University of Asia Pacific
Department of Computer Science and Engineering
Final Examination, Fall 24
Program: B.Sc. in CSE
4th year 1st Semester

Course Title: Topics of Current Interest | Course Code: CSE 427
Time: 3:00 hours | Credit Hour: 3.00 | Full Marks: 50

There are Five Questions. Answer all of them. Part marks are shown in the margins.

QUESTION 1 [10 MARKS]

a. **Differentiate** between classification and clustering on the following aspects: definition, core concept, input, output, and relevant example. [5]

b. **Apply** the principles of SVM to handle linearly inseparable data, including the formulation and key concepts involved. [5]

QUESTION 2 [10 MARKS]

Table 1. Purchase behavior dataset.

Age Group	Income	Education Level	Purchased (Label)
Young	Low	High School	No
Young	High	College	Yes
Middle-aged	Medium	College	Yes
Old	Low	High School	No
Middle-aged	High	Masters	Yes
Old	Medium	College	No
Young	High	Masters	Yes
Old	Low	High School	No
Middle-aged	High	College	Yes
Young	Medium	High School	No

a. Given the purchase behavior dataset in **Table 1**, **apply** the Naïve Bayes classifier to predict whether a person will purchase a product, given the following data point: *Age Group = Middle-aged, Income = High, Education Level = Masters*. [5]

b. **Identify** the first feature to split on when constructing a decision tree using the purchase behavior dataset in **Table 1**. [5]

QUESTION 3 [10 MARKS]

Apply the perceptron learning algorithm over 2 epochs to determine the decision boundary for a dataset containing 4 rows representing a 2-input OR gate. The structure of the perceptron is given in **Figure 1**. Use the following parameters and hyperparameters: $w_1 = 0.3$, $w_2 = 0.4$, $\alpha = 0.01$, $b = -0.2$. 10

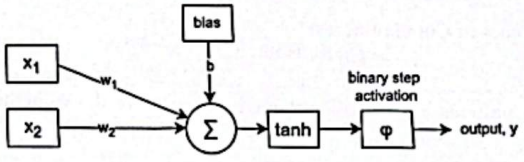


Figure 1. Perceptron Structure.

QUESTION 4 [10 MARKS]

Construct the output of each layer in the CNN illustrated in **Figure 2**, assuming 10

QUESTION 3 [10 MARKS]

Apply the perceptron learning algorithm over 2 epochs to determine the decision boundary for a dataset containing 4 rows representing a 2-input OR gate. The structure of the perceptron is given in Figure 1. Use the following parameters and hyperparameters: $w_1 = 0.3$, $w_2 = 0.4$, $\alpha = 0.01$, $b = -0.2$. 10

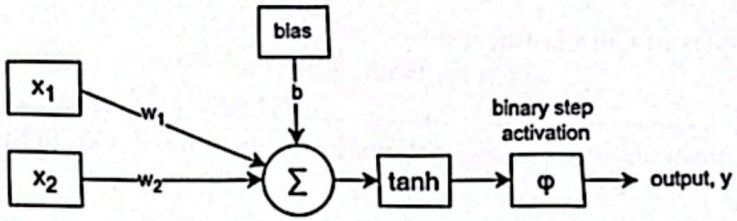


Figure 1. Perceptron Structure.

QUESTION 4 [10 MARKS]

Construct the output of each layer in the CNN illustrated in Figure 2, assuming white cells represent 0 and black cells represent 1. 10

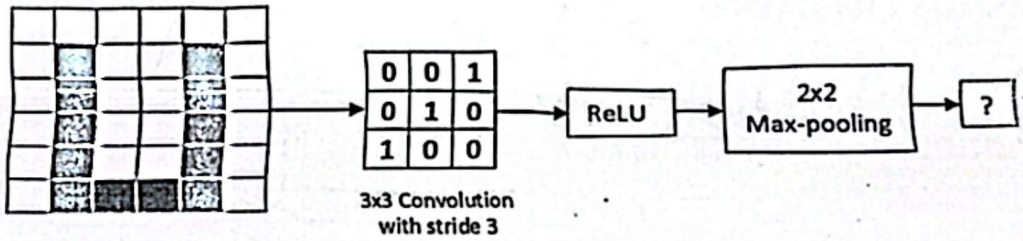


Figure 2. CNN Architecture.

QUESTION 5 [10 MARKS]

Analyze the Table 2 to calculate the Positive Predictive Value (PPV) and True Positive Rate (TPR). Also, show the confusion matrix. 10

Table 2. Binary Classification Results.

