EGT3
ENGINEERING TRIPOS PART IIA

CRIB

Module 3F8

INFERENCE

STATIONERY REQUIREMENTS

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM

10 minutes reading time is allowed for this paper.

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%]

a) let data be denoted $Y = \{y_n\}_{n=1}^N$
parameters denoted θ

$p(Y|\theta)$ = probability of data given parameters
&
likelihood of the parameters

maximum likelihood estimates for the parameters, θ_{ML}, are found via

$$\theta_{ML} = \arg\max_{\theta} p(Y|\theta) = \arg\max_{\theta} \log p(Y|\theta)$$

(b) A source emits N signals x_n drawn independently from a Gaussian distribution with mean 1 and variance 1. The signals are measured by a receiver a fixed distance d metres away. The signals are exponentially attenuated and corrupted by independent Gaussian noise so that the measurements are given by $y_n = \exp(-d)x_n + \varepsilon_n$. The noise ε_n has zero mean and variance 1.

The formula for a one dimensional Gaussian distribution with mean μ and variance σ^2 is

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right).$$

.

(i) Compute the mean and the variance of a single measurement y_n under the probabilistic model. [30%]

b) $y_n = w x_n + \varepsilon_n$ where $w = e^{-d}$ $\quad \varepsilon_n \sim \mathcal{N}(0,1)$ $\quad x_n \sim \mathcal{N}(1,1)$

i) $\mathbb{E}_{p(x_n,\varepsilon_n)}[y_n] = w\,\mathbb{E}_{p(x_n)}(x_n) + \mathbb{E}_{p(\varepsilon_n)}(\varepsilon_n) = w = e^{-d} = \mu_y$

$$\mathbb{E}_{p(x_n,\varepsilon_n)}[y_n^2] = \mathbb{E}_{p(x_n,\varepsilon_n)}[w x_n + \varepsilon_n]^2$$

$$= \mathbb{E}_{p(x_n,\varepsilon_n)}[w^2 x_n^2 + 2w x_n \varepsilon_n + \varepsilon_n^2]$$

$$= w^2\,\mathbb{E}_{p(x_n,\varepsilon_n)}[x_n^2] + 2w\,\mathbb{E}_{p(x_n)}[x_n]\,\mathbb{E}_{p(\varepsilon_n)}(\varepsilon_n) + \mathbb{E}_{p(\varepsilon_n)}(\varepsilon_n^2)$$

$$= w^2 \times 2 + 0 + 1$$

$$= 2w^2 + 1$$

Variance is given by

$$\sigma_y^2 = \mathbb{E}_{p(x_n,\varepsilon_n)}[y_n^2] - \left[\mathbb{E}_{p(x_n,\varepsilon_n)}(y_n)\right]^2 = w^2 + 1$$
$$= e^{-2d} + 1$$

(ii) Use your answer to (i) to compute the likelihood of the parameter d when N measurements have been made. [20%]

ii) $$p(y_{1:N}|d) = \prod_n \frac{1}{\sqrt{2\pi\sigma_y(d)}} e^{-\frac{1}{2\sigma_y^2}(y_n - \mu_y(d))^2}$$

$$= \left(2\pi\sigma_y(d)\right)^{-N/2} e^{-\frac{1}{2\sigma_y^2}\sum_n (y_n - \mu_y(d))^2}$$

$$= \left(2\pi\left(1+w^2(d)\right)\right)^{-N/2} e^{-\frac{1}{2(1+w^2(d))}\sum_n (y_n - w(d))^2}$$

where $w(d) = e^{-d}$

(iii) Four measurements are made $\{y_n\}_{n=1}^{N} = \{1, -2, -1, 2\}$. Find the maximum likelihood setting of the parameter d for these data. You may find it simpler to first find the maximum likelihood setting for $w = \exp(-d)$ and then rearrange to find the estimate of d. [30%]

iii) $\log p(y_{1:N}|w) = -\frac{N}{2}\log\left(2\pi(1+w^2)\right) - \frac{1}{2(1+w^2)}\sum_n (y_n - w)^2$

$$= -\frac{N}{2}\log\left(2\pi(1+w^2)\right) - \frac{1}{2(1+w^2)}\left[\sum_n y_n^2 - 2w\sum_n y_n + Nw^2\right]$$

