Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2023	MATH3431-WE01

Title: Machine Learning and Neural Networks III

Time:	2 hours	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Students must use the mathematics specific answer book.

Revision:	

ED01/2023

University of Durham Copyright

SECTION A

Q1 (a) The World Almanac and Book of Facts, 1975 reports the height (in inches) and weight (in pounds) of multiple American women aged 30 to 39. Six of these data points are given in the table below, with height denoted (x) and weight denoted (y).

i	1	2	3	4	5	6
x_i	58	59	60	61	62	63
y_i	115	117	120	123	126	129

We describe the relationship between height and weight with a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

(i) Calculate the Least Squares estimates of the model parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$.

(ii) The estimate s_e of the error standard deviation takes the value 0.345. Construct 95% confidence intervals for β_0 and β_1.

(iii) The simple linear model for this data has an R^2 value of 0.9967. A friend tells you the value of R^2 will get even higher if more data is collected, as

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

and SST will tend to zero as more data is collected.
Is your friend correct? Give a reason for your answer.

(b) Assume we have a problem with a single predictor variable x. We turn it into a problem with two predictor variables by replicating x. In other words, in our training set we will have $x_{i,1} = x_{i,2}$ for all i. Now we fit a linear regression model with *no* intercept, using ridge regression. We want to investigate the effect of this exact collinearity.

(i) Write out the ridge regression optimization problem in this setting.

(ii) Suppose there are only two observed data records, i.e. $n = 2$. Prove that the ridge coefficient estimates are identical.

Q2 Consider a prediction rule $h : \mathbb{R}^d \to \mathbb{R}_+^q$ with $h(x) = (h_1(x), ..., h_q(x))^\top$ which receives inputs $x = (x_1, ..., x_d)^\top \in \mathbb{R}^d$ and which is modeled as a feedforward neural network (NN) with equation

$$h_k(x) = \sigma_2\left(\sum_{j=1}^{c} w_{2,k,j}\sigma_1\left(\sum_{i=1}^{d} w_{1,j,i}x_i\right)\right)$$

for $k = 1, ..., q$. We consider activation functions $\sigma_1(a) = \frac{1}{1+\exp(-a)}$ and $\sigma_2(a) = \log(1 + \exp(a))$. Parameters $c \in \mathbb{N}_+$, and $d \in \mathbb{N}_+$ are considered as known, while the weights $\{w_{\cdot,\cdot,\cdot}\}$ of the NN are unknown. To learn the unknown weights $\{w_{\cdot,\cdot,\cdot}\}$, we specify the loss function

$$\ell(w, z = (x, y)) = \frac{1}{2}\|h(x) - y\|_2^2 = \frac{1}{2}\sum_{k=1}^{q}(h_k(x) - y_k)^2$$

where $z = (x, y)$ denotes an example, $x \in \mathbb{R}^d$ is the input vector (features), and $y = (y_1, ..., y_q)^\top \in \mathbb{R}^q$ is the output vector (targets).

University of Durham Copyright

(a) Perform the forward pass of the back-propagation procedure to compute the activations which may be denoted as $\{a_{t,i}\}$ and outputs which may be denoted as $\{o_{t,i}\}$ at each layer t.

(b) Perform the backward pass of the back-propagation procedure in order to compute the gradient

$$\nabla_w \ell\left(w,(x,y)\right) = \left(\left(\frac{\partial}{\partial w_{1,j,i}}\ell\left(w,(x,y)\right)\right)_{j=1,i=1}^{c,d}, \left(\frac{\partial}{\partial w_{2,k,j}}\ell\left(w,(x,y)\right)\right)_{k=1,j=1}^{q,c}\right)$$

of the loss function $\ell\left(w,z\right)$ with respect to w for any example $z=(x,y)$. Clearly state the steps of the procedure and the quantities

$$\frac{\partial}{\partial w_{1,k,j}}\ell\left(w,(x,y)\right), \text{ and } \frac{\partial}{\partial w_{2,j,i}}\ell\left(w,(x,y)\right)$$

for all $k=1,...,q$, $j=1,...,c$, and $i=1,...,d$.