Sl.	1	2	3	4	5	6	7	8	9	10	11	12
y	0	1	0	1	0	1	0	1	1	0	1	1
ŷ	0	0	1	1	0	ID%2	1	0	1	0	ID%2	0

OR,

Analyze the Table 2 to calculate the True Negative Rate (TNR) and F1-score. Also, show the confusion matrix. 10

University of Asia Pacific

Department of Computer Science and Engineering

Program: B.Sc. in CSE

Final Examination | **Spring - 2024** | **4rd Year 1st Semester**

Course Code: CSE 427 | **Course Title: Topics of Current Interest** | **Credit: 3.00**

Time: 3.00 Hour. | **Full Mark: 50**

There are Five Questions. Answer all of them. Part marks are shown in the margins.

1. "K-means clustering is a supervised algorithm" on the other hand, "the K-NN algorithm is an unsupervised algorithm." – Do you agree with these statements? Explain your answer. [5] [CO2]

2. Two models, Logistic Regression and Decision Tree, are used to predict whether a customer will default on a loan. In this context, the defaulter is considered the positive class, while the non-defaulter is considered the negative class. The confusion matrices for the predictions are as follows: [10] [CO4]

Confusion Matrix for Logistic Regression

Predicted Value \ Actual Value	Defaulter	Not Defaulter
Defaulter	40	5
Not Defaulter	10	45

Confusion Matrix for Decision Tree

Predicted Value \ Actual Value	Defaulter	Not Defaulter
Defaulter	38	7
Not Defaulter	8	47

CamScanner

Analyze the accuracy, precision, recall, and F1-score for both models. Based on these metrics, determine which model performs better and justify your choice.

OR

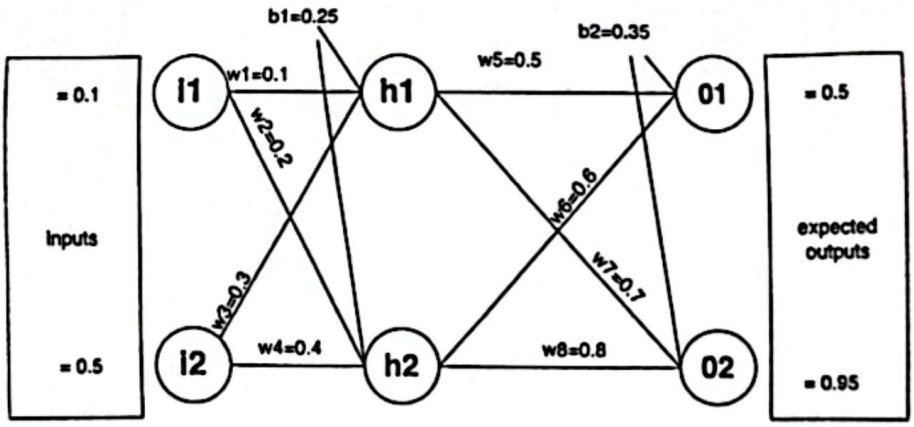
A company wants to build a recommendation system for a new e-commerce platform. The system must: [10] [CO4]

1. Handle categorical and numerical data (e.g., user age, gender, purchase history, and product categories).
2. Be interpretable for understanding the impact of different features on recommendations.
3. Provide fast predictions for real-time recommendations.

Given the requirements, which algorithm would you choose: SVM, Decision Tree, Naive Bayes, or KNN? Explain your reasoning by addressing the suitability of each algorithm for the task and the requirements.

3. a. Discuss the drawbacks of using the ReLU activation function in neural networks, providing mathematical reasoning. [4] [CO1]

b. Assume that the neurons have a sigmoid activation function. Perform a forward pass on the given network and also calculate error. [10] [CO3]



4 a. Describe a scenario where K-fold cross-validation might not be suitable and suggest an alternative method. [5] [CO2]

CamScanner

b. Table 1 provides data collected from a group of students about their study habits, academic performance, and participation in class activities. The attributes include whether a student studies regularly (Yes/No), completes homework (Yes/No), participates in class (Yes/No), and their combined CGPA (High/Medium/Low). Based on these attributes, an output label, Pass Exam?, indicates whether the student successfully passes their current course (Yes) or not (No). [10] [CO2]

Table 1

Studies Regularly	Completes Homework	Participates in Class	CGPA	Pass Exam?
Yes	Yes	Yes	HIGH	Yes
Yes	No	No	LOW	No
No	No	Yes	MEDIUM	No
Yes	Yes	No	HIGH	Yes
No	Yes	Yes	MEDIUM	Yes
No	No	No	LOW	No
Yes	No	Yes	MEDIUM	Yes
No	Yes	No	LOW	No

Using the Naïve Bayes algorithm, determine whether a student who studies regularly, participates in class but does not complete homework, and has low CGPA will pass the exam.