in our case

$$\sum_n y_n = 0 \qquad \sum_n y_n^2 = 1+4+1+4 = 10 \qquad N = 4$$

$$\therefore \log p(y_{1:N}|w) = -2\log\left(2\pi(1+w^2)\right) - \frac{1}{2(1+w^2)}\left[10+4w^2\right]$$

$$0 = \frac{d}{dw}\log p(y_{1:N}|w) = \frac{-2\cdot 2w}{1+w^2} + \frac{1}{\cancel{2}(1+w^2)^2}\cdot\cancel{2}w\left[10+4w^2\right]$$

$$- \frac{8w}{2(1+w^2)}$$

$$0 = -8w(1+w^2) + w(10+4w^2)$$

$w=0$ is a minimum

$$0 = 2 - 4w^2$$

$$\Rightarrow w_{ML} = \sqrt{\frac{1}{2}} \qquad \Rightarrow d_{ML} = \frac{1}{2}\log_e(2) = 0.347$$
(3 sf)

Assessor's comments: The most popular question (all but two candidates attempted it). Generally very well answered, but it was clear that some candidates spent a large amount of time on this question.

2 (a) Compare and contrast *regression* and *classification* tasks in machine learning. [30%]

a) in both regression & classification you get given a training set of input/output pairs $\{\underline{x}_n, y_n\}_{n=1}^{N}$ and the goal is to predict outputs y^* associated with unseen (new) inputs x^*.

inputs outputs

In regression the outputs are real valued, $y_n \in \mathbb{R}e$

In classification the outputs are discrete valued, $y_n \in \{1 \ldots K\}$ where K = # of classes

(b) A dataset comprises pairs of real valued inputs x_n and real valued outputs y_n shown in Fig. 1. Suggest a suitable probabilistic model for these data that could be used to predict an output from new input. Explain your reasoning. [30%]

data are noisy & Gaussian looks reasonable with variance $\approx 5^2$

b) $$p(y_n|x_n) = \mathcal{N}\left(y_n; 2\exp\left(\frac{x_n}{20}\right) + 25, 5^2\right)$$

data lie on an exponential trend with length scale ≈ 20

at $x=0$ the trend passes through ≈ 25

Only rough numerical values are required, the reasoning is the key aspect

($\Rightarrow$ regression problem with non-linear basis function)

(other sensible choices of model are ok too)

(c) A second set of discrete valued outputs z_n, shown in Fig. 2, were measured simultaneously with y_n so that the training data is now $\{x_n, y_n, z_n\}_{n=1}^{N}$. Extend the probabilistic model you proposed for part (b) so that it can be used to jointly predict both outputs from a new input. Explain your reasoning. [30%]

c) Assumption : y_n & z_n are independent given x_n

Note : $z_n \in \{1, 2, 3\}$ i.e. takes discrete values $\therefore$ suggest

$$p(z_n = k \mid x_n) = \frac{e^{\underline{w}_k^T \underline{\Psi}(x_n) + b_k}}{\sum_{l=1}^{K} e^{\underline{w}_l^T \underline{\Psi}(x_n) + b_l}} \quad \text{(softmax)}$$

Linear basis functions will not be sufficient here since class 1 occurs around $x_n < 20$ & $80 < x_n$. Quadratic, or some other sensible non-linear basis function would be appropriate i.e. $\Psi(x_n) = \begin{bmatrix} x_n \\ x_n^2 \end{bmatrix}$

($\Rightarrow$ multi class classification with non linear basis functions)

(other sensible solutions are fine to)

(d) Consider the extended model you have proposed in part (c). Do the second set of outputs provide useful information about the parameters of the original component described in part (b)? Explain your reasoning. [10%]

d) The likelihood for the model described in parts b & c above is:

$$\prod_n p(y_n|x_n,\theta_1)\, p(z_n|x_n,\theta_2) = g_y(\theta_1)\, h_z(\theta_2)$$

If maximum likelihood is used to fit the model, then $g_y(\theta_1)$ can be optimized to find θ_1 & $h_z(\theta_2)$ to find $\theta_2 \Rightarrow \{y_n\}$ do not provide useful information about θ_2 & vice versa. Since

$$\log p(z,y|x,\theta) = \log g_y(\theta_1) + \log g_z(\theta_2)$$

Other model proposals for parts b & c and inference schemes can lead to different answers here eg if correlated noise is assumed between y_n & z_n, if generative models are suggested or if prior distributions over parameters are used that are not independent: $p(\theta_1, \theta_2) \neq p(\theta_1)p(\theta_2)$.

