Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	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. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2025

SECTION A

Q1 We have the following observations of categorical variables genus (x_1), habitat (x_2) and number of legs (x_3) alongside the response variable life expectancy in years (y).

Animal	x_1	x_2	x_3	y
rat	mammalia	urban	4	2
cat	mammalia	urban	4	15
salmon	actinopterygii	river	0	5
frog	amphibia	river	4	8
eagle	aves	mountain	2	25

(a) Calculate the distance matrix for these observations based on the Hamming distance (treat the number of legs as a category here rather than a number).

(b) Recode the variables using one-hot encoding and calculate the distance matrix based on the Manhattan distance (treat the number of legs as a number in this case).

(c) Due to the number of categories and variables, we decide to use the Hamming distance version. We have two new animals to predict the life expectancy for:

Animal	x_1	x_2	x_3
dog	mammalia	urban	4
otter	mammalia	river	4

Use a k-nearest neighbour algorithm with $k = 2$ to predict the life expectancy for these two animals.

Q2 Consider the regression problem, with a predictive rule $h_w : \mathbb{R}^d \to (0,1)$ which receives inputs $x = (x_1, ..., x_d)^\top \in \mathbb{R}^d$ and returns values in $(0,1)$. Let $h_w(x)$ be modeled as a feedforward neural network (FNN) with equation

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

We consider activation functions

$$\sigma_1(\xi) = \begin{cases} \exp(\xi) - 1 & , \xi \leq 0 \\ 1 & , \xi > 0 \end{cases}$$

and

$$\sigma_2(\xi) = \exp\left(-\xi^2\right)$$

for $\xi \in \mathbb{R}$. The parameters $c, d \in \mathbb{N}_+$ are 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

$$\ell(w, z = (x,y)) = -\log(1 + h_w(x) - y) + \log(1 - h_w(x) + y)$$

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

(a) Describe the algorithm necessary to perform the forward pass of the back-propagation procedure to compute the activations (which may be denoted by $\{\alpha_{t,i}\}$) and outputs (which may be denoted by $\{o_{t,i}\}$) at each layer t.

(b) Describe the algorithm necessary to perform the backward pass of the back-propagation procedure in order to compute the gradient

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

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

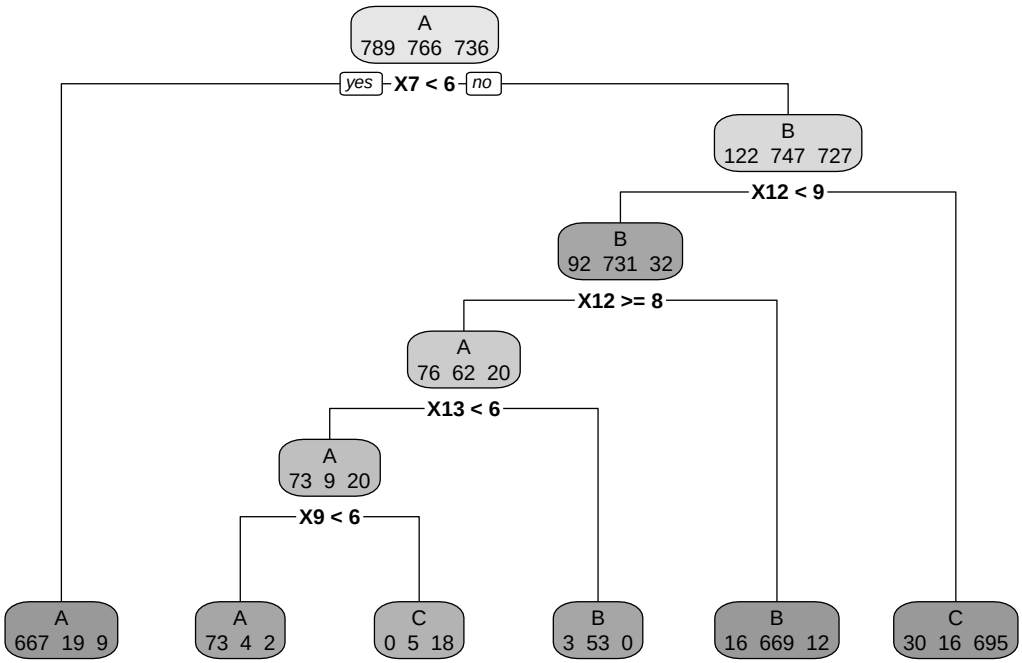
$$\frac{\partial}{\partial w_{1,j,i}}\ell(w,(x,y)), \text{ and } \frac{\partial}{\partial w_{2,1,j}}\ell(w,(x,y))$$

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

SECTION B

Q3 A researcher is attempting to set up a classification system for images of handwritten letters. As a first attempt, they are trying out decision tree models to see if they can predict the letters "A", "B" and "C" from 16 measurements of the letters $(x_1, ..., x_{16})$.

