IMPERIAL

MATH70049

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
Summer 2025

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Introduction to Statistical Learning

Date: Monday, May 19, 2025

Time: Start time 10:00 – End time 12:30 (BST)

Time Allowed: 2.5 hours

This paper has 4 Questions.

Please Answer All Questions in 1 Answer Booklet

This is a closed book examination.

Candidates should start their solutions to each question on a new sheet of paper.

Supplementary books may only be used after the relevant main book(s) are full.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Allow margins for marking.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO DO SO

1. (a) Suppose β and v are p-dimensional vectors of real numbers. Let $u = \beta^T v$. Prove

$$\frac{\partial u}{\partial \beta} = v. \tag{1}$$

Suppose that A is a symmetric $p \times p$ matrix of real numbers. Let $w = \beta^T A \beta$. Prove

$$\frac{\partial w}{\partial \beta} = 2A\beta. \tag{2}$$

(3 marks)

(b) Suppose Y is an $n \times 1$ response vector, X is an $n \times p$ design matrix of explanatory variables, β a $p \times 1$ vector of parameters and ϵ an $n \times 1$ vector of errors. The linear model is

$$Y = X\beta + \epsilon. \tag{3}$$

List the standard assumptions associated with this model. (1 mark)

(c) Write down the objective criterion associated with ridge regression in matrix/vector form, with penalty parameter of λ. Then solve the optimisation problem and show that the ridge regression estimator is given by

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T Y, \tag{4}$$

where I_p is the p-dimensional identity matrix. (4 marks)

(d) Using any notation already established, briefly explain what principal components regression is and compare and contrast it to ridge regression. (You can assume X is centred).

(5 marks)

(e) Suppose that X is an orthogonal matrix and that $X^T X = I_p$. Derive the least squares estimator and the ridge regression estimator for the orthogonal case, and, additionally, express the latter in terms of the former. [Hint: the formula for the general least squares estimator is $\hat{\beta}^{\text{ls}} = (X^T X)^{-1} X^T Y$.] (2 marks)

Suppose further we now wish to conduct a regression procedure where the objective function consists of the *sum* of ridge regression criterion function you wrote down in part (c) *and* the Lasso penalty $2\mu \sum_{i=1}^{p} |\beta_j|$, where 2μ is the Lasso penalty parameter. Show that the optimal estimator $\hat{\beta}^{\text{net}}$ with this new objective function is given by

$$\hat{\beta}_j^{\text{net}} = (1+\lambda)^{-1}(\hat{\beta}_j^{\text{ls}} - \mu), \tag{5}$$

where the j subscript indicate the jth component and assuming that $\hat{\beta}_j^{\text{ls}} > 0$ and *remembering that it is crucial that* $\hat{\beta}^{\text{net}} > 0$. (8 marks)

Interpret this new estimator and how it compares to both ridge and Lasso estimators.

(2 marks)

(Total: 25 marks)

2. (a) Let C be a set of objects. Suppose $d(g, h)$ is a metric for $g, h \in C$. Let $f > 0$ be a real number. Show that $d(g, h)/f$ is also a metric. (2 marks)

(b) Let B be a set of n-digit binary strings, where $n > 0$. Suppose we have two binary strings $u, v \in B$, where the mutual presence or absence of binary digits 0 or 1 are given by the table

Object		u	
		1	0
v	1	a	b
	0	c	d

The simple matching coefficient is given by $s_M(u, v) = (a + d)/(a + b + c + d)$. Define the simple matching distance by $d_M(u, v) = 1 - s_M(u, v)$ for $u, v \in B$. Show that $d_M(u, v)$ is a metric for all $u, v \in B$. [Hint: If $\{d_\alpha\}_{\alpha \in \mathcal{A}}$ is a family of metrics, then $\sum_{\alpha \in \mathcal{A}} d_\alpha$ is a metric.]
(6 marks)

(c) Let A be a set of objects with attributes, not necessarily binary strings. Given objects $X, Y \in A$, the number of presence/absence attributes that match or not are

Object		X	
		Present	Absent
Y	Present	a	b
	Absent	c	d

So, for example, a is the number of attributes that both objects X and Y possess.

(i) Which of a, b, c or d is/are not part of the Jaccard distance formula and what is the reason for that? Give an example that exemplifies the reason.
(2 marks)

(ii) Define the Jaccard distance $d_J(X, Y)$ for $X, Y \in A$. (1 mark)

(iii) Let $r, s, t \geq 0$ be integers and let $r \leq s$. Prove the Jaccard Triangle Assistance Lemma $r/s \leq (r + t)/(s + t)$. (1 mark)

(iv) Prove that the Jaccard distance is a metric.

[Hint1: draw Venn diagram showing objects X, Y, Z with numbers of attributes that are common or not to each of them individually, as pairs or all three.

Hint2: Define W to be the virtual object defined by $W = (X \cap Y) \cup (Y \cap Z) \cup (X \cap Z)$].
(8 marks)

continued, next page. . .

(d) Let $\{X_i, Y_i\}_{i=1}^n$, be a set of pairs of independently-distributed random variables, where the marginal and joint densities of $X_i, Y_i \in \mathbb{R}$ are $f(x), f(y)$ and $f(x, y)$ for $i = 1, \ldots, n$. Explain how the Nadaraya-Watson estimator

$$\hat{\mathbb{E}}(Y|X = x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)} \tag{6}$$

is derived, where $K_h(x) = h^{-1}K(x/h)$, $K : \mathbb{R} \to [0, \infty)$ is a kernel function and the bandwidth $h > 0$. Explain how the Nadaraya-Watson estimator can be viewed as a distance-based estimator.

(5 marks)

(Total: 25 marks)

3. This question concerns a set of simultaneous audio recordings made over n time points at p different microphones at Dr Dolby's party. Let $X_{i,j}$ is the value of an audio recording signal at time $i = 1, \ldots, n$ for microphone $j = 1, \ldots, p$. Assume that the $X_{.,j}$ variables are centred.

After the party, Dr Dolby wants to know what each speaker was saying, but he cannot decipher the individual conversations after listening to the recording from each microphone, as the speakers are talking over each other. Dr Wiener says that the problem is that the $X_{.,j}$ variables are correlated and that the problem can be solved by representing it as a factor model given by

$$X_{i,1} = a_{1,1}S_{i,1} + a_{1,2}S_{i,2} + \cdots + a_{1,p}S_{i,p} \tag{7}$$
$$X_{i,2} = a_{2,1}S_{i,1} + a_{2,2}S_{i,2} + \cdots + a_{2,p}S_{i,p} \tag{8}$$
$$\vdots \qquad\qquad\qquad \vdots$$
$$X_{i,p} = a_{p,1}S_{i,1} + a_{p,2}S_{i,2} + \cdots + a_{p,p}S_{i,p}, \tag{9}$$

for $i = 1, \ldots, n$ and insists that the variables $S_j = (S_{1,j}, S_{2,j}, \ldots, S_{n,j})$ for $j = 1, \ldots, p$ are uncorrelated. Dr Wiener says that the $S_{i,j}$ over time $i = 1, \ldots, n$ might be able to retrieve clear speech.