Assessor's comments: A large number of adequate solutions, but few were very good or very poor. Many candidates failed to identify that the second dataset in part (c) was a classification dataset. Some candidates wrote that classification and/or regression were examples of unsupervised learning, rather than supervised learning.

3 (a) Describe what *clustering* is and give an example application where a clustering algorithm might be used. [20%]

a) Clustering is an unsupervised machine learning problem in which data $\{y_n\}_{n=1}^N$ are assigned into one of a number of clusters $\{s_n\}_{n=1}^N$ where $s_n \in \{1,\dots,K\}$ in such a way that "near by" or "similar" data points are assigned to the same cluster and "far away" or "dissimilar" points are assigned to different clusters.

Example application: segmentation of image pixels

(b) A simple one dimensional dataset, $\{y_n\}_{n=1}^N = \{-10.1, -9.9, 9.9, 10.1\}$, is modelled using a mixture of Gaussians. The mixture comprises two components with class membership probabilities $p(s_n = 1) = \alpha$ and $p(s_n = 2) = 1 - \alpha$. The component distributions are given by $p(y_n|s_n = 1, \theta) = \mathcal{N}(y_n; \mu_1, \sigma_1^2)$ and $p(y_n|s_n = 2, \theta) = \mathcal{N}(y_n; \mu_2, \sigma_2^2)$ where

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

The parameters of the model are collectively denoted $\theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \alpha\}$.

(i) The model is fitted using the Expectation Maximisation (EM) algorithm. The posterior distribution over the class labels for the nth data point is denoted $r_{n,k} = p(s_n = k|y_n)$. Write down the algorithm's M-Step update equations. [40%]

b i) $\mu_k = \dfrac{\sum_{n=1}^N r_{nk} y_n}{\sum_{n=1}^N r_{nk}}$ $\sigma_k^2 = \dfrac{\sum_{n=1}^N r_{nk} (y_n - \mu_k)^2}{\sum_{n=1}^N r_{nk}}$

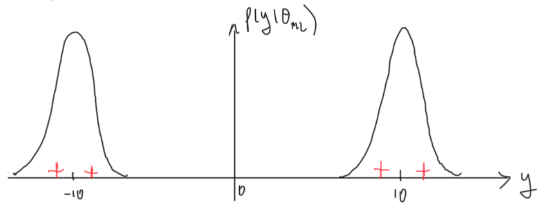
$$\pi_k = \frac{1}{N} \sum_{n=1}^N r_{nk}$$

(ii) The EM algorithm is run to convergence returning a parameter estimate θ_{EM}. The estimate is found to depend on the initialisation. Sketch the various Gaussian mixture model fits that the EM algorithm returns, i.e. sketch the densities $p(y|\theta_{\text{EM}})$ as a function of y. Where possible, indicate the estimated parameter values approximately. Identify which estimates correspond to global optima of the likelihood. [40%]

ii)

local optimum

+ = data



$\mu_1 \simeq -10$ $\mu_2 \simeq 10$

$\sigma_1^2 \simeq 0.1^2$ $\sigma_2^2 \simeq 0.1^2$

$\pi_1 \simeq 1/2$ $\pi_2 \simeq 1/2$

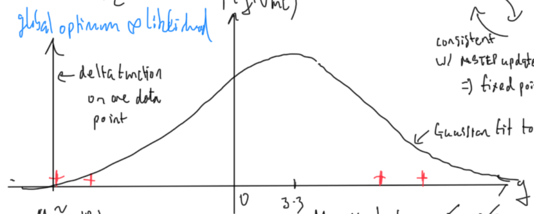
consistent w/ M-STEP updates ⇒ fixed point of EM

assignments will be

n	y_n	r_{nk}
1	-10.1	$\approx [1,0]$ ← as Gaussian well separated
2	-9.9	$\approx [1,0]$
3	9.9	$\approx [0,1]$
4	10.1	$\approx [0,1]$

