THE UNIVERSITY OF MELBOURNE

Student number

Semester 2 Assessment, 2021

School of Mathematics and Statistics

MAST90083 Computational Statistics & Data Science

Reading time: 30 minutes — Writing time: 3 hours — Upload time: 30 minutes

This exam consists of 4 pages (including this page)

Permitted Materials

- This exam and/or an offline electronic PDF reader, blank loose-leaf paper and a non-programmable calculator.
- No books or other material are allowed. Only one double side A4 page note (handwritten or printed) is allowed.

Instructions to Students

- If you have a printer, print the exam. If using an electronic PDF reader to read the exam, it must be disconnected from the internet. Its screen must be visible in Zoom. No mathematical or other software on the device may be used. No file other than the exam paper may be viewed.
- Ask the supervisor if you want to use the device running Zoom.

Writing

- There are 6 questions with marks as shown. The total number of marks available is 55.
- Write your answers on A4 paper. Page 1 should only have your student number, the subject code and the subject name. Write on one side of each sheet only. Each question should be on a new page. The question number must be written at the top of each page.

Scanning

- Put the pages in question order and all the same way up. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4. Make sure that you upload the correct PDF file and that your PDF file is readable.

Submitting

- **You must submit while in the Zoom room.** No submissions will be accepted after you have left the Zoom room.
- Go to the Gradescope window. Choose the Canvas assignment for this exam. Submit your file. Wait for Gradescope email confirming your submission. Tell your supervisor when you have received it.

Question 1 (10 marks)

Given the model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ is full rank p and $\boldsymbol{\epsilon} \in \mathbb{R}^n \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ be the column representation of X where we further assume that the columns are mutually orthogonal.

(a) Derive the expression of, $\hat{\beta}_j$ the j^{th} component of the least square estimate of $\boldsymbol{\beta}$, as a function of $\mathbf{x}_j$

(b) Is the least square estimate of β_j modified if any of the other components β_l, $(l \neq j)$ is forced to zero ?

(c) Provide the expression of the residual sum of squares and discuss its change when a component is put to zero, $\beta_j = 0$

Assume now that instead of having orthogonal columns, the x_{ij} are standardized so that for $j = 1, ..., p$

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = c$$

(d) Derive the expression of the covariance of the least square estimator of $\boldsymbol{\beta}$

(e) Derive the expression of $\sum_{j=1}^{p} \text{var}\left(\hat{\beta}_j\right)$ as a function of σ^2 and λ_j, $j = 1, ..., p$, the eigenvalues of $C = X^\top X$

(f) Use these results to show that $\sum_{j=1}^{p} \text{var}\left(\hat{\beta}_j\right)$ is minimized when X is orthogonal

Note: For parts (a)-(c), the columns of X are assumed orthogonal and not orthonormal; i.e., $\mathbf{x}_i^\top \mathbf{x}_j = 0$ but $\mathbf{x}_i^\top \mathbf{x}_i = \|\mathbf{x}_i\|^2 \neq 1$

Question 2 (9 marks)

Consider a positive sample $x_1, ..., x_n$ from an exponential distribution

$$f(x|\theta) = \theta e^{-\theta x}, \quad x \geq 0, \quad \theta > 0.$$

Suppose we have observed $x_1 = y_1, ..., x_m = y_m$ and $x_{m+1} > c, ..., x_n > c$ where m is given, $m < n$ and $y_1, .., y_m$ are given numerical values. This implies that $x_1, \cdots, x_m$ are completely observed whereas $x_{m+1}, \cdots, x_n$ are partially observed in that they are right-censored. We want to use an EM algorithm to find the MLE of θ.

(a) Find the complete-data log-likelihood function $\ell(\theta) = \log L(\theta)$.

(b) In the E-step, we calculate

$$Q(\theta, \theta^{(k)}) = \text{E}[\ln L(\theta) \mid x_1 = y_1, \cdots, x_m = y_m, x_{m+1} > c, \cdots, x_n > c; \theta^{(k)}]$$

where $\theta^{(k)}$ is the current estimate of θ. Show that

$$Q(\theta, \theta^{(k)}) = n \log \theta - \theta \left[\sum_{i=1}^{m} y_i + (n-m) \left(\frac{c\theta^{(k)} + 1}{\theta^{(k)}} \right) e^{-\theta^{(k)} c} \right]$$

(c) In the M-step, we maximise $Q(\theta, \theta^{(k)})$ with respect to θ to find an update $\theta^{(k+1)}$ from $\theta^{(k)}$. Show that

$$\theta^{(k+1)} = n \left[\sum_{i=1}^{m} y_i + (n-m) \left(\frac{c\theta^{(k)} + 1}{\theta^{(k)}} \right) e^{-\theta^{(k)} c} \right]^{-1}$$

(d) Suppose the sequence $\{\theta^{(k)};\, k=1,2,\cdots\}$ converges to the MLE $\hat{\theta}$ when $k \to \infty$. Establish the equation allowing the derivation of $\hat{\theta}$.