(a) The following decision tree has been fitted to the data.



(i) In which regions of input space would the tree classify as a "C"?

(ii) What is the accuracy of this classifier with respect to all classes?

(b) In constructing a decision tree for a classification problem, two standard impurity measures are used: entropy and Gini impurity. Show that these both are special cases of Tsallis entropy for partition t:

$$I_q(t) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{c} p(i|t)^q\right),$$

where $q > 1$ is a parameter and c is the number of classes

(c) As an alternative, a random forest classifier is proposed.

(i) Write out the algorithm for fitting a random forest model.

(ii) For these data, the accuracy for the random forest model is 0.98 (to 2 d.p.). Which of the two models would you prefer to use and why?

Q4 (a) Consider a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$. Assume that the loss function $\ell(w, z)$ is convex, β-smooth, and nonnegative with $z \in \mathcal{Z}$ and $w \in \mathcal{H}$ where $\|w\| \leq B$. Let g denote the real data generating process. If we run the Stochastic Gradient Descent (SGD) algorithm with constant learning rate η, for total number of iterations T, and with purpose to minimize the risk function $R_g(w)$, then we have that for every $w^* \in \mathcal{H}$

$$\mathrm{E}\left(R_g\left(w_{\mathrm{SGD}}\right)\right) \leq \frac{1}{1-\eta\beta}\left(R_g\left(w^*\right) + \frac{\|w^*\|^2}{2\eta T}\right)$$

where w_{SGD} is the output of the SGD. In addition assume that $\ell(0, z) \leq 1$ for all z. Show that by running online SGD with $\eta = \frac{1}{\beta(1+3/\epsilon)}$ for $T \geq 12B^2\beta/\epsilon$ iterations, where $\epsilon > 0$, we obtain agnostic Probably Approximately Correct (PAC)-like guarantees.

(b) Consider the regression problem with inputs $x \in \mathcal{X} \equiv \mathbb{R}^2$, target $y \in \mathcal{Y} \equiv \mathbb{R}$, and prediction rule $h_w : \mathbb{R}^2 \to \mathbb{R}$ with $h_w(x) = w^\top x = w_1 x_1 + w_2 x_2$, $w = (w_1, w_2)^\top$. Consider a loss function $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ with

$$\ell(w, z = (x, y)) = \left\|y - w^\top x\right\|_2^2 + \alpha \|w_2\|_1 + (1-\alpha)\|w_2\|_2^2 \tag{1}$$

for some given value $\alpha \in (0, 1)$. Assume there is available a training data set $S_m = \{z_i = (x_i, y_i)\,;\, i = 1, ..., m\}$ of size m. Let g denote the real data generating process. Write down the algorithm of the Stochastic Gradient Descent (SGD) with constant learning rate $\eta > 0$, batch sample size equal to 1, and termination criterion $t > T$ for some $T > 0$ (t denotes the t-th iteration of SGD), that aims to compute w^*, where

$$w^* = \arg\min_w \left(\mathrm{E}_{z\sim g}\left(\ell(w, z = (x, y))\right)\right) \tag{2}$$

The formulas in your algorithm should be specific to the loss function in (1).

(c) Let X be an instance set and let ψ be a feature mapping of $\mathcal{X}$ into some Hilbert feature space $\mathcal{V}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function that implements inner products in the feature space $\mathcal{V}$.

Consider the binary classification algorithm that predicts the label of an unseen instance according to the class with the closest average. Formally, given a training sequence $\mathcal{S} = \{(x_1, y_1), ..., (x_m, y_m)\}$, for every $y \in \mathcal{Y} = \{-1, +1\}$ we define

$$c_y = \frac{1}{m_y} \sum_{i:y_i = y} \psi(x_i)$$

where $m_y = |\{i : y_i = y\}|$. We assume that m_{+1} and m_{-1} are nonzero. Then, the algorithm outputs the value of the following decision rule:

$$h(x) = \begin{cases} 1 & , \|\psi(x) - c_{+1}\|_2 \leq \|\psi(x) - c_{-1}\|_2 \\ -1 & , \text{otherwise.} \end{cases}$$

(i) Let $w = c_{+1} - c_{-1}$ and let $b = \frac{1}{2}\left(\|c_{-1}\|_2^2 - \|c_{+1}\|_2^2\right)$. Show that

$$h(x) = \mathrm{sign}\left(\langle w, \psi(x)\rangle + b\right)$$

(ii) Express $h(x)$ in terms of the kernel function, and without accessing individual entries of $\psi(x)$ or w.