(a) Let $X = (X_{i,j})$ be the $n \times p$ data matrix containing the signals from all of the microphones and write the singular value decomposition of X as $X = UDV^T$, where U is an $n \times p$ matrix with $U^TU = I_p$, D is a diagonal $p \times p$ matrix and V is a $p \times p$ orthogonal matrix.

(i) Use the singular value decomposition to show how to write X as the factor model $X = SA^T$, by defining S and A in terms of singular value decomposition components. (2 marks)

(ii) Then show that the empirical covariance of S is the identity matrix I_p, which shows that the S variables are uncorrelated. (1 mark)

(iii) Dr Dolby is skeptical of Dr Wiener's idea and says that the variables S are not unique and that there exists a rotated version S^* of S that has the same properties as S. Show that Dr Dolby is correct. (4 marks)

(b) Dr Rényi joins the discussion and says that it would be better to have $q < p$ factors rather than p, since Dr Dolby used many microphones, not many people were at Dr Dolby's party and not all would have been speaking at once. Dr Rényi recommends using independent components analysis working with the following model:

$$X_{i,1} = a_{1,1}S_{i,1} + a_{1,2}S_{i,2} + \cdots + a_{1,q}S_{i,q} + \epsilon_{i,1} \tag{10}$$
$$X_{i,2} = a_{2,1}S_{i,1} + a_{2,2}S_{i,2} + \cdots + a_{2,q}S_{i,q} + \epsilon_{i,2} \tag{11}$$
$$\vdots \qquad\qquad\qquad \vdots$$
$$X_{i,p} = a_{p,1}S_{i,1} + a_{p,2}S_{i,2} + \cdots + a_{p,q}S_{i,q} + \epsilon_{i,p}. \tag{12}$$

The $\epsilon_{i,p}$ are zero mean and uncorrelated errors and that the S variables have to be *independent*.

continued, next page...

(i) *(Part b continued ...)* Dr Rényi says that there is no point in considering S to be Gaussian as is often assumed in factor analysis. Why? (1 mark)

(ii) Dr Rényi then says that the X matrix needs to be sphered. Explain what this means. (1 mark)

(c) Let $g : \mathbb{R} \to [0, \infty)$ be a probability density function. The entropy of Y with probability density g is defined to be

$$H(Y) = H(g) = -\int_{\mathbb{R}} g(x) \log\{g(x)\}\, dx. \tag{13}$$

The joint entropy of two random variable X, Y with densities f, g is given by

$$H(X, Y) = -\int \int f(x, y) \log\{f(x, y)\} dxdy. \tag{14}$$

Show that $H(X, Y) = H(X) + H(Y)$, if X, Y are independent. (2 marks)

The conditional entropy of X given Y is

$$H(X|Y) = -\int \int f(x, y) \log\{f(x|y)\}\, dxdy. \tag{15}$$

Show that $H(X|Y) = H(X, Y) - H(Y)$. (2 marks)

What is $H(X|Y)$, if X, Y are independent random variables? (1 mark)

(d) Dr Dolby says that they like to use a mode-counting estimator when performing independent components analysis, not the entropy in the continuous random variable setting. That is, given a density, the criterion is to count the number of modes. Dr Dolby says that's good because many modes is interesting. Dr Rényi disagrees. Which estimator would you prefer? Explain your reasoning. (2 marks)

(e) In discrete exploratory projection pursuit, the discrete entropy is given by

$$G(X) = G(p) = -\sum_{i=1}^{n} p_i \log(p_i), \tag{16}$$

where the random variable X can take on any of the values x_i with probability p_i, $i = 1, \ldots, n$. Show that the discrete entropy is maximised by the uniform distribution $p_i = n^{-1}$, for $i = 1, \ldots, n$. [Hint: assume that any second derivatives that you might come across result in a maximum for G] (4 marks)

(f) A survey of the area around the Chew Valley Lake in Somerset, UK, was carried out by an aircraft. The aircraft scanned the ground at six frequencies simultaneously as it flew over Chew Valley Lake and the town of Bishop Sutton. The scanning frequencies were both in the visible light range and infra-red. Figure 1 shows an image of the ground in the first frequency band. The image shows part of the Lake, the town, some buildings and mostly agricultural fields.

continued, next page...

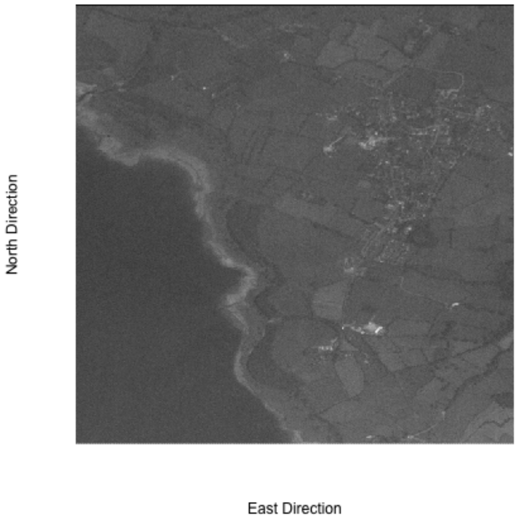

Figure 1: Single band grey scale image of the Chew Valley Lake region in Somerset, UK. The lake is the dark expanse to the bottom left of the image. The town of Bishop Sutton is on the right-hand side mostly towards the top.

(part f, continued) Each image can be represented as a matrix with 512 rows and 512 columns or as a single vector with 262144 entries. Each image vector can be stacked together to form a 262144×6 data matrix X. Independent Component Analysis was applied to X, which found an unmixing matrix W to form independent components $S = XW$. For the Chew Valley set, X, the unmixing matrix for the first THREE independent components was:

$$W_{6\times 3} = \begin{pmatrix} -12.2 & -2.8 & 1.7 & 9.2 & 23.5 & -23.6 \\ 90.8 & 52.2 & 23.7 & -5.8 & -65.3 & 2.4 \\ 156.3 & 100.9 & 56.6 & 15.8 & -58.3 & -1.5 \end{pmatrix}^T . \tag{17}$$

(i) Provide an interpretation of the W matrix in the context of this imaging problem. (2 marks)

(ii) The other five 'original' images (not shown) are similar to the first in Figure 1. Figure 2 shows the first three independent components. Compare and contrast the 3 independent component images to the original and briefly explain any observations. The second and third components show some bright spots in similar locations. Suggest what the bright spots might correspond to on the ground and why there are not more of them in the parts of the image corresponding to the town. (3 marks)

