Version JMHL/5

EGT2
ENGINEERING TRIPOS PART IIA

Wednesday 26 April 2023 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.

You may not remove any stationery from the Examination Room.

1 A data scientist uses a generative classifier to predict the binary label $y \in \{0, 1\}$ from the scalar feature $x \in \mathbb{R}$. The forms of $p(y)$ and $p(x|y)$ are known and given by

$$p(y) = \begin{cases} 0.4 & \text{if} \quad y = 1 \\ 0.6 & \text{if} \quad y = 0 \end{cases}, \qquad p(x|y) = \mathcal{N}(x|m_y, 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - m_y)^2}{2}\right\},$$

where m_0 and m_1 are two scalar mean parameters.

(a) Assuming that m_0 and m_1 are known, write down an expression for the classifier's predictive probability $p(y = 1|x)$. Write your expression using the notation $\mathcal{N}(x|m, v)$ to represent a Gaussian density with mean m and variance v evaluated at x. [25%]

(b) The data scientist introduces the following prior for m_0 and m_1:

$$p(m_0, m_1) = \mathcal{N}(m_0|0, 1) \times \mathcal{N}(m_1|0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{m_0^2}{2}\right\} \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{m_1^2}{2}\right\}.$$

Given a data point (x_n, y_n), write down an expression for the posterior distribution of m_0 and m_1 as a product of Gaussian densities using the notation $\mathcal{N}(x|m, v)$ as before. Do not calculate the value of the normalisation constant for the posterior and write Z instead. [25%]

(c) The data scientist observes a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ containing N data points. According to the generative classifier, the log-likelihood for m_0 and m_1 is

$$L(m_0, m_1) = \sum_{n=1}^{N} -I[y_n = 0] \times \frac{1}{2}(x_n - m_0)^2 - \sum_{n=1}^{N} I[y_n = 1] \times \frac{1}{2}(x_n - m_1)^2 + \text{constant},$$

where $I[\cdot]$ is equal to 1 if its input is true and 0 otherwise. Using this result, write down an expression for the maximum a posteriori estimates of m_0 and m_1 given $\mathcal{D}$. [25%]

(d) The maximum a posteriori estimates of m_0 and m_1 are given by $\hat{m}_0 = 1$ and $\hat{m}_1 = -1$. The data scientist then uses the classifier to make predictions using these estimates and the following loss (negative reward) function:

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if} \quad y = 0 \quad \text{and} \quad \hat{y} = 0 \\ 2 & \text{if} \quad y = 0 \quad \text{and} \quad \hat{y} = 1 \\ 1 & \text{if} \quad y = 1 \quad \text{and} \quad \hat{y} = 0 \\ 0 & \text{if} \quad y = 1 \quad \text{and} \quad \hat{y} = 1 \end{cases},$$

where y is the true class label and $\hat{y}$ is the classifier's prediction. What is the classifier's decision boundary in this case? Hint: The decision boundary is the input x at which predicting one class or the other yields the same expected loss. [25%]

2 A product of experts model makes predictions for a quantity of interest by multiplying the predictive densities of individual experts. Consider a regression problem and a total of E experts given by linear models with additive Gaussian noise, parameter vectors $\mathbf{w}_1, \ldots, \mathbf{w}_E$ and non-linear feature transformation functions $\phi_1(\cdot), \ldots, \phi_E(\cdot)$, respectively. In this case, the product of experts predictions are given by

$$p(y_n|\mathbf{x}_n) \propto \prod_{e=1}^{E} \mathcal{N}(y_n|\mathbf{w}_e^{\mathrm{T}}\phi_e(\mathbf{x}_n), \sigma_e^2)\,,$$

where σ_e^2 is the variance of the zero-mean additive Gaussian noise assumed by expert e.

(a) It can be shown that $p(y_n|\mathbf{x}_n)$ is Gaussian:

$$p(y_n|\mathbf{x}_n) = \mathcal{N}(x|m, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-m)^2}{2v}\right).$$

Give an expression for m and v in terms of the means and variances of the individual experts, that is, $\mathbf{w}_e^{\mathrm{T}}\phi_e(\mathbf{x}_n)$ and σ_e^2, $e = 1, \ldots, E$. For this, use the following property of the product of Gaussian densities: $\mathcal{N}(x|m_1, v_1) \times \mathcal{N}(x|m_2, v_2) \propto \mathcal{N}(x|m_3, v_3)$, where $v_3^{-1} = v_1^{-1} + v_2^{-1}$ and $m_3/v_3 = m_1/v_1 + m_2/v_2$. Will v increase or decrease as one more expert is added into the model? [25%]

(b) Under some specific choice of transformation functions $\phi_1, \ldots, \phi_E$, the training error of a product of experts model that is fitted to the data by maximum likelihood can be made arbitrarily small by just increasing the number of experts in the model. However, using a very large number of experts may not be a good idea in practice. Why? [25%]

(c) Consider now an alternative mixture of experts model that uses the same experts as the model above and has uniform mixing weights. The predictive distribution is now

$$p(y_n|\mathbf{x}_n) = \frac{1}{E}\left\{\sum_{e=1}^{E} \mathcal{N}(y_n|\mathbf{w}_e^{\mathrm{T}}\phi_e(\mathbf{x}_n), \sigma_e^2)\right\}.$$

What are the mean and variance of this predictive distribution? [25%]

