_n) = \mathcal{N}(s_n; 0, 1)$.

(a) The retailer wants to fit the parameters of the model $\theta = \{\mathbf{w}, \sigma_y\}$ using maximum likelihood. Compute the likelihood of the parameters $p(\{\mathbf{y}_n\}_{n=1}^N | \theta)$. [30%]

(b) Having fit the model, the retailer would like to estimate the size of each customer. Compute the posterior distribution over the relative-size variable given the measurements $p(s_n | \mathbf{y}_n, \theta)$. [40%]

(c) The retailer would like to enhance the model by including two additional pieces of information. First, they know the shoe size of their customers, which takes integer values in the range $\{2 \ldots 12\}$. Second, they know the age of their customers which affects their relative-size. Describe how the model can be altered to incorporate this information. [30%]

Q2

a) $s_n \sim N(0,1)$

$\underline{y}_n | s_n \sim N(\underline{w} s_n, \sigma_y^2 \underline{\underline{I}})$

$\Rightarrow p(\{\underline{y}_n\}_{n=1}^N | \theta) = \prod_n p(\underline{y}_n | \theta)$ ← Gaussian

$\mathbb{E}_{p(s_n, \underline{y}_n)}[\underline{y}_n] = \underline{0}$

$\mathbb{E}_{p(s_n, \underline{y}_n)}[\underline{y}_n \underline{y}_n^T] = \mathbb{E}_{p(s_n, \underline{\varepsilon}_n)}[\underline{w} s_n s_n \underline{w}^T + \underline{\varepsilon}_n \underline{\varepsilon}_n^T] = \underline{w}\underline{w}^T + \sigma_y^2 \underline{\underline{I}}$

$\therefore p(\{\underline{y}_n\}_{n=1}^N | \theta) = \prod_n N(\underline{y}_n ; \underline{0}, \underline{w}\underline{w}^T + \sigma_y^2 \underline{\underline{I}})$

b) $p(s_n | \underline{y}_n, \theta) \propto p(s_n) p(\underline{y}_n | s_n, \theta)$

$\propto e^{-\frac{1}{2}s_n^2 - \frac{1}{2\sigma_y^2}(y_{1n} - w_1 s_n)^2 - \frac{1}{2\sigma_y^2}(y_{2n} - w_2 s_n)^2}$

$\propto e^{-\frac{1}{2}s_n^2 (1 + w_1^2/\sigma_y^2 + w_2^2/\sigma_y^2) + s_n \frac{1}{\sigma_y^2}(w_1 y_{1n} + w_2 y_{2n})}$

$\therefore p(s_n | \underline{y}_n, \theta) = N(s_n ; \mu_{s_n|\underline{y}_n}, \sigma^2_{s_n|\underline{y}_n})$

$\sigma^2_{s_n|\underline{y}_n} = \frac{1}{1 + w_1^2/\sigma_y^2 + w_2^2/\sigma_y^2} = \frac{\sigma_y^2}{\sigma_y^2 + w_1^2 + w_2^2}$

$\mu_{s_n|\underline{y}_n} = \sigma^2_{s_n|\underline{y}_n} \frac{1}{\sigma_y^2}(w_1 y_1 + w_2 y_2)$

$= \frac{1}{\sigma_y^2 + w_1^2 + w_2^2}(w_1 y_1 + w_2 y_2)$

lots of potential answers, the following is just one possibility

c) Assume the noise in the shoe size is independent from that in y_1 & y_2

let y_3 = shoe size, a_n = age of customer

prior now depends on age

$$p(s_n|a_n) = N(s_n; \mu(a_n), \sigma^2(a_n))$$

mean and possibly variance depend on age

could use regression model for this e.g. with radial basis functions

$$p(y_{1n}|s_n) = N(y_{1n}; w_1 s_n, \sigma_y^2)$$

$$p(y_{2n}|s_n) = N(y_{2n}; w_2 s_n, \sigma_y^2)$$

Now define $p(y_{3n}|s_n)$ as follows

let $z_n = w_3 s_n + \mu_3 + \varepsilon_{3n}$ $\quad \varepsilon_3 \sim N(0, \sigma_3^2)$