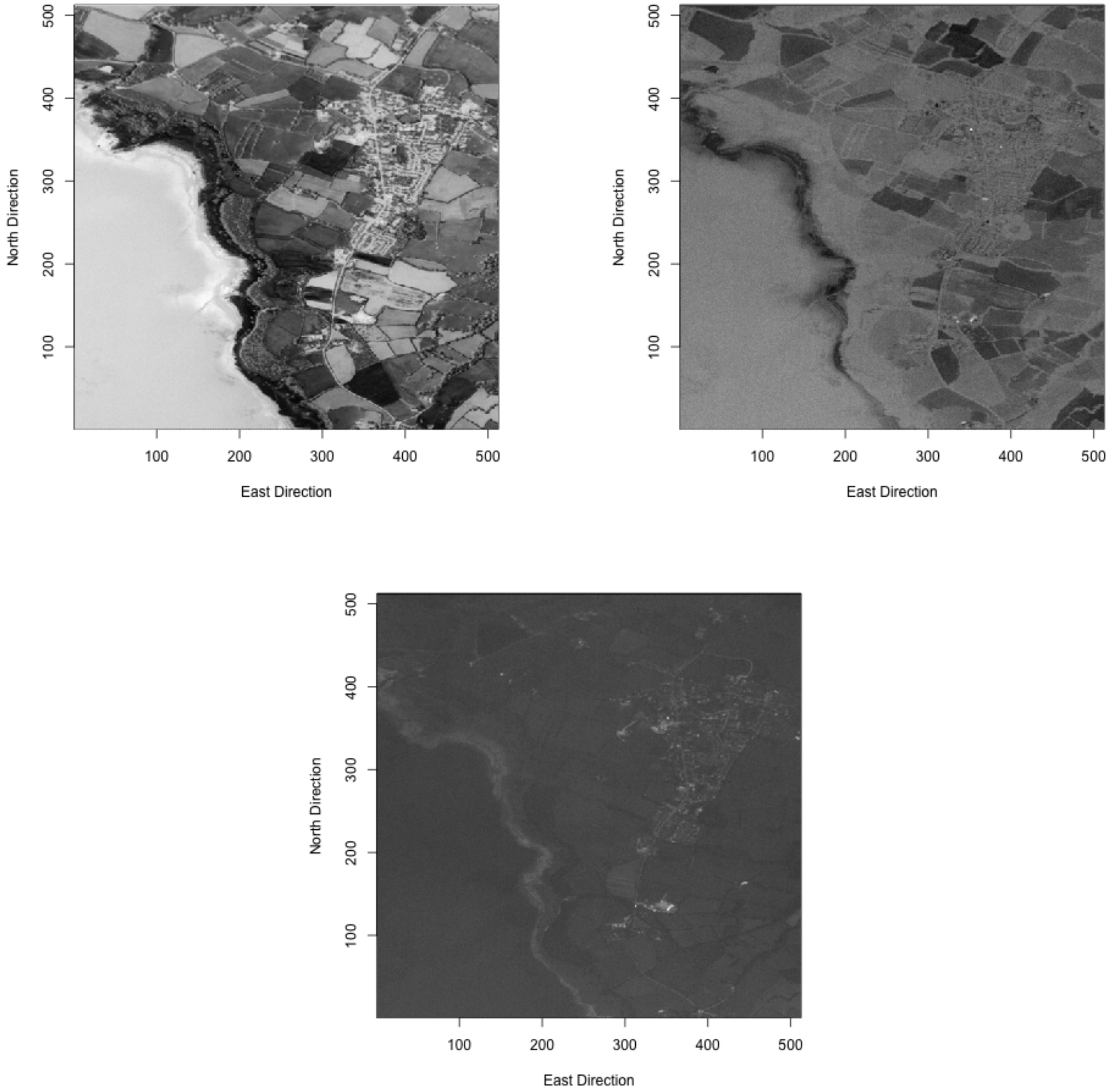

Figure 2: First three output dimensions after applying `fastICA` to the full six dimensions (frequency bands) of the Chew Valley Lake data. Top left: first component; top right: second component; bottom: third component.

(Total: 25 marks)

4. MASTERY QUESTION

A single layer perceptron for classification model produces a set of outputs, $f_k(X)$, based on a set of inputs, $X_1, \ldots, X_p$, which contain real numbers with n entries (n data points), where

$$Z_m = \sigma(\alpha_{0,m} + \alpha_m^T X), \quad m = 1, \ldots, M, \tag{18}$$

$$T_k = \beta_{0,k} + \beta_k^T Z, \quad k = 1, \ldots, K, \tag{19}$$

$$f_k(X) = g_k(T), \quad k = 1, \ldots, K, \tag{20}$$

where $Z = (Z_1, \ldots, Z_M)$ and $T = (T_1, \ldots, T_K)$ for some integers $M, K > 0$, where $\sigma(v) = 1/\{1 + \exp(-v)\}$ is the logistic sigmoid activation function, $g_k(T) = \exp(T_k)/\{\sum_{\ell=1}^K \exp(T_\ell)\}$ is the softmax function, and weights $\{\theta\}$, are:

$$\{\alpha_{0,m}, \alpha_m : m = 1, \ldots, M\} \qquad [M(p+1)\text{weights}], \tag{21}$$

$$\{\beta_{0,k}, \beta_k ; k = 1, \ldots, K\} \qquad [K(M+1)\text{weights}]. \tag{22}$$

We use the usual sum of squares criterion for goodness of fit:

$$R(\theta) = \sum_{k=1}^K \sum_{i=1}^n \{y_{i,k} - f_k(x_i)\}^2, \tag{23}$$

where the $y_{i,k}$ are the relevant training set outputs for $i = 1, \ldots, n, k = 1, \ldots, K$.

(a) Derive the back-propagation equations and explain how the forward and backward pass operations work. (13 marks)

(b) A market analyst decides to reanalyse the Boston housing data using a neural network computed via the `neuralnet` package in R. The Boston data set contains the median value (in $1000s) of owner-occupied homes in 506 suburbs in the Boston area on 14 variables. The marketer divides the data at random into a training set (75%) and a test set (the remaining 25%). For comparison, the marketer fits the median value to a set of variables identified as significant in a linear model (LM). Then, after scaling the data, the marketer applies a neural network (called NNA) with two hidden layers with five neurons in the first and three in the second.

A data scientist told the marketer that they would get better results if they applied a bigger neural network. So, the marketer fits a second neural network (called NNB) with 100 first layer neurons and 60 in the second.

Figure 3 shows the predicted values from the three models (LM, NNA and NNB) against the true data from the test set.

(i) Why do all the predictions on the far right, for both the LM and the neural networks, have such high variability? (2 marks)

continued, next page. . .

