THE UNIVERSITY OF MELBOURNE

| Student number | |
|---|---|

Semester 2 , 2020

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 and blank loose-leaf paper.
- No books, notes or other printed or handwritten material are permitted.
- Calculators are not permitted.

### Instructions to Students

- There are 6 questions with marks as shown. The total number of marks available is 80.
- During writing time you may only interact with the device running the Zoom session with supervisor permission. The screen of any other device must be visible in Zoom from the start of the session.



- Write your answers on A4 paper. The first page should contain only your student number, the subject code and the subject name. Write on one side of each sheet only. Start each question on a new page and include the question number at the top of each page.



- Assemble your single-sided solution pages in correct order and the correct way up. Use a mobile phone scanning application to scan all pages to a single PDF file. Scan from directly above to reduce keystone effects. Check that all pages are clearly readable and cropped to the A4 borders of the original page. Poorly scanned submissions may be impossible to mark.



- Submit your PDF file to the Canvas Assignment corresponding to this exam using the Gradescope window. Before leaving Zoom supervision, confirm with your Zoom supervisor that you have Gradescope confirmation of submission.

**Question 1 (12 marks)**

Suppose that you collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. Then, you fit a **linear regression model** to the data, as well as a separate **cubic regression**

(a) Suppose that the true relationship between $X$ and $Y$ is linear. Consider the training residual sum of squares (RSS) for the **linear regression**, and also the training RSS for the **cubic regression**. Would you expect one to be lower than the other, would you expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between $X$ and $Y$ is not linear, but you don't know how far it is from linear. Consider the training RSS for the **linear regression**, and also the training RSS for the **cubic regression**. Would you expect one to be lower than the other, would you expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.



**Question 2 (12 marks)**

Consider **Quadratic Discriminant Analysis** (QDA) model, in which the observations within each class are drawn from a normal distribution with a class specific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$, i.e., there is only one feature. Suppose that we have $K$ classes, and that if an observation belongs to the $k$th class then $X$ comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right).$$

Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

**Hint:** For this problem, you should follow the arguments laid out in

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)},$$

but without making the assumption that $\sigma_1^2 = \cdots = \sigma_k^2$.

**Question 3 (12 marks)**

For parts (a) through (c), indicate which of i. through iv. is correct. Justify your answer.

(a) The **Lasso**, relative to least squares, is:

i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

(b) Repeat (a) for **Ridge regression** relative to least squares.

(c) Repeat (a) for **non-linear methods** relative to least squares.

(d) If variable selection is important for your problem, will you choose **Ridge** or **Lasso**? Why? Justify your answer.

**Question 4 (15 marks)**



As discussed in class, a cubic regression spline with one knot at $\xi$ can be obtained using a basis of the form

$$1, \quad x, \quad x^2, \quad x^3, \quad (x-\xi)_+^3, \quad \text{where } (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi, \\ 0 & \text{otherwise.} \end{cases}$$



You will now try to show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3$$



is indeed a **cubic regression spline**, regardless of the values of $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$.

(a) Find a **cubic polynomial**

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express $a_1$, $b_1$, $c_1$ and $d_1$ in terms of $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$.

(b) Find a **cubic polynomial**

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$

such that f(x)= $f_2$(x) for all x $> \xi$. Express $a_2$, $b_2$, $c_2$ and $d_2$ in terms of $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$ and $\beta_4$. We have now established that $f(x)$ is a **piecewise polynomial**.

(c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at $\xi$.

(d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at $\xi$.

(e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at $\xi$.

Therefore, $f(x)$ is indeed a **cubic spline**.

**Question 5 (16 marks)**
Here you will explore the **Maximal Margin Classifier** on a toy data set.

(a) You are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

| **Obs.** | **X1** | **X2** | **Y** |
|---|---|---|---|
| 1 | 3 | 4 | Red |
| 2 | 2 | 2 | Red |
| 3 | 4 | 4 | Red |
| 4 | 1 | 4 | Red |
| 5 | 2 | 1 | Blue |
| 6 | 4 | 3 | Blue |
| 7 | 4 | 1 | Blue |

Sketch the observations in a 2D graph.

(b) Sketch the optimal separating hyperplane, and provide the equation for this hyperplane.

(c) Describe the classification rule for the **maximal margin classifier**. Provide the values for $\beta_0$, $\beta_1$, and $\beta_2$.

(d) On your sketch, indicate the margin for the maximal margin hyperplane.



(e) Indicate the support vectors for the **maximal margin classifier**.

(f) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.



(g) Sketch a hyperplane that is not the optimal separating hyperplane, and provide the equation for this hyperplane.



(h) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

**Question 6 (13 marks)**
Consider **K-means clustering algorithm**.
(a) Prove equation (6.1) (within-cluster variation for the $k$th cluster):

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{ij} - \overline{x}_{kj}\right)^2 \tag{6.1}$$

(b) On the basis of this identity, argue that the **K-means clustering algorithm** decreases the objective (6.2) at each iteration:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \right\}. \tag{6.2}$$

**End of Exam — Total Available Marks = 80**