13

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION **SUMMER SEMESTER, 2021-2022**

DURATION: 3 HOURS **FULL MARKS: 150**

CSE 4621: Machine Learning

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

Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of questions whereas corresponding CO and PO are written within parentheses. Symbols have their usual meaning.

---

1. a) The standard form of L2-regularized loss function for linear regression is: 3+3 (CO2) (PO2)

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2 + \frac{\lambda}{m}\theta^T\theta$$

i. Suppose you have accidentally defined: $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2 + \frac{\lambda}{m}\sum_{i=1}^{m}\left\|y^i\right\|^2$

What kind of regulization effect will you have?

ii. Suppose we use the correct expression but accidentally choose $\lambda<0$. Will you either have overfitting or underfitting? Justify your answer.

b) Suppose you are training a robot in a lumber yard, and the robot must learn to discriminate Oak wood from Pine wood. For your robot construct a decision tree with the following data in Table 1. Show all calculations for split generation. 14 (CO3) (PO3)

Table 1: Dataset for Question 1.b)

Density	Grain	Hardness	Class
Heavy	Small	Hard	Oak
Heavy	Large	Hard	Oak
Heavy	Small	Hard	Oak
Light	Large	Soft	Oak
Light	Large	Hard	Pine
Heavy	Small	Soft	Pine
Heavy	Large	Soft	Pine
Heavy	Small	Soft	Pine

c) Differentiate between the entropy measures for classification and regression trees. 5 (CO2) (PO2)

2. a) Suppose you have used linear activation functions in all layers of a Perceptron neural network. What advantages and disadvantages will you face? Explain with the help of a three-layer neural network with necessary illustrations. 5 (CO2) (PO2)

b) Suppose Figure 1 represents the forward calculation involved in a feed-forward neural network with a logistic function, σ. 8 (CO1) (PO1)

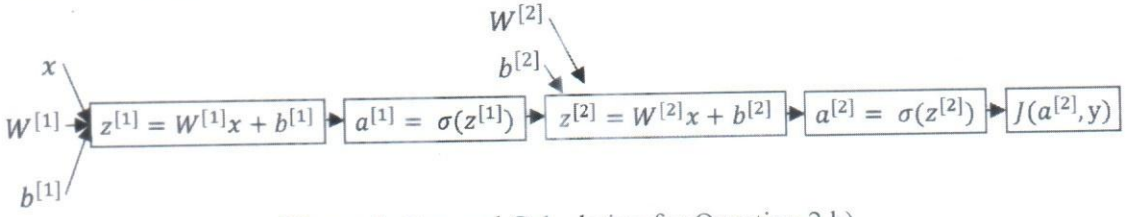

Figure 1: Forward Calculation for Question 2.b)

Here, $J(\hat{y}, y)$ is the binary cross-entropy cost function. Derive the mathematical expression of the derivative: $\frac{dJ}{dz^{[1]}}$. Give the weight update equation for $W^{[1]}$.

c) Consider the hypothetical convolution neural network defined by the layers in the left column in Table 2. Calculate the shape of the output volume and the number of parameters (weights and biases) at each layer. You can write the activation shapes in the format (H, W, C), where H, W, C are the *height*, *width* and *channel* dimensions, respectively. Unless specified, assume *padding=1* and *stride=1* where appropriate.

The notation follows the convention:

- CONV*f*-N denotes a convolutional layer with *N* filters, each them of size $f \times f$.
- POOL-n denotes a n×n max-pooling layer with stride of n and 0 padding.
- FLATTEN flattens its input.
- FC-N denotes a fully-connected layer with N neurons.

Table 2: List of Layers for Question 2.c)

Layer	Activation Volume Dimensions	Number of Parameters
Input	**128*128*6**	
CONV5-8		
ReLU		
CONV3-16		
POOL-2		
CONV2-64		
CONV1-32 with padding 0		
POOL-2		
POOL-2		
FLATTEN		
FC-27		

3. a) Compare between Generative and Discriminative models with examples. 6 (CO2) (PO2)

b) You decide to make a text classifier. In the beginning, you attempt to classify documents as either sport or politics. You decide to represent each document as a (row) vector of attributes, x, describing the presence or absence of words. 14 (CO1) (PO1)

$$x = [\text{goal}, \text{football}, \text{golf}, \text{defence}, \text{offence}, \text{wicket}, \text{office}, \text{strategy}]^T$$

Training data from sport documents and politics documents is represented below using a matrix in which each row represents a sample document as a (row) vector of the 8 attributes.

$$X_{politics} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \quad X_{sport} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Using a Naive Bayes classifier, classify the document $x = [1,0,0,1,1,1,1,0]^T$. Calculate each individual class-conditional probability.

What are the assumptions made in the Naive Bayes classifier? 5 (CO1) (PO1)

a) What is the curse of dimensionality? How can you deal with this problem with a linear projection method? 1+9 (CO1) (PO1)

b) Principal Component Analysis (PCA) transforms a set of correlated variables into a new set of uncorrelated variables. How is the transformation carried out with principal components? 10 (CO1) (PO1)

c) Suppose a 3×3 covariance matrix **A** has trace value of 4, and two of its Eigen values are 1 and 0. Interpret on the distribution pattern of the original samples in the 3D feature space from which that covariance matrix is computed. 5 (CO1) (PO1)

5. a) What is Cluster Analysis? Briefly describe the major considerations for cluster analysis. 2+6 (CO1) (PO1)

b) Consider the data set as given in Table 3 consisting of the scores of two variables on each of six individuals. 12 (CO1) (PO1)

Table 3: Dataset for Question 5.b)

Sample	x_1	x_2
1	**1.0**	**1.0**
2	**1.5**	**2.0**
3	**3.0**	**4.0**
4	**5.0**	**7.0**
5	**3.5**	**5.0**
6	**4.5**	**5.0**

Apply the *k*-means clustering algorithm with the value $k = 2$. Provide all required calculations up to two cluster-center updates.

c) Which clustering algorithm is more robust in presence of outliers: *k*-means or *k*-medoids? Explain why. 5 (CO2) (PO2)

6. a) "Irrespective of the dimensionality of the data space, a data set consisting of just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane" – Do you agree or disagree with this statement. Justify your answer. 8 (CO2) (PO2)

b) Build the mathematical expression of the objective function [along with the constraint(s)] which needs to be minimized in order to find the decision boundary with maximum margin. 7 (CO1) (PO1)

c) If the training samples of a two-class problem cannot be linearly classified in the original feature space, how does Support Vector Machine (SVM) try to classify them? Show with a help of a kernel. 10 (CO1) (PO1)