Now define 12 bins:

$$z_n \leq z_{min} \Rightarrow y_{3n} = 2$$

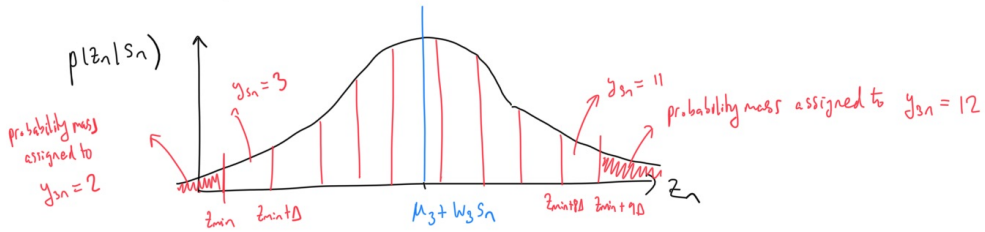
$$z_{min} < z_n < z_{min}+\Delta \Rightarrow y_{3n} = 3$$

$$\vdots$$

$$z_{min}+8\Delta < z_n < z_{min}+9\Delta \Rightarrow y_{3n} = 11$$

$$z_{min}+9\Delta < z_n \Rightarrow y_{3n} = 12$$

Or schematically:



use error functions to compute these integrals as for probit function.

3 The *binary latent feature model* first draws two binary latent variables s_1 and s_2 from independent Bernoulli distributions. That is, $p(s_1 = 1|\theta) = \pi_1$ and $p(s_2 = 1|\theta) = \pi_2$. Observations $\mathbf{y}$, which are real valued and D dimensional, are produced by multiplying the latent variables by the associated weights ($\mathbf{w}_1$ and $\mathbf{w}_2$), adding these contributions, and corrupting with isotropic Gaussian noise of variance σ_y^2. That is, $p(\mathbf{y}|s_1, s_2, \theta) = \mathcal{N}(\mathbf{y}; \mathbf{w}_1 s_1 + \mathbf{w}_2 s_2, \mathrm{I}\sigma_y^2)$.

Above, the model parameters have been denoted $\theta = \{\pi_1, \pi_2, \mathbf{w}_1, \mathbf{w}_2, \sigma_y^2\}$.

(a) Mathematically define a *mixture of Gaussians model*. Your definition should identify the *mixing proportions*, *component means* and *component variances*. [20%]

(b) Express the binary latent feature model, described at the start of this question, as a mixture of Gaussians stating the *mixing proportions*, *component means* and *component variances* in terms of the parameters θ. [40%]

(c) A machine learner will employ the EM algorithm to learn the parameters of the *binary latent feature model* defined above.

(i) Compute the E-step update by deriving the posterior distribution over the latent variables given an observed variable, $p(s_1, s_2|\mathbf{y}, \theta)$. [15%]

(ii) Describe how you would mathematically derive the M-step update. Your description should outline the steps involved, but does not require detailed calculation. [25%]

Q3

a) MoG model

$p(z_n = k) = \tilde{\pi}_k$ for $k = 1 \dots K$ $\quad \sum_k \tilde{\pi}_k = 1$

(mixing proportions; number of components)

$p(\underline{y}_n | z_n = k) = N(\underline{y}_n ; \underline{\mu}_k, \underline{\underline{\Sigma}}_k)$

(component covariance; component mean)

$$\Rightarrow p(\underline{y}_n | \theta) = \sum_{k=1}^{K} \tilde{\pi}_k \, N(\underline{y}_n ; \underline{\mu}_k, \underline{\underline{\Sigma}}_k)$$

b)

original model

State s_1, s_2	prior $p(s_1, s_2)$	conditional $p(\underline{y} \mid s_1, s_2, \theta)$
$s_1 = 0, s_2 = 0$	$(1-\pi_1)(1-\pi_2)$	$N(\underline{y}; \underline{0}, \underline{\underline{I}}\sigma_y^2)$
$s_1 = 1, s_2 = 0$	$\pi_1 (1-\pi_2)$	$N(\underline{y}; \underline{w}_1, \underline{\underline{I}}\sigma_y^2)$
$s_1 = 0, s_2 = 1$	$(1-\pi_1)\pi_2$	$N(\underline{y}; \underline{w}_2, \underline{\underline{I}}\sigma_y^2)$
$s_1 = 1, s_2 = 1$	$\pi_1 \pi_2$	$N(\underline{y}; \underline{w}_1 + \underline{w}_2, \underline{\underline{I}}\sigma_y^2)$