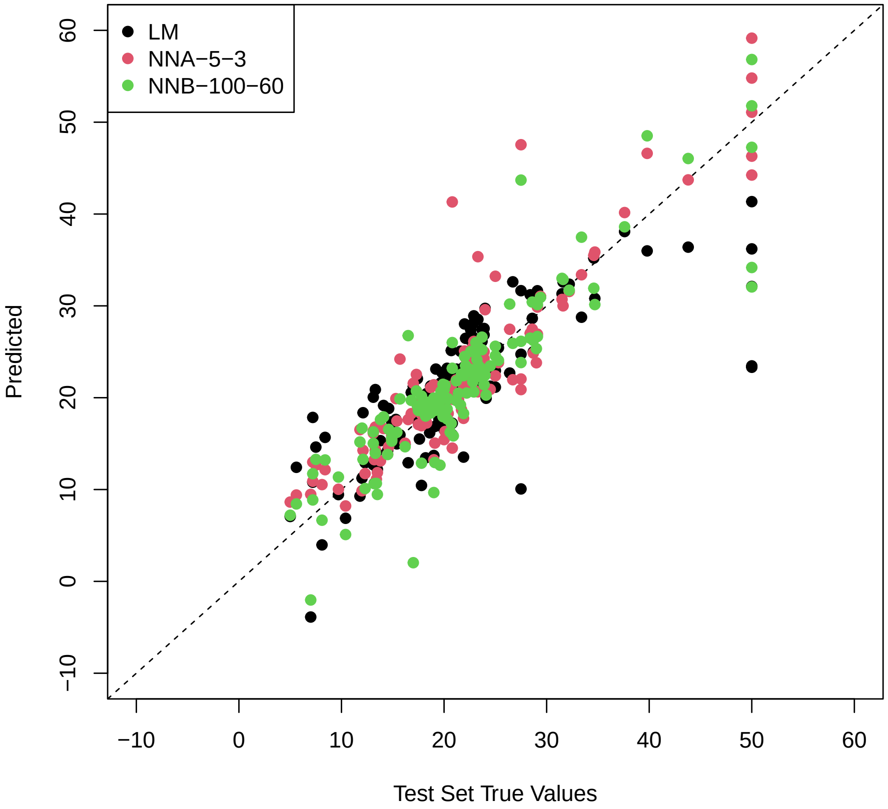

Figure 3: Results of linear model and two neural networks. Plotting predicted value versus actual value on the test set.

(ii) *part b, continued...* The residual sum of squares (RSS) for the three methods is shown in the following table:

Method	LM	NNA	NNB
RSS	63.4	45.5	64.7

The marketer is disappointed that the larger neural networks performs less well than both a simple linear model and the smaller neural network. Suggest an explanation as to why NNB does not perform as well. (2 marks)

(c) (i) The vanishing gradient problem occurs when the derivative of the activation function tends to zero for large positive or negative values of its input. Show that the logistic activation function suffers from this, but the ReLU function defined by

$$\mathrm{ReLU}(x) = \max(0, x) \tag{24}$$

does not. (6 marks)

(ii) Why could a neuron with ReLU activation die? That is, not ever produce any output? (2 marks)

(Total: 25 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2025

This paper is also taken for the relevant examination for the Associateship.

Math 60049/70049

Introduction to Statistical Learning (Solutions)

Setter's signature	Checker's signature	Editor's signature
. . . .	. . . .	. . . .

1. (a) Let $u = \beta^T v = \sum_{i=1}^p \beta_i v_i$, where v is a p-vector. $\boxed{\text{seen} \Downarrow}$

Then

$$\frac{\partial u}{\partial \beta_j} = \sum_{i=1}^p v_i \frac{\partial \beta_i}{\partial \beta_j} = v_j,$$

so $\frac{\partial u}{\partial \beta} = v$. $\boxed{1, \text{A}}$

Let the quadratic form $w = \beta^T A \beta$ for some symmetric matrix A. We can write

$$\begin{align} w &= \sum_{i=1}^p \beta_i \sum_{j=1}^p A_{i,j}\beta_j \tag{1} \\ &= \sum_{i=1}^p \beta_i \left\{ A_{i,i}\beta_i + \sum_{j=1, j\neq i}^p A_{i,j}\beta_j \right\} \tag{2} \\ &= \sum_{i=1}^p A_{i,i}\beta_i^2 + \sum_{i=1}^p \beta_i \sum_{j=1, j\neq i}^p A_{i,j}\beta_j \tag{3} \end{align}$$

For the first term the derivative wrt β_k is $2A_{k,k}\beta_k$. For the second term the derivative wrt β_k is

$$\sum_{j=1, j\neq k} A_{k,j}\beta_j + \sum_{i=1, i\neq k}^p \beta_i A_{i,k} = 2\sum_{j=1, j\neq k}^p A_{j,k}\beta_j, \tag{4}$$

due to symmetry of A. $\boxed{2, \text{A}}$

Hence,

$$\frac{\partial w}{\partial \beta_k} = 2\sum_{j=1}^p A_{k,j}\beta_j. \tag{5}$$

Hence, $\frac{\partial w}{\partial \beta} = 2A\beta$. $\boxed{\text{seen/sim.seen} \Downarrow}$

(b) The assumptions are the $\mathbb{E}(\epsilon) = 0$, $\text{var}(\epsilon_i) = \sigma^2$ constant for $i = 1, \ldots, n$ and that $\{\epsilon_i\}_{i=1}^n$ is a mutually independent set. $\boxed{1, \text{A}}$

(c) The ridge regression objective is $\boxed{\text{seen} \Downarrow}$

$$(Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta. \tag{6}$$

The ridge regression optimization is to minimise the objective over all real p vectors β. We first expand the criterion: $\boxed{1, \text{A}}$

$$R(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta + \lambda \beta^T \beta. \tag{7}$$

To minimise, we differentiate wrt β using our results from part (a) and set to zero.

$$\begin{align} \frac{\partial R(\beta)}{\partial \beta} &= -2X^T Y + 2X^T X\beta + 2\lambda\beta \tag{8} \\ &= -2X^T Y + 2(X^T X + \lambda I_p)\beta = 0 \tag{9} \end{align}$$

Then, solving for β gives $\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T Y$. $\boxed{2, \text{B}}$

The second derivative is $X^T X + \lambda I_p$, which is positive definite for $\lambda > 0$ and so the estimator is at a minimum of the objective function. $\boxed{1, \text{A}}$

(d) To do PC regression, first compute SVD of X, i.e. $X = UDV^T$, where U is $n \times p$, with $U^T U = I_p$, V is a $p \times p$ orthogonal matrix and D is a $p \times p$ diagonal matrix with entires $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$. The covariance matrix of X is $X^T X$ (since X is centred) and

seen ⇓

$$X^T X = V D^T U^T U D V^T = V D I_p D V^T = V D^2 V^T, \tag{10}$$

and since V is orthogonal and D^2 is diagonal this is the eigendecomposition of $X^T X$.

The principal components, z_m, of X are the projection of X onto the columns of V, e.g. define

$$z_m = X v_m, \tag{11}$$

for $m = 1, \ldots, p$.