global optimum of likelihood

n	y_n	r_{nk}
1	-10.1	$[1,0]$
2	-9.9	$[0,1]$
3	9.9	$[0,1]$
4	10.1	$[0,1]$



$\mu_1 \simeq -10.1$

$\sigma_1^2 \approx 0$

$\pi_1 \approx 1/4$

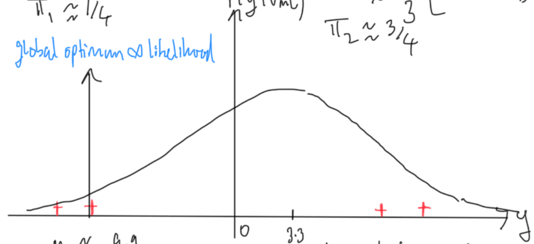
$\mu_2 \approx \frac{1}{3}(-\cancel{9.9}+\cancel{9.9}+10.1) \approx 3.3$

$\sigma_2^2 \approx \frac{1}{3}\left[(-9.9-3.3)^2+(9.9-3.3)^2+(10.1-3.3)^2\right]$

$\pi_2 \approx 3/4$

global optimum of likelihood

n	y_n	r_{nk}
1	-10.1	$[0,1]$
2	-9.9	$[1,0]$
3	9.9	$[0,1]$
4	10.1	$[0,1]$



$\mu_1 \approx -9.9$

$\sigma_1^2 \approx 0$

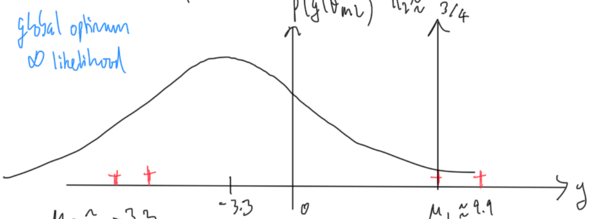
$\pi_1 \approx 1/4$

$\mu_2 \approx \frac{1}{3}(-10.1+9.9+10.1) = 3.3$

$\sigma_2^2 \approx \frac{1}{3}\left[(-10.1-3.3)^2+(9.9-3.3)^2+(10.1-3.3)^2\right]$

$\pi_2 \approx 3/4$

global optimum of likelihood



$\mu_2 \approx -3.3$

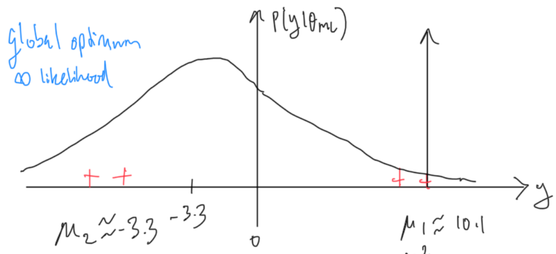
$\sigma_2 = \frac{1}{3}\left((-10.1+3.3)^2+(-9.9+3.3)^2+(10.1+3.3)^2\right)$

$\pi_2 = 3/4$

$\mu_1 \approx 9.9$

$\sigma_1^2 \approx 0$

$\pi_1 \approx 1/4$



$\mu_2 \approx -3.3$ $\qquad$ $\mu_1 \approx 10.1$

$\sigma_2^2 \approx \frac{1}{3}\left((-10.1+3.3)^2 + (-9.9+3.3)^2 + (9.9+3.3)^2\right)$ $\qquad$ $\sigma_1^2 \approx 0$

$\pi_2 \approx 3/4$ $\qquad$ $\pi_1 \approx 1/4$

NB for each of the above solutions, there is also a solution that switches $\{\mu_1, \sigma_1, \pi_1\}$ to $\{\mu_2, \sigma_2, \pi_2\}$ & vice versa.

Assessor's comments: This question was the hardest one on the exam. Many candidates gave full derivations for the M-Step equations in part (bi). This was not asked for in the question and it was sufficient to write them directly. Some candidates struggled with part (bii) failing to identify that the maximum likelihood solutions will place a single Gaussian of zero width on one of the data points, and use the other Gaussian to model the remaining points.

4 (a) Explain what the terms *Markov property*, *filtering* and *stationary distribution* refer to in the context of *Hidden Markov Models*? [30%]

a) Markov property : s_n is independent of $s_{1:n-2}$ given s_{n-1} i.e.

equivalently $p(s_{1:n}) = p(s_1) \prod_{n=2}^{N} p(s_n|s_{n-1})$

Filtering : forming the posterior distribution over the hidden state s_n given the observations up to that time point $y_{1:n}$ ie

$p(s_n|y_{1:n})$

Stationary distribution : if the marginal distribution of the hidden & observed variables in a chain, $p(y_n, s_n)$, converges as $n \to \infty$ then $p(y_\infty, s_\infty)$ is the stationary distribution.