MoG view of model

state z	prior $p(z=k)$	conditional $p(\underline{y} \mid z, \theta) = N(\underline{y}; \underline{\mu}_z, \underline{\underline{\Sigma}}_z)$
$z=1$	$\tilde{\pi}_1 = (1-\pi_1)(1-\pi_2)$	$N(\underline{y}; \underline{\mu}_1 = \underline{0}, \underline{\underline{\Sigma}}_1 = \underline{\underline{I}}\sigma_y^2)$
$z=2$	$\tilde{\pi}_2 = \pi_1(1-\pi_2)$	$N(\underline{y}; \underline{\mu}_2 = \underline{w}_1, \underline{\underline{\Sigma}}_2 = \underline{\underline{I}}\sigma_y^2)$
$z=3$	$\tilde{\pi}_3 = (1-\pi_1)\pi_2$	$N(\underline{y}; \underline{\mu}_3 = \underline{w}_2, \underline{\underline{\Sigma}}_3 = \underline{\underline{I}}\sigma_y^2)$
$z=4$	$\tilde{\pi}_4 = \pi_1 \pi_2$	$N(\underline{y}; \underline{\mu}_4 = \underline{w}_1 + \underline{w}_2, \underline{\underline{\Sigma}}_4 = \underline{\underline{I}}\sigma_y^2)$

c i) let $p(s_{1n}=0, s_{2n}=0 \mid \underline{y}_n) = \dfrac{p(\underline{y}_n \mid s_{1n}=0, s_{2n}=0)\, p(s_{1n}=0, s_{2n}=0)}{\sum_{s_1,s_2} p(\underline{y}_n \mid s_{1n}, s_{2n})\, p(s_{1n}, s_{2n})} \triangleq \dfrac{r_{00}^{(n)}}{r^{(n)}}$

similarly

$$p(s_{1n}=1, s_{2n}=0 \mid \underline{y}_n) = \frac{r_{10}^{(n)}}{r^{(n)}}$$

$$p(s_{1n}=0, s_{2n}=1 \mid \underline{y}_n) = \frac{r_{01}^{(n)}}{r^{(n)}}$$

$$p(s_{1n}=1, s_{2n}=1 \mid \underline{y}_n) = \frac{r_{11}^{(n)}}{r^{(n)}}$$

where $r^{(n)} = r_{00}^{(n)} + r_{10}^{(n)} + r_{01}^{(n)} + r_{11}^{(n)}$

and $r_{00}^{(n)} = (1-\pi_1)(1-\pi_2)\, \mathcal{N}(\underline{y}_n ; \underline{0}, \underline{\underline{I}}\sigma_y^2)$

$r_{10}^{(n)} = \pi_1(1-\pi_2)\, \mathcal{N}(\underline{y}_n ; \underline{w}_1, \underline{\underline{I}}\sigma_y^2)$

$r_{01}^{(n)} = (1-\pi_1)\pi_2\, \mathcal{N}(\underline{y}_n ; \underline{w}_2, \underline{\underline{I}}\sigma_y^2)$

$r_{11}^{(n)} = \pi_1\pi_2\, \mathcal{N}(\underline{y}_n ; \underline{w}_1 + \underline{w}_2, \underline{\underline{I}}\sigma_y^2)$

ii) M-Step computes $\underset{\theta}{\operatorname{argmax}} \sum_n \mathbb{E}_{q_n(\underline{s}_n)}\left[\log p(s_{1n}, s_{2n} \mid \theta) + \log p(\underline{y}_n \mid s_{1n}, s_{2n}, \theta)\right]$

- differentiate this objective and either
  a) set to zero to find optimum (using Lagrange multipliers for π_1 / π_2)
  b) use a gradient based optimization (reparametrising π_1, π_2 & σ_y^2 to avoid constraints)

4 (a) Define a *hidden Markov model* that has a discrete hidden state. Your definition should identify the *initial state probabilities*, *transition probabilities* and the *emission distribution*. [20%]

(b) A sequence model comprises a discrete latent variable s_t and a continuous observed variable y_t. The discrete state always begins with value 1, that is $p(s_1 = 1) = 1$. The state then evolves according to following rule

$$p(s_t = k|s_{t-1} = k') = \begin{cases} 0.9 & \text{for } k = k' + 1 \\ 0.1 & \text{for } k = 1 \\ 0 & \text{for all other } k \end{cases} .$$

The observations are generated from Gaussian distributions $p(y_t|s_t) = \mathcal{N}(y_t; s_t^2, 0.1^2)$.