PC regression can be written as the sum of univariate regressions (since the regression variables are orthogonal) as

$$\hat{y}^{\mathsf{pcr}}_{(M)} = \bar{Y} 1_n + \sum_{m=1}^{M} \hat{\theta}_m z_m, \tag{12}$$

where $\hat{\theta}_m =< z_m, y > / < z_m, z_m >$, the usual regression coefficient, but using $\{z_m\}_{m=1}^{M}$ variables not the Xs. Since $z_m = X v_m$ and $\hat{\theta}_m$ is a scalar we can write:

4, B

$$\hat{y}^{\mathsf{pcr}}_{(M)} = \bar{Y} 1_n + \sum_{m=1}^{M} \hat{\theta}_m X v_m = \bar{Y} 1_n + X \sum_{m=1}^{M} \hat{\theta}_m v_m, \tag{13}$$

so we can think of $\hat{\beta}^{\mathsf{pcr}} = \sum_{m=1}^{M} \hat{\theta}_m v_m$.

1, C

[Note: students might not write down the overall level estimate $\bar{Y} 1_n$, which is fine, since they might assume that the Y are centred too.]

seen ⇓

(e) If X is orthogonal, then the least squares estimator in this case is just $\hat{\beta}^{\mathsf{ls}} = X^T Y$ (using the hint).

1, A

For ridge it is

$$\hat{\beta}^{\mathsf{ridge}} = (1 + \lambda)^{-1} X^T Y = (1 + \lambda)^{-1} \hat{\beta}^{\mathsf{ls}}. \tag{14}$$

1, B

unseen ⇓

The new objective function, width ridge and Lasso penalties is:

$$R^*(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta + \lambda \beta^T \beta + 2\mu \sum_{i=1}^{p} |\beta_i|. \tag{15}$$

We need to minimize R^* over β, this is equivalent to minimising

1, D

$$-2\beta^T \hat{\beta}^{\mathsf{ls}} + (1 + \lambda)\beta^T \beta + 2\mu \sum_{i=1}^{p} |\beta_i|, \tag{16}$$

as X is orthogonal and because $X^T Y$ is the least squares estimator, when X is orthogonal.

1, D

seen/sim.seen ⇓

Formula (16) can be written as sums over components:

$$
\begin{aligned}
& -2\sum_{i=1}^{p}\beta_i\hat{\beta}_i^{\mathsf{ls}} + (1+\lambda)\sum_{i=1}^{p}\beta_i^2 + 2\mu\sum_{i=1}^{p}|\beta_i| && (17)\\
= & \sum_{i=1}^{p}\left(-2\beta_i\hat{\beta}_i^{\mathsf{ls}} + (1+\lambda)\beta_i^2 + 2\mu|\beta_i|\right) && (18)\\
= & \sum_{i=1}^{p} M_i, && (19)
\end{aligned}
$$

where M_i is the term in the sum of (18). Hence, we can minimize the objective function term by term. $\boxed{2,\ \mathrm{C}}$

We are told that $\hat{\beta}_i^{\mathsf{ls}} > 0$, this means that the minimizing β_i has to be positive. If it were negative, then because of the first term in M_i we could reduce the size of M_i by swapping the sign. Hence, because this this we need to minimize

$$M_i = -2\beta_i\hat{\beta}_i^{\mathsf{ls}} + (1+\lambda)\beta_i^2 + 2\mu\beta_i, \tag{20}$$

which we do by differentiating and setting to zero: $\boxed{1,\ \mathrm{C}}$

$$\left.\frac{\partial M_i}{\partial \beta_i}\right|_{\beta=\hat{\beta}^{\mathsf{net}}} = -2\hat{\beta}_i^{\mathsf{ls}} + 2(1+\lambda)\hat{\beta}_i^{\mathsf{net}} + 2\mu = 0. \tag{21}$$

Hence, $\boxed{2,\ \mathrm{D}}$

$$(1+\lambda)\hat{\beta}_i^{\mathsf{net}} = \hat{\beta}_i^{\mathsf{ls}} - \mu, \tag{22}$$

then

$$\hat{\beta}_i^{\mathsf{net}} = (1+\lambda)^{-1}\left(\hat{\beta}_i^{\mathsf{ls}} - \mu\right), \tag{23}$$

$\boxed{1,\ \mathrm{D}}$

This estimator shrinks and also sets coefficients equal to zero just like Lasso, since the estimator can never be negative and gets set to zero if μ is bigger than the least squares estimator. $\boxed{2,\ \mathrm{D}}$

2. (a) This part is unseen, but it is easy. Define $e(g,h) = d(g,h)/f$. Clearly $e(g,h) \geq 0$. Also $e(g,g) = d(g,g)/f = 0$, since d is a metric. Also $e(h,g) = d(h,g)/f = d(g,h)/f = e(g,h)$ as d is symmetric because it is a metric. Finally, the triangle inequality, for $g,h,j \in C$: unseen ⇓

$$\begin{aligned} e(g,h) + e(h,j) &= d(g,h)/f + d(h,j)/f && (24)\\ &= \{d(g,h) + d(h,j)\}/f && (25)\\ &\geq d(g,j)/f && (26)\\ &= e(g,j), && (27) \end{aligned}$$

by the triangle inequality for d, as it is a metric. 2, A

(b) The simple matching distance must be $(b+c)/(a+b+c+d)$. The numerator is the Hanning distance (which was defined and explored in lectures) and the Hanning distance can be written as $\sum_{i=1}^{n} d_i(u,v)$, where d_i examines the distance between ith binary digits by meth seen ⇓

$$d_i(u,v) = \begin{cases} 0 & \text{if } u_i = v_i, \\ 1 & \text{otherwise,} \end{cases} \qquad (28)$$

where u_i, v_i are the ith binary digit of u and v respectively.

One mark for each property of metric shown. Class A for the easy bits, B for the table.

We show d_i is a metric. Clearly, $d_i(u,v) \geq 0$ and $d_i(u,v) = 0$ if $u = v$. Also, clearly d_i is symmetric. 3, A

To show the triangle inequality, we consider the ith binary digit of three strings x, y, z, which each have to be either 0 or 1. Then, we consider each combination of these in turn and use a table to show that the triangle inequality holds in each case (and hence all cases in this situation).

x_i	y_i	z_i	$d_i(x,y)$	$d_i(y,z)$	Sum	$d_i(x,z)$
0	0	0	0	0	0	0
0	0	1	0	1	1	1
0	1	0	1	1	2	0
0	1	1	1	0	1	1
1	0	0	1	0	1	1
1	0	1	1	1	2	0
1	1	0	0	1	1	1
1	1	1	0	0	0	0

2, B

Since d_i is a metric and Hanning is the sum of these (for $i = 1, \ldots, n$) using the hint, this means that the Hanning distance is a metric. For this situation, the simple matching distance with binary strings of length n, is just the Hanning distance divided by n which, according to part(a), is also a metric. 1, B

(c) (i) The integer d is not part of the Jaccard distance formula. seen/sim.seen ⇓

The reason is because d counts the number of times that an attribute does not appear in both objects, and hence is irrelevant to the comparison between them. For example, one would not be interested in the joint absence of iPhones in an archeological grave site. 1, A 1, B