Question 3 (9 marks)

Consider scatterplot data (x_i, y_i), $1 \le i \le n$ such that

$$y_i = f(x_i) + \epsilon_i$$

where $y_i \in \mathcal{R}$, $x_i \in \mathcal{R}$, $\epsilon_i \in \mathbb{R} \sim \mathcal{N}(0, \sigma^2)$ and are i.i.d. The function

$$f(x) = E(y|x)$$

characterizing the underlying trend in the data is some unspecified smooth function that needs to be estimated from (x_i, y_i), $1 \le i \le n$. For approximating f we propose to use quadratic spline basis with truncated quadratic functions $1, x, x^2, (x-k_1)_+^2, ..., (x-k_K)_+^2$.

(a) Provide the quadratic spline model for f and define the set of unknown parameters that need to be estimated

(b) Derive the matrix form of the model and the associated penalized spline fitting criterion

(c) Derive the expression for the penalized least squares estimator for the unknown parameters of the model and the associated expression for the best fitted values.

(d) Find the degrees of freedom of the fit (effective number of parameters) obtained with the proposed model and its extremes or limit values when the regularization parameter λ varies from 0 to $+\infty$.

(e) Find the optimism of the fit and its relation with the degrees of freedom.

Question 4 (10 marks)

Let $Y = (\mathbf{y}_1, ..., \mathbf{y}_n)$ be a set of n vector observations of dimension q such that $\mathbf{y}_i = (y_{1i}, ..., y_{qi})^\top \in \mathbb{R}^q$. For modeling these observations we propose to use the parametric model given by

$$\mathbf{y}_i = \Phi_1 \mathbf{y}_{i-1} + \Phi_2 \mathbf{y}_{i-2} + ... + \Phi_p \mathbf{y}_{i-p} + \boldsymbol{\epsilon}_i$$

where ϵ_i are independent identically distributed normal random variables with mean vector zero and $q \times q$ variance-covariance matrix Σ modeling the approximation errors and the Φ_j, $j = 1, ..., p$ are $q \times q$ coefficient or parameter matrices.

(a) How many vector observations need to be lost to work with this model ? And what is the effective number of observation ?

(b) Provide a linear matrix form for the model where the parameters are represented in a $(pq) \times q$ matrix form $\Phi = [\Phi_1, ..., \Phi_p]^\top$, derive the least square estimator of Φ and the maximum likelihood estimate of Σ

(c) What could you describe as an inconvenience of this model and find the number of parameters involved in the model

(d) Derive the expression of the log-likelihood for this model

(e) Use the obtained log-likelihood expression to obtain the expressions of AIC and BIC

(f) What consequences this model has on selection criteria ?

Question 5 (9 marks)

Let $x_1, ..., x_n$ be a set of independent and identically distributed samples from a population distribution F_0 and let

$$\mu = \int x dF_0(x)$$

denotes the mean of this population assumed to be a scalar. We are interested in $\theta_0 = \theta(F_0) = \mu^2$

(a) Provide the form of the nonparametric estimator obtained from the empirical distribution F_1

(b) Derive the expression of the bias $b_1 = \mathbb{E}\left(\hat{\theta} - \theta_0\right)$.

(c) Derive the expression of the bootstrap estimate of b_1.

(d) Use this expression to derive the bootstrap bias-reduced estimate $\hat{\theta}_1$ of θ

(e) Derive the expression of the bias $b_2 = \mathbb{E}\left(\hat{\theta}_1 - \theta_0\right)$

(f) Compare b_1 and b_2

Question 6 (8 marks)

Suppose we have a two-layer network with r input nodes x_m, $m = 1, ..., r$, a single layer $(L = 1)$ of t hidden nodes Z_j, $j = 1, ..., t$ and s output nodes Y_k, $k = 1, ..., s$. Let β_{mj} be the weight of the connection $X_m \rightarrow Z_j$ with bias β_{0j} and let α_{jk} be the weight of the connection $Z_j \rightarrow Y_k$ with bias α_{0k}. The functions $f_j(.)$, $j = 1, ..., t$ and $g_k(.)$, $k = 1, ..., s$ are the activation functions for the hidden and output layers nodes respectively.

(a) Derive the expression for the value of the k^{th} output node of the network as function of α_{0k}, g_k, α_{jk}, f_j, β_{0j}, β_{mj} and X_m

(b) Derive the matrix form for the vector of output of the network

(c) Under which conditions this network becomes equivalent to a single-layer perceptron

(d) What is the special case model obtained when the activation functions for the hidden and output nodes are taken to be identity functions.

End of Exam — Total Available Marks = 55