(i) Sketch a typical sample from the model for $T = 20$ time steps. Your sketch should show both the latent process and the corresponding observed process. Explain the salient features. [30%]

(ii) An algorithm has been used to compute the posterior distribution over the latent variable s_t from t observations $y_{1:t}$. The posterior is found to be

$$p(s_t = k|y_{1:t}) = \begin{cases} \frac{2}{3} & \text{for } k = 3 \\ \frac{1}{3} & \text{for } k = 4 \\ 0 & \text{for all other } k \end{cases} .$$

Compute the predictive distribution over the next observed variable i.e. $p(y_{t+1}|y_{1:t})$. Explain each step in your calculation. [40%]

(iii) What algorithm could be used to compute the posterior distribution $p(s_t = k|y_{1:t})$? Would the standard implementation of this algorithm need to be modified to handle long sequences? Explain your reasoning. [10%]

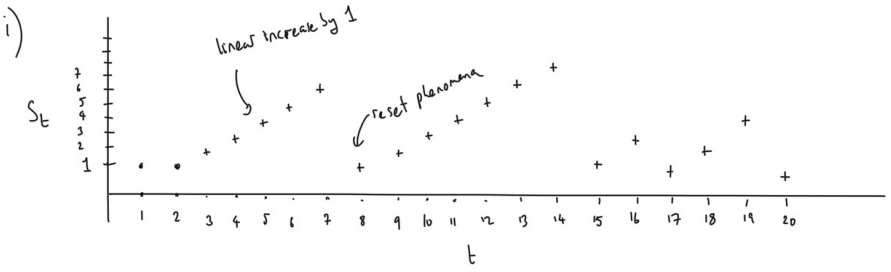
Q4

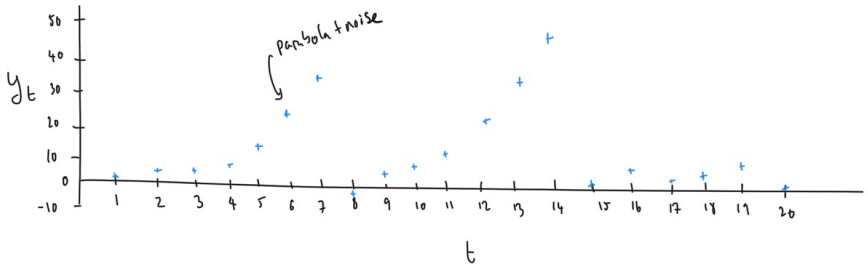
a) Hidden Markov Model : $p(S_1 = k) = \Pi_k$ for $k = 1 \dots K$ ← initial state probability

$p(S_t = k \mid S_{t-1} = k') = T_{kk'}$ ← transition probability

$p(y_t \mid S_t)$ ← emission distribution

b) i)





ii) $$p(y_{t+1} \mid y_{1:t}) = \sum_{s_t, s_{t+1}} p(y_{t+1}, s_{t+1}, s_t \mid y_{1:t})$$

$$= \sum_{s_t, s_{t+1}} p(y_{t+1} \mid s_{t+1})\, p(s_{t+1}, s_t \mid y_{1:t})$$

enumerate this

Four options for the hidden states:

$$p(s_t = 3, s_{t+1} = 4 \mid y_{1:t}) = \frac{2}{3} \cdot \frac{9}{10}$$

$$p(s_t = 3, s_{t+1} = 1 \mid y_{1:t}) = \frac{2}{3} \cdot \frac{1}{10}$$

$$p(s_t = 4, s_{t+1} = 5 \mid y_{1:t}) = \frac{1}{3} \cdot \frac{9}{10}$$

$$p(s_t = 4, s_{t+1} = 1 \mid y_{1:t}) = \frac{1}{3} \cdot \frac{1}{10}$$

$$\therefore\ p(y_{t+1} \mid y_{1:t}) = \frac{3}{5} N(y_t; 16, 0.1^2) + \frac{1}{10} N(y_t; 1, 0.1^2) + \frac{3}{10} N(y_t; 25, 0.1^2)$$

iii) Forward algorithm is used to compute $p(s_t = k \mid y_{1:t})$. Typically this is implemented using matrix multiplies with a transition matrix $T_{k,k'}$ (see part a). However, the number of states K in the current model is set by the duration T. This could result in very large matrices. Best to rewrite the matrix multiplication in a bespoke way that leverages the sparsity / special structure in the transition matrix.

END OF PAPER