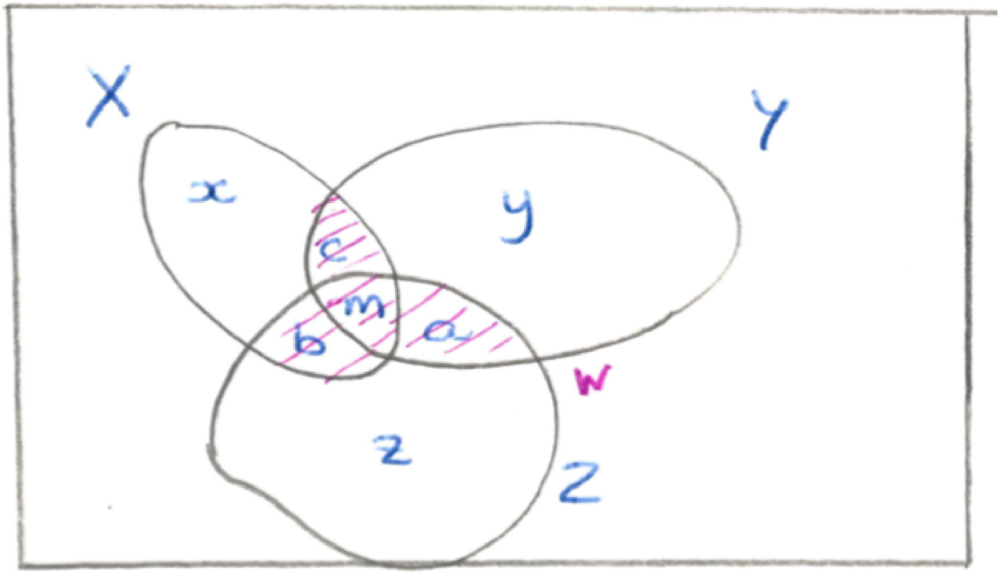

Figure 1: Venn diagram showing numbers of attributes. E.g. x is the number of attributes that only object X possesses, a is the number of attributes that both Y and Z possess, but are not found in X, and so on.

(ii) The Jaccard distance is given by

seen ⇓

$$d_J(x, y) = (b + c)/(a + b + c). \tag{29}$$

1, A

(iii) Let $r, s, t \geq 0$ and $r \leq s$. Show $r/s \leq (r + t)/(s + t)$.
Proof: Since $r \leq s$, we have $0 \leq w = r/s \leq 1$. Hence:

seen ⇓

$$\begin{aligned} w\frac{t}{s} &\leq \frac{t}{s} && (30)\\ \implies w + w\frac{t}{s} &\leq w + \frac{t}{s} && (31)\\ \implies w(1 + t/s) &\leq w + \frac{t}{s} && (32)\\ \implies w &\leq \frac{w + t/s}{1 + t/s} && (33)\\ \implies r/s &\leq (r + t)/(s + t), && (34) \end{aligned}$$

as required.

1, B

(iv) *Seen, but challenging.*

seen ⇓

See example Venn diagram shown in Figure 1.

1, A

Also, recall from Hint2 that $W = (X \cap Y) \cup (Y \cap Z) \cup (X \cap Z)$.
We first prove a version of the triangle inequality for X, Z and W. Using the Venn diagram, we have

$$d_J(X, W) = (x + a)/(x + a + b + c + m), \tag{35}$$

and

2, A

$$d_J(W, Z) = (z + c)/(z + c + a + b + m). \tag{36}$$

Hence,

1, A

$$\begin{align*}
\text{the sum} &\to d_J(X,W)+d_J(W,Z) \tag{37}\\
&= (x+a)/(x+a+b+c+m)+(z+c)/(z+c+a+b+m) \tag{38}\\
&\geq (x+a)/(x+a+b+c+m+z)+(z+c)/(z+c+a+b+m+x) \notag\\
&= (x+c+z+a)/(x+a+b+c+m+z) \tag{39}\\
&= d_J(X,Z). \tag{40}
\end{align*}$$

Secondly, $\boxed{2,\ \mathrm{B}}$

$$\begin{align*}
d_J(X,W) &= (x+a)/(x+a+b+c+m)\\
&\leq (x+a+b)/(x+a+b+c+m)\\
&\leq (x+a+b+y)/(x+y+a+b+c+m)\\
&= d_J(X,Y),
\end{align*}$$

the last two steps use the Assistance Lemma. $\boxed{1,\ \mathrm{A}}$
Similarly, $d_J(Z,W) \leq d_J(Z,Y)$.
Hence $d_J(X,Y)+d_J(Y,Z) \geq d_J(X,W)+d_J(Z,W) \geq d_J(X,Z)$, which shows Jaccard distance satisfies the triangle inequality. $\boxed{1,\ \mathrm{A}}$

(d) Suppose now we have variables X, Y and we want to find $\mathbb{E}(Y|X)$. We proceed via the joint density $f_{X,Y}(x,y)$ by examining $\boxed{\text{seen} \Downarrow}$

$$\begin{align*}
\mathbb{E}(Y|X) &= \int y f(y|x)\,dy \tag{41}\\
&= \int y \frac{f(x,y)}{f(x)}\,dy \tag{42}\\
&= \frac{\int y f(x,y)\,dy}{f(x)}. \tag{43}
\end{align*}$$

Now we make use of our kernel density estimate for X

$$\hat{f}(x) = n^{-1}\sum_{i=1}^{n} K_h(x-X_i), \tag{44}$$

where $K_h(x) = h^{-1}K(x/h)$. We recall the 2D kernel density estimator

$$\hat{f}(x,y) = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)K_h(y-Y_i). \tag{45}$$

So an estimator can be formed by

$$\begin{align*}
\hat{\mathbb{E}}(Y|X=x) &= \frac{\int y\hat{f}(x,y)\,dy}{\hat{f}(x)} \tag{46}\\
&= \frac{\sum_{i=1}^{n} K_h(x-X_i)\int yK_h(y-Y_i)\,dy}{\sum_{i=1}^{n} K_h(x-X_i)}, \tag{47}
\end{align*}$$

and the n^{-1} cancel. Now let $v = y - Y_i$, then

$$\begin{align*}
\int yK_h(y-Y_i)\,dy &= \int (v+Y_i)K_h(v)\,dv \tag{48}\\
&= \cancelto{0}{\int vK_h(v)\,dv} + Y_i\cancelto{1}{\int K_h(v)\,dv}, \tag{49}\\
&= Y_i \tag{50}
\end{align*}$$

since the kernel is symmetric and integrates to 1.