(b) A probabilistic model for a time-series containing binary valued observations y_n employs binary state variables s_n. The transition matrix and emission matrix of the model are denoted

$$T = \begin{bmatrix} p(s_{n+1}=0|s_n=0) & p(s_{n+1}=0|s_n=1) \\ p(s_{n+1}=1|s_n=0) & p(s_{n+1}=1|s_n=1) \end{bmatrix}, \quad E = \begin{bmatrix} p(y_n=0|s_n=0) & p(y_n=0|s_n=1) \\ p(y_n=1|s_n=0) & p(y_n=1|s_n=1) \end{bmatrix}.$$

The forward filtering recursions have been used to process N observations, $y_{1:N}$, in order to return the posterior distribution over the Nth state variable,

$$\rho = \begin{bmatrix} p(s_N=0|y_{1:N}) \\ p(s_N=1|y_{1:N}) \end{bmatrix}.$$

(i) Explain how to transform the posterior distribution over the Nth state into a forecast for the observations one time step into the future, i.e. express $p(y_{N+1}|y_{1:N})$ in terms of ρ. [25%]

b i) $p(y_{N+1}|y_{1:N}) = \underline{\underline{E}}\,\underline{\underline{T}}\,\underline{\rho}$

since $\underline{\underline{T}}\,\underline{\rho} = p(s_{N+1}|y_{1:N})$

(ii) Now provide a forecast for the observations τ time steps into the future by expressing $p(y_{N+\tau}|y_{1:N})$ in terms of ρ. [15%]

ii) $p(y_{N+\tau}|y_{1:N}) = \underline{\underline{E}}\,\underline{\underline{T}}^{\tau}\underline{\rho}$

Since $\underline{\underline{T}}^{\tau}\underline{\rho} = p(s_{N+\tau}|y_{1:N})$

(iii) Compute the forecast $p(y_{N+\tau}|y_{1:N})$ in the limit $\tau \to \infty$ when

$$T = \begin{bmatrix} 3/4 & 1/2 \\ 1/4 & 1/2 \end{bmatrix},\ E = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}.$$

Explain your reasoning. [30%]

iii) in the limit the forecast will tend to that given by the stationary distribution of the chain. We find the stationary distribution as follows:

Let $p(s_\infty) = \underline{\alpha}$ where $\underline{\alpha} = \begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix}$ ($p(s_\infty = 0)$, $p(s_\infty = 1)$)

Then $\underline{\alpha} = \underline{\underline{T}}\,\underline{\alpha}$ (definition of stationary dist)

$$\Rightarrow \begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix} = \begin{bmatrix} 3/4 & 1/2 \\ 1/4 & 1/2 \end{bmatrix}\begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix} = \begin{bmatrix} 3/4\,\alpha + 1/2(1-\alpha) \\ 1/4\,\alpha + 1/2(1-\alpha) \end{bmatrix}$$

$$\Rightarrow \left(1/4 + 1/2\right)\alpha = 1/2$$

$$\Rightarrow \alpha = 2/3$$

$$\therefore p(s_\infty) = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

$$\therefore p(y_\infty) = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}\begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} \frac{6}{12} + \frac{1}{12} \\ \frac{2}{12} + \frac{3}{12} \end{bmatrix} = \frac{1}{12}\begin{bmatrix} 7 \\ 5 \end{bmatrix}$$

($p(y_\infty = 0)$, $p(y_\infty = 1)$)

(Makes sense as transitions from $s_t = 0$ to $s_{t+1} = 0$ are more probable than transitions from $s_t = 1$ to $s_{t+1} = 1$ ($3/4$ vs $1/2$) & the emission probability from the two states are $\begin{bmatrix} 3/4 \\ 1/4 \end{bmatrix}$ vs $\begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}$.)

Assessor's comments: Well answered in the main. A number of candidates could not define what filtering was. Some made analytic errors when calculating the stationary distribution of the hidden variables.

END OF PAPER