17

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION SUMMER SEMESTER, 2022-2023

DURATION: 1 HOUR 30 MINUTES FULL MARKS: 75

CSE 4621: Machine Learning

Programmable calculators are not allowed. Do not write anything on the question paper.

Answer **all 3 (three)** questions. Figures in the right margin indicate full marks of questions with corresponding COs and POs in parentheses.

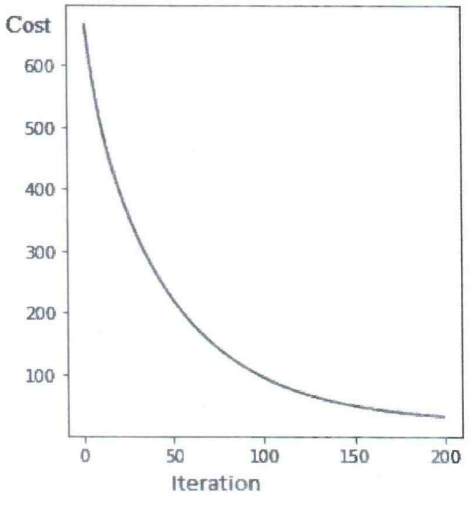
---

1. a) Briefly explain the basic workflow of a typical supervised learning algorithm. 6 (CO1) (PO1)

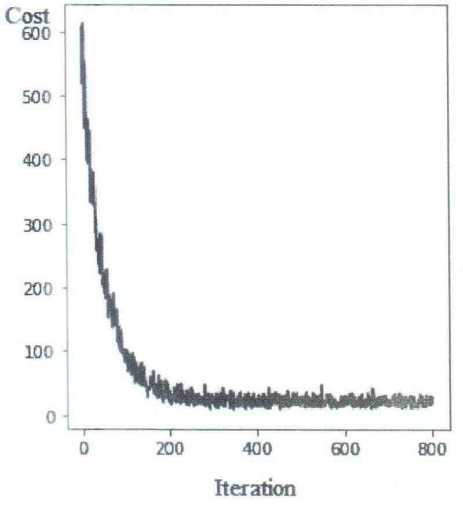
   b) With necessary figures identify why feature scaling is important before applying the gradient descent technique. Write the equation to rescale the range of a feature to a new scale $[-2, 3]$. 6 + 2 (CO1) (PO1)

   c) Mathematically show that the binary cross-entropy function with L2 regularization term is convex. 5 (CO1) (PO1)

   d) Consider the two different cost functions $J(\theta)$ plotted over many iterations as shown in Figure 1. Which graph corresponds to the use of Batch gradient descent (BGD) and which one corresponds to use of Mini-Batch gradient descent (mBGD). Justify your choice. 6 (CO2) (PO2)



(a) Cost values for case 1.



(b) Cost values for case 2.

Figure 1: Cost function curves for Question 1.d

2. a) For a binary classification model, we predict for the positive class (i.e., $y = 1$) when $P(y = 1|x) \geq 0.5$, where a logistic function is used to convert to a probabilistic value. Mathematically show that this model is linear. 7 (CO1) (PO1)

   b) How can you extend a binary classifier into a multi-class classifier considering a one-versus-rest strategy? What are the limitations of this design? Draw figures as necessary. 6 + 2 (CO1) (PO1)

c) Consider a rare scenario where the d-dimensional feature vectors contains binary features, i.e. $x \in \{0,1\}^d$, where x_1 is mostly 0, and happens to be 1 in the training set with only positive class ($y = 1$) samples. Roughly estimate the value of $\hat{\theta}_1$. Justify your answer — 5 (CO1) (PO1)

d) Is it possible to get a closed form solution for the parameter $\hat{\theta}$ when the cost function $J(\theta)$ is the log-loss function for binary classification? Explain why. — 5 (CO1) (PO1)

3. a) Your teammates are training a neural network for a classification task using a feed-forward architecture with one hidden layer. They decide to tune the number of neurons in the hidden layer using cross-validation, and compare train, validation and test error. They run cross-validation for numbers of neurons in the hidden layer of [50, 100, 200]. Upon examining the test error curve, you notice that the test error increases when the number of neurons decreases beyond 50. After evaluating the models, they choose 200 neurons in the hidden layer because it resulted in the lowest test error. — 5 (CO1) (PO1)

This observation suggests that there may be an optimal number of neurons that was not explored. You suggest extending the search to include additional values for the number of neurons, such as [250, 300, 350], to ensure a more comprehensive search for the best-performing neural network architecture.

Do you believe you gave the right recommendation to your team considering the behavior of the test error? Justify your answer.

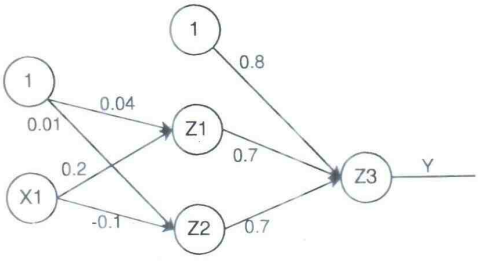


Figure 2: Neural Network for Question 3.b

b) Consider the 2-layer neural network with its weights as shown in Figure 2. In the hidden and output layers, sigmoid activation functions are used. For a sample input $x = 0.7$, the actual output is $y = 1$. Answer the followings: — 2 + 6 + 2 (CO1) (PO1)

i. Compute the predicted output $\hat{y}$.

ii. Find the updated weight matrix $W^{[1]}$ after one epoch.

iii. Also, find the updated bias $b^{[1]}$.

c) Consider a very simple neural network with N layers that takes in a scalar input $x = a_0$, and for each layer $i \in \{1, 2, .., N\}$, we have — 3 + 2 + 5 (CO1) (PO1)

$$z_i = w_i * a_{i-1} + b_i \quad (1)$$

$$a_i = \sigma(z_i) \quad (2)$$

Note that w_i, b_i, z_i, a_i are all scalars. Answer the followings:

i. Using the chain rule of calculus, give the mathematical expression for $\frac{\partial a_i}{\partial a_{i-1}}$. Your expression should only contain $\sigma'(z_i)$ and w_i.

ii. Since the derivative of the sigmoid function $\sigma'(.)$ is at most $\frac{1}{4}$. Give an upper bound value for $\frac{\partial a_N}{\partial w_1}$

iii. From Question 3.c)ii, how can you prove that for a neural network with too many layers, this deep network will lead to slow or unstable learning?