Hence.

$$\hat{\mathbb{E}}(Y|X=x) = \frac{\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)} \tag{51}$$

is the *Nadaraya-Watson* kernel regression. $\boxed{4,\text{ A}}$

The N-W estimator can be seen as distance based as it weights (X_i, Y_i) points that are nearer to x more highly (and less highly further away). $\boxed{\text{unseen} \Downarrow}$ $\boxed{1,\text{ D}}$

3. (a) (i) We can write $S = \sqrt{n}U$, which is $n \times p$ and $A^T = DV^T/\sqrt{n}$.

seen ⇓

2, A

(ii) Since the data matrix X is centred, then so is S. We can write the empirical covariance of S as

$$n^{-1}S^TS = U^TU = I_p. \tag{52}$$

1, A

(iii) Let's rotate the variables of S. To do this, we use a $p \times p$ rotation (orthogonal) matrix R and write $S^* = SR$ and note

$$X = SA^T = SRR^TA^T = S^*(A^*)^T, \tag{53}$$

where $A^* = AR$. So, X can be decomposed into S^*, A^* in the same way as S, A. Moreover, the new S^* representation has empirical covariance given by

2, A

$$n^{-1}(S^*)^TS^* = n^{-1}R^TS^TSR = R^TI_pR = R^TR = I_p. \tag{54}$$

2, B

(b) (i) This is because for S Gaussian there is a direct equivalence between independence and uncorrelated, so we run into the same problem about non-unique factors S as in the previous part.

seen ⇓

1, B

(ii) A sphered data matrix is both centred and has identity variance-covariance matrix.

1, A

(c) If X and Y are independent then

seen/sim.seen ⇓

$$\begin{align}
H(X,Y) &= -\int\int f(x,y)\log\{f(x,y)\}dxdy \tag{55}\\
&= -\int\int f(x)f(y)\log\{f(x)f(y)\}dxdy \tag{56}\\
&= -\int f(x)\int f(y)\left[\log\{f(x)\} + \log\{f(y)\}\right]dxdy \tag{57}\\
&= -\int f(x)\log\{f(x)\}dx\int f(y)dy - \int f(y)dy\int f(x)\log\{f(x)\}dx \notag\\
&= H(X) + H(Y). \tag{58}
\end{align}$$

The conditional entropy of X given Y is

2, C

$$\begin{align}
H(X|Y) &= -\int\int f(x,y)\log\{f(x|y)\}\,dxdy \tag{59}\\
&= -\int\int f(x,y)\log\{f(x,y)/f(y)\}\,dxdy \tag{60}\\
&= -\int\int f(x,y)\log\{f(x,y)\}dxdy + \int\int f(x,y)\log\{f(y)\}dxdy \notag\\
&= H(X,Y) + \int\log\{f(y)\}\int f(x,y)dxdy \tag{61}\\
&= H(X,Y) + \int f(y)\log\{f(y)\}dy \tag{62}\\
&= H(X,Y) - H(Y). \tag{63}
\end{align}$$

Clearly, if X, Y are independent, then $H(X|Y) = H(X,Y) - H(Y) = H(X) + H(Y) - H(Y) = H(X)$, so the information in X given Y is just the information in X if X, Y are independent.

2, C

1, C

(d) The mode counting approach is probably not a good one. This is because one could have many modes (potentially interesting), but they could all be very small. For example, if applied to a kernel density estimate, then a mode could be produced by each data point (if a small bandwidth was used) resulting in a lot of (small) modes. However, this does not reflect the bulk of the data. Such an index would seek out such components. The entropy takes much more account of bulk (because of the integration of mass), but still picks up on multimodality. unseen ⇓ 2, C

(e) There are several ways to prove this. One way is to use Lagrange multipliers: we want to maximise $G(p)$ subject to $\sum_{i=1}^{n} p_i = 1$. The Lagrangian is unseen ⇓

$$L(p) = -\sum_{i=1}^{n} p_i \log(p_i) + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right). \tag{64}$$

Differentiating with respect to p_j gives

$$\frac{\partial L}{\partial p_j} = -\log(p_j) - 1 + \lambda. \tag{65}$$

Setting this equal to zero and solving for p_j implies

$$p_j = \exp(\lambda - 1), \tag{66}$$

for all $j = 1, \ldots, n$. Now maximising wrt λ gives $\sum_{i=1}^{n} p_j = 1$. Substituting (66) gives

$$\begin{aligned} \sum_{i=1}^{n} \exp(\lambda - 1) = 1 \quad &\Longrightarrow \quad n \exp(\lambda - 1) = 1 && (67) \\ &\Longrightarrow \quad \lambda - 1 = \log(n^{-1}) = -\log(n) && (68) \\ &\Longrightarrow \quad \lambda = 1 - \log(n). && (69) \end{aligned}$$

and hence, substituting this back into (66) gives $p_j = n^{-1}$, which is the (discrete) uniform distribution over $1, \ldots, n$. 4, D

[The following is not part of the question or solution, but the bordered Hessian matrix for the Lagrangian can be shown to be as follows. $\frac{\partial L}{\partial p_j \partial p_k} = 0$ for $k \neq j$ and $-p_k^{-1}$ for $j = k$. Further $\frac{\partial L}{\partial p_j \partial \lambda} = 1 = \frac{\partial L}{\partial \lambda \partial p_j}$ and $\frac{\partial L}{\partial \lambda^2} = 0$. So, the Hessian looks like

$$\begin{pmatrix} -p_1^{-1} & 0 & 0 & \cdots & 0 & 1 \\ 0 & -p_2^{-1} & 0 & \cdots & 0 & 1 \\ 0 & 0 & -p_3^{-1} & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -p_n^{-1} & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \tag{70}$$

I believe an inductive approach, or otherwise, shows a maximum.] seen/sim.seen ⇓

(f) (i) The columns of W are contrasts. So, the first col is predominantly a contrast between the 5th original dimension with a weighted average of the first and last dimension (and a 1:2 weight). The second col contrasts a (roughly 3:2:1) weighted average of the first three original dimensions against the fifth, etc. 2, B

(ii) The independent component images seem to be much ‘sharper’ and better highlight different land use types. E.g. there are sharp distinctions between different fields in the first independent component image, whereas they blur into one in the original. This is because the independent components are trying to get away from normality and so prefer components that are non-normal, and often multimodal. With the second and third images this also happens, but some of the modes are linked to the bright spots, which are almost certainly buildings with a certain roof type (e.g. iron). This kind of roof (whatever it is) is not very common in this area (which is mostly tiled roof), and that is probably why there are not more bright spots in the town area.

unseen ⇓

3, D

4. **MASTERY QUESTION** seen ⇓

(a) Minimisation happens by gradient descent, which is called *back propagation* here. The gradient derivation is as follows.

Let $z_{m,i} = \sigma(\alpha_{0,m} + \alpha_m^T x_i)$ from (18) in the exam paper and let $z_i = (z_{1,i}, \dots, z_{M,i})$.