(d) Jensen's inequality says that for a convex function φ, numbers $x_1, \ldots, x_n$ in its domain, and positive weights a_i, $\varphi(\sum_i a_i x_i / \sum_i a_i) \leq \sum a_i \varphi(x_i) / \sum_i a_i$. Using this and the fact that the quadratic function is convex, or otherwise, show that the variance of the predictive distribution for the previous mixture of experts model has to be greater than or equal to the minimum of $\sigma_1^2, \ldots, \sigma_E^2$. Using this fact and the response to part (a) above, justify why the product of experts model may be preferred over the mixture of experts one. [25%]

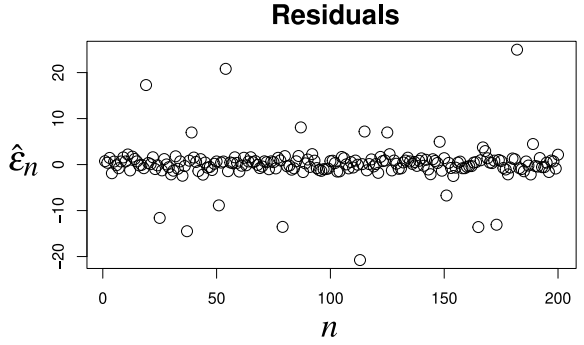
3 Consider a regression dataset $\{(x_n, y_n)\}_{n=1}^N$ and a linear regression model with coefficients $\boldsymbol{\theta} \in \mathbb{R}^d$ and Gaussian additive noise with variance σ^2. Let $\boldsymbol{\Omega} = \{\boldsymbol{\theta}, \sigma^2\}$ be the set of model parameters. The likelihood for the n-th data point is then given by

$$p(y_n|x_n, \boldsymbol{\Omega}) \;=\; \mathcal{N}(y_n|f(x_n;\boldsymbol{\theta}), \sigma^2)\,, \quad \text{where} \quad f(x_n;\boldsymbol{\theta}) \;=\; \sum_{i=1}^{d} \phi_i(x_n)\theta_i$$

with the ϕ_i being non-linear basis functions. The model is fitted with respect to both $\boldsymbol{\theta}$ and σ^2 using maximum likelihood, giving the following estimates for these parameters:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \;=\; (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}\,, \qquad \hat{\sigma}^2_{\text{ML}} \;=\; \frac{1}{N}\sum_{n=1}^{N}(y_n - f(x_n;\hat{\boldsymbol{\theta}}_{\text{ML}}))^2\,,$$

where $\boldsymbol{\Phi}$ is the $N \times d$ matrix given by $[\boldsymbol{\Phi}]_{ni} \;=\; \phi_i(x_n)$ and $\mathbf{y}$ is an N-dimensional vector containing all the y_n. After fitting the model, the training residuals $\hat{\varepsilon}_n = y_n - f(x_n;\hat{\boldsymbol{\theta}}_{\text{ML}})$ are calculated and it is noticed that a small proportion of them have much larger magnitude than the others as shown in the following plot:



(a) Explain the negative effect that these large residuals will have in $\hat{\boldsymbol{\theta}}_{\text{ML}}$ and $\hat{\sigma}^2_{\text{ML}}$ as compared to the case when no such large residuals are present in the data. [25%]

(b) To alleviate the problems caused by the large residuals, a latent binary variable h_n is introduced for each data point to distinguish between regular magnitude ($h_n = 0$) and large magnitude ($h_n = 1$) residual data points. The same f is used for both types of data points. Write an expression for the new likelihood $p(y_n|x_n, \boldsymbol{\Omega}, h_n)$ and a probability distribution for the prior $p(h_n)$. What parameters, in addition to $\boldsymbol{\theta}$, does this new model have? [25%]

(c) The new model is fitted using the EM algorithm. State the E-step. [25%]

(d) State the M-step for $\boldsymbol{\theta}$ when all the other model parameters are kept fixed. [25%]

4 A scalar linear dynamical system evolves according to

$$x_t = \lambda_1 x_{t-1} + \lambda_2 x_{t-2} + \sigma\epsilon_t, \tag{1}$$

where the ϵ_t are independent and identically distributed with $\epsilon_t \sim \mathcal{N}(0, 1)$.

(a) What is the Markov order of this system, and why? [20%]

(b) If λ_1 were set to be zero in equation (1), what would then be the relationship between states x_t at odd and even times? [20%]

(c) Re-write the system in equation (1) as a first order system in a new vector-valued variable $\mathbf{z}_t$ such that

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{e}_t, \quad \text{where} \quad \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}) .$$

In this expression, $\mathbf{A}$ and $\mathbf{Q}$ are matrices and $\mathbf{z}_t$ and $\mathbf{e}_t$ vectors. Write $\mathbf{z}_t$, $\mathbf{A}$ and $\mathbf{Q}$ in terms of quantities from equation (1). [20%]

(d) A dynamical system is unstable if its state grows without bound with time. What are the conditions of stability for the previous system in terms of the eigenvalues of $\mathbf{A}$? [20%]

(e) An observation vector $\mathbf{y}_t$ is obtained from $\mathbf{z}_t$ as follows:

$$\mathbf{y}_t = \mathbf{B}\mathbf{z}_t + \mathbf{s}_t , \quad \text{where} \quad \mathbf{s}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{T})$$

and $\mathbf{B}$ and $\mathbf{T}$ are matrices. Write down an expression for $p(\mathbf{y}_t|\mathbf{z}_{t-1})$. [20%]

END OF PAPER

THIS PAGE IS BLANK