University of Durham Copyright

SECTION B

Q3 (a) Consider three individual price indices, $\boldsymbol{P} = (P_1, P_2, P_3)^T$, recorded monthly in the UK from 1990–2009 in the categories P_1="Health", P_2="Recreation & Culture", and P_3 ="Hotels, Cafés and Restaurants". These three indices are to be combined into a single index with name W ("Wellbeing"), through an appropriate linear combination

$$W = a_1 P_1 + a_2 P_2 + a_3 P_3.$$

The unit vector $\boldsymbol{a} = (a_1, a_2, a_3)^T$ is to be chosen such that the projection $\boldsymbol{a}^T \boldsymbol{P}$ has maximal variance among all linear combinations of P_1, P_2, and P_3. A principal component analysis of $\boldsymbol{P}$ is carried out, with results summarized in usual notation as

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.610 & -0.215 & -0.763 \\ 0.278 & 0.959 & -0.048 \\ 0.742 & -0.183 & 0.645 \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} 578.013 & 0 & 0 \\ 0 & 14.775 & 0 \\ 0 & 0 & 1.454 \end{pmatrix}.$$

(i) Give the vector $\boldsymbol{a}$.

(ii) What proportion of the total variance of $\boldsymbol{P}$ is explained by W?

(iii) Compute the variance of "Health", $\mathrm{Var}(P_1)$.

(b) Write down the conditions that should be satisfied so that the following function is a natural cubic spline on the interval $[0, 2]$ and determine the values of the coefficients a, b, c, d, and e under the conditions.

$$s(x) = \begin{cases} 1 + x - ax^2 + bx^3, & x \in [0, 1], \\ c + d(x-1) + e(x-2)^2 + (x-2)^3, & x \in [1, 2]. \end{cases}$$

(c) Suppose we want to fit a generalized additive model for a regression problem with two predictors X_1 and X_2. Suppose that we are using a cubic spline with three knots for each predictor, which means our model can be expressed as a linear regression model after the right basis expansion (for example using truncated power basis functions). Suppose that we fit our model by the following three steps:

(i) First fit our cubic spline model for the response y against X_1, obtaining the fit $\hat{f}_1(x)$ and the corresponding residuals $r_i = y_i - \hat{f}_1(X_{i,1})$ where i refers to the i^{th} observed data record.

(ii) Then, fit a cubic spline model for r against X_2 to obtain $\hat{f}_2(x)$.

(iii) Finally construct fitted value $\hat{y}_i = \hat{f}_1(X_{i,1}) + \hat{f}_2(X_{i,1})$.

Will we get the same fitted values as we would if we fit the additive model for y against X_1 and X_2 jointly? Justify your answer.

(d) The Generalised Additive Models, Random Forests and Boosting Trees are all constructed from a number of individual models. Please discuss their differences in terms of tackling regression problems.

University of Durham Copyright

Q4 Consider the binary classification learning problem: Let the set of targets be $\mathcal{Y} = \{-1, +1\}$, let the set of inputs be $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq B\right\}$ for some scalar $B > 0$, let the prediction rule be $h_w(x) = x^\top w$, and let the loss function ℓ be

$$\ell(w, z = (x, y)) = \log\left(1 + \exp\left(-yx^\top w\right)\right),$$

for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $w \in \mathcal{W}$ where $\mathcal{W} = \left\{w \in \mathbb{R}^d : \|w\|_2 \leq B\right\}$.

(a) Show that the resulting learning problem is convex-Lipschitz-bounded. Specify the parameter of Lipschitzness.

(b) Show that the above loss $\ell(w, z = (x, y))$ is $B^2/4$ smooth.

Hint: You may use the mean value theorem which states that (under pre-assumed conditions), $f(b) - f(a) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)\big|_{x=c} (b - a)$ for $a \leq c \leq b$.

(c) Consider a risk function $R_g(w) = \mathrm{E}_{z\sim g}(\ell(w, z = (x, y)))$ where g denotes the unknown data generating process. Assume there is a set of available examples $\mathcal{D} = \{z_i = (x_i, y_i); i = 1, ..., n\}$. Now assume that $w \in \mathbb{R}^d$. To learn w, we aim to compute $w^* \in \mathbb{R}^d$ such that

$$w^* = \arg\min_w (f(w))$$

where $f(w) = R_g(w) + \frac{\lambda}{2}\|w\|_2^2$.

(i) Show that the stochastic gradient descent algorithm with batch size one and with learning rate

$$\eta_t = \frac{1}{\lambda t}$$

at iteration $t \in \mathbb{N}_+$ which is used to address the learning problem under consideration has a recursion that can be written in the form

$$w^{(t+1)} = -\frac{1}{\lambda t}\sum_{j=1}^{t} v_j$$

where $\{v_j\}$ is the gradient of the loss function at certain values of w and example. Show your working.

(ii) Compute the exact formula of v_j as a function of λ, t, $\mathcal{D}$, and $\left\{w^{(t)}\right\}$. Show your working.

University of Durham Copyright