Then

$$R(\theta) = \sum_{i=1}^{n} R_i(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \{y_{i,k} - f_k(x_i)\}^2, \tag{71}$$

with derivatives via chain rule: 1, M

$$\frac{\partial R_i(\theta)}{\partial \beta_{k,m}} = -2\{y_{i,k} - f_k(x_i)\} g_k'(\beta_k^T z_i + \beta_{0,k}) z_{m,i}, \tag{72}$$

$$\begin{aligned}\frac{\partial R_i(\theta)}{\partial \alpha_{m,\ell}} &= -2 \sum_{k=1}^{K} \{y_{i,k} - f_k(x_i)\} g_k'(\beta_k^T z_i + \beta_{0,k}) \beta_{k,m} \\ &\times \sigma'(\alpha_m^T x_i + \alpha_{0,m}) x_{i,\ell}. \end{aligned} \tag{73}$$

4, M

From these derivatives, we can form a back-propagation/gradient descent algorithm by

$$\beta_{k,m}^{(r+1)} = \beta_{k,m}^{(r)} - \gamma_r \sum_{i=1}^{n} \frac{\partial R_i(\theta)}{\partial \beta_{k,m}^{(r)}}, \tag{74}$$

and

$$\alpha_{m,\ell}^{(r+1)} = \alpha_{m,\ell}^{(r)} - \gamma_r \sum_{i=1}^{n} \frac{\partial R_i(\theta)}{\partial \alpha_{m,\ell}^{(r)}}, \tag{75}$$

where γ_r is called the *learning rate*. 1, M

Now write (72) and (73) as $\delta_{k,i} z_{m,i}$ and $s_{m,i} x_{i,\ell}$ respectively, these can be seen as the "errors" of the current model at the output and hidden layers, respectively. 1, M

Putting it all together we can get

$$s_{m,i} = \sigma'(\alpha_m^T x_i + \alpha_{0,m}) \sum_{k=1}^{K} \delta_{k,i} \beta_{k,m} \tag{76}$$

These are called the *back-propagation equations*. 2, M

The algorithm proceeds in two stages:

Forward pass: the current weights are fixed, and the predicted values $\hat{f}_k(x_i)$ are computed from (18) to (20) in the exam paper. Then 2, M

Backward pass: the errors $\delta_{k,i}$ are computed from (72) and then back-propagated via (76) to give the errors $s_{m,i}$ and these form the gradients for the updates in (74) and (75). 2, M

(b) (i) These are points far away from the mean of the data and in areas of low density. Hence, there is little information in those outlying areas. This contrasts to near the centroid of the data, where the high density of points, and having points both left and right to one's location ties down the prediction more securely. In the LM model one can either refer to a 'see-saw' in the regression line or bigger confidence intervals at the end. The NN presumably because there is just so little information to go on in the training set at that location as it is located at an x extreme. seen/sim.seen ⇓ 2, M

(ii) It is probably overfitted, and thus cannot adapt to the test set well. As a rough guide, there are 160 parameters, but the data set only contains $0.75 \times 506 \times 14 = 5313$ observations, so only 33 observations per parameter and even this is being generous as there are correlations in the Boston data that reduce the *effective* number of observations further. $\boxed{\text{unseen} \Downarrow}$ $\boxed{\text{2, M}}$

(c) (i) The logistic activation function is $\sigma(v) = \{1 + \exp(-v)\}^{-1}$. Then $\boxed{\text{unseen} \Downarrow}$

$$\sigma'(v) = -\{1 + \exp(-v)\}^{-2} \times - \exp(-v) \tag{77}$$

$$= \exp(-v)/\{1 + \exp(-v)\}^2. \tag{78}$$

As $v \to \infty$ we have $\exp(-v) \to 0$, and then $\sigma'(v) \to 0$ since the numerator tends to zero, and the denominator tends to one. $\boxed{\text{1, M}}$ For $v \to -\infty$ then $\sigma'(v) \sim \exp(-v)/\exp(-v)^2 = \exp(v) \to 0$. Hence, the logistic sigmoid activation function suffers from the vanishing derivative problem. $\boxed{\text{3, M}}$

For ReLU(v) the derivative is $\mathrm{ReLU}'(v) = \mathbb{I}(v > 0)$, which never vanishes for any $v > 0$ $\boxed{\text{2, M}}$

(ii) If the inputs to the ReLU are consistently negative, then the output is negative and hence the derivative with respect to the parameters will also be negative, unless there is a big shock to the inputs, and the ReLU output will be zero for all of this. Hence, the specific neuron won't be ever activated and hence dead. $\boxed{\text{unseen} \Downarrow}$ $\boxed{\text{2, M}}$

Review of mark distribution:
Total A marks: 30 of 30 marks
Total B marks: 19 of 19 marks
Total C marks: 11 of 11 marks
Total D marks: 15 of 15 marks
Total marks: 100 of 100 marks (including mastery)
Total Mastery marks: 25 of 25 marks

MATH60049 Introduction to Statistical Learning Markers Comments

Question 1

(a) Most students did the derivative of u ok, but a surprising number tripped up on w.

(b) The normal distribution assumption is not really standard for this basic model when. Some students mentioned it as an option, which was fine.

(c) Nearly all students did this well, although a few students had trouble putting the differentiated beta in the right place.

(d) About a third of students did well here. Some more could describe PCR in words.

(e) First part was done very well by most. Many did well on the elastic net bit, although struggled on how to differentiate through the absolute value. Not enough students produced good answers on the interpretation, but marks were given for reasonable attempts.

Question 2

(a) Nearly all students did this part well.

(b) Most students did this well, some struggled on the triangle inequality and tied to make something out of the table, which is not easy to do (better to recognise d as related to the Hamming distance and use the table related to that.

(c)(i) Most did this very well. (ii) Nearly all did this ok. (iii) Most did this, but surprisingly, some struggled or omitted it, given that it is fairly easy (IMHO). (iv) Most could show positivity and symmetry. Several proved the triangle inequality ok, which was pleasing, as it’s challenging. Some again used the table to try and make something, which did not lead anywhere. Some tried to use the hint from (b), but this does not work here as the denominators are not constant over different pairs of objects.

(d) Many did this ok. Some forgot that a bivariate estimator was required.

Question 3

(a)(i) A surprising number used the wrong decomposition, something like UD, and then convinced themselves that the product in the next part was the identity, but it was not true (the maths did not work).
(ii) Most did this if they got the decomposition in (a) right , otherwise, they struggled.
(iii) Most did something sensible here, but many forgot to express the decomposition in terms of the rotated factors.

(b)(i) Several students seemed to misunderstand the question and a wide variety of incorrect answers were supplied (and some right ones).
(ii) Many got this right.

(c) This part was well done by most people. One or two forgot to answer the very last part, which was a shame.

(d) A variety of answers were given here. I gave credit even if, in my view, the answer was wrong, as you might have a reason for counting modes and if you gave it, and if it was sensible you got marks.

(e) Almost nobody got this part right, which was surprising. About five people did get it right as they recognised it as a constrained optimisation. One or two students did so, but then, unfortunately, did a simple differentiation wrong, but they got some credit for this.

(f) Most people engaged well with this question and the answers to part (ii) were better than (i). Some did not mention W in part (i), which was odd.

Question 4

A moderate number of students managed to do part (a) well, a few did very well. Essentially, (a) is about repeated application of the chain rule, keeping the calcs organised and leaving high level functions as g and \sigma. Most students got (b)(i) correct and nearly all (b)(ii). c(i) Most students did this well, but some forgot to work with the derivative. (ii) Most students did this ok, even though it was unseen, which was pleasing.