43

Ph.D. / M.Sc. Engg. CSE 15 May 2024

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION SUMMER SEMESTER, 2022-2023

DURATION: 3 HOURS FULL MARKS: 150

CSE 6229: 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.

1. Consider a binary classification problem where you have a dataset with two features, x_1, x_2, and binary classes as shown in Table 1. Suppose that you want to solve the classification problem using the Support Vector Machine (SVM) algorithm.

Table 1: Dataset for Question 1.a.

Sample	x_1	x_2	Class
1	2	3	0
2	3	4	0
3	4	5	1
4	5	6	1
5	6	7	0

a) What is the linearity of the dataset in this binary classification problem? How can it be determined? 7

b) Explain the concept of the kernel trick in Support Vector Machine (SVM) classification. If the classification problem is non-linear, outline the utilization of the kernel trick. Furthermore, write an appropriate kernel function for solving non-linear classification problem, and provide step-by-step instructions along with necessary diagrams to illustrate the solution approach. 10

c) Explain why the SVM algorithm aims to maximize the margin with the help of support vectors and how it contributes to the model's robustness. 8

2. Suppose you want to construct and train a neural network to solve the classification problem based on the dataset in Table 1. The task involves designing a simple feed-forward neural network using suitable activation function(s) and appropriate loss function.

a) Draw a schematic diagram of the neural network architecture suitable for solving the problem. Clearly label the input layer, hidden layer(s), and output layer. Indicate the number of neurons in each layer. 6

b) Write the importance of weight initialization in neural networks. 2

c) Show the required calculations involved in forward propagation for a given input sample, including the calculation of activations using your chosen activation functions. 7

d) Explain the purpose of activation functions in neural networks and their importance in introducing non-linearity and the role of loss function in quantifying the model's performance. 5

e) Explain the concept of backward propagation and the role of gradient descent in updating the weights of the neural network during training. 5

3. You have been tasked with constructing and benchmarking a Bangla word-level sign language dataset for classification using classical machine learning models. The dataset consists of videos captured at different frame rates (15, 24, 30 fps) with varying signer speeds, and key points are extracted using the Mediapipe library. After building the model, evaluating its performance is crucial for assessing its effectiveness.

   a) Discuss in detail the feature engineering tasks (i.g. Normalization, handling missing values, dealing with temporal variations, segmentation, feature extraction, etc.) required to prepare the dataset for classification. 15

   b) Elaborate on the evaluation metrics (e.g. classification metrics, cross-validation, confusion matrix, Receiver Operating Characteristic (ROC)/Area Under the Curve (AUC) curves, etc.) and techniques you will be using to measure the performance of the model. 10

4. a) Given the scenario in Question 3, how the clustering techniques can be used for feature representation? How the clustering technique can be used for gesture similarity analysis? Explain your answer. 9

   b) Perform the iterations of agglomerative clustering (single linkage) using the provided distance matrix as given in Table 2 and draw the corresponding Dendrogram. 16

Table 2: Dataset for Question 4.b.

	A	B	C	D
A	-	2	5	1
B	-	-	3	4
C	-	-	-	6
D	-	-	-	-

Show each step of the clustering process, including the merging of the two closest classes and the corresponding updated distance matrix. Provide clear explanations for each step.

5. a) Consider the dataset in Table 3 consisting of information about the movie teaser views (in millions) and likes (in thousands) of five movies along with their respective popularity categories (High, Medium, Low).

Table 3: Dataset for Question 5.a.

Movie ID	Teaser Views (millions)	Likes (thousands)	Popularity
1	23	1.3	High
2	15	0.3	Low
3	34	2.0	High
4	13	1.1	Medium
5	10	0.5	Low

Answer the following:

   i. For a new movie with teaser views of 20 million and likes of 1.2 thousand, using the K-nearest neighbor (K-NN) algorithm with $k = 3$, find the popularity category of the new movie. Show the step-by-step process. 12

   ii. Discuss how changing the value of k might impact the prediction and the importance of feature scaling in this k-NN application. 5

b) Differentiate between training accuracy and testing accuracy. Why is it important to evaluate a model on a separate test set? Discuss scenarios where a high training accuracy may not necessarily translate to good generalization performance on unseen data. 8

6. Explain the considerations when selecting machine learning algorithms, identifying dataset characteristics, problem complexity, interpretability, and computational resources. Utilize a real-life machine learning problem scenario, detailing dataset, problem statement, and objectives, while elucidating stages of the machine learning pipeline (data preprocessing, feature selection, model training, evaluation, and deployment). Discuss decisions made at each stage regarding algorithm selection, parameter tuning, and model evaluation, and analyze the impact of algorithm choice on overall performance and solution success. Provide key takeaways regarding the importance of algorithm selection for effectively solving real-world problems. 25