OR

b. A security system is analyzing factors that influence whether an email is classified as spam. The dataset includes attributes such as email length (short/long), whether the email contains links (yes/no), and whether it contains attachments (yes/no). Based on these attributes, the label Spam? indicates whether the email is spam (Yes) or not (No). The dataset is in Table 2: [10] [CO2]

Table 2

Email Length	Contains Links	Contains Attachments	Spam?
Short	Yes	No	Yes
Long	No	Yes	No
Short	Yes	Yes	Yes
Long	Yes	No	No
Short	No	No	No

Construct the decision tree step by step, showing all calculations for Information Gain (using Entropy). Determine whether the email, where the length is long, does not contain links, and does not include attachments, is classified as spam.

CamScanner

5. Evaluate the performance of a predictive model using the outcomes mentioned in Table 3. Calculate the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score. [6] [CO4]

Table 3

$\hat{y}$	32	38	33	52	47	50	58	67	68	72
y	30	40	35	50	45	55	60	65	70	75

University of Asia Pacific
Department of Computer Science and Engineering
Program: B.Sc. in CSE

Final Examination | **Fall-2023** | **4th year 1st Semester**

Course Code: CSE 427 | Course Title: Machine Learning | Credit: 3

Time: 3.00 Hour. | Full Mark: 50

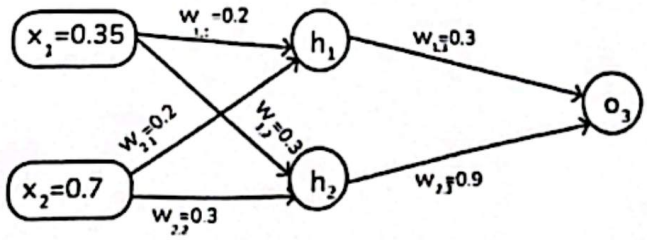
There are Five Questions. Answer all of them. Part marks are shown in the margins.

1. a. "An activation function decides the output of each layer with respect to the input in a neural network" – justify using proper mathematical equations and diagrams. [4] [CO2]

   b. Suppose you have a dataset for a classification problem with three features A, B, and C with a range of [0,1], [500,10000], and [-50,50] respectively which makes feature B greatly affect the drawing of the decision boundary. Assess the importance of feature scaling to resolve this issue through proper illustration of the problem. [4] [CO2]

   c. Explain how the increase in amount of data affect the performance of ML and DL models. [2] [CO2]

2. Applying back propagation, update the weights $w_{1,1}$ and $w_{1,3}$ of the back propagation neural network shown in the following figure after one epoch. Sigmoid function is used as an activation function. Learning rate α=0.6, actual output from node o_3 is 0.5, and bias for the hidden b_1=0.55. [10] [CO4]



3. a. Suppose a KNN and an SVM model make prediction of a patient having the risk of a heart attack. The confusion matrix obtained by them are as follows [4] [CO5]

$$confusion_{KNN} = \begin{bmatrix} 35 & 2 \\ 8 & 25 \end{bmatrix}, confusion_{SVM} = \begin{bmatrix} 30 & 7 \\ 8 & 25 \end{bmatrix}$$

Analyze the accuracy, precision, recall, and f1-score of these models and decide which one is more robust.

b. Evaluate the performance of the model using the outcomes mentioned in the following table with the help of MSE, RMSE, and R squared error. [6] [CO5]

$\hat{y}$	28	37	41	36	27	32	45	51	50	48	55	57
y	32	35	42	45	33	35	51	45	48	55	51	49

4. Construct a decision tree from the dataset given in table 1 considering information gain as attribute selection measure. [10] [CO3]

OR

Construct a decision tree from the dataset given in table 1 considering gini index as attribute selection measure. [10] [CO3]

Table 1: Sample dataset

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Rain	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Rain	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Cool	Normal	Weak	Yes
D14	Sunny	Mild	High	Strong	No

5. Develop a one-dimensional dataset using principal component analysis from the following sample dataset. [10] [CO3]

x	4	8	13	7
y	11	4	8	14

University of Asia Pacific

Department of Computer Science and Engineering

Program: B.Sc. in CSE

Final Examination | **Spring-2023** | **4th year 1st Semest**

Course Code: CSE 427 | Course Title: Topic of Current Interest (ML) | Credit: 3

Time: 3.00 Hour. | Full Mark: 50

There are Five Questions. Answer all of them.

1. a Explain the eps and MinPts parameter in the DBSCAN algorithm. Explain MinPts's significance in determining whether a data point is considered a core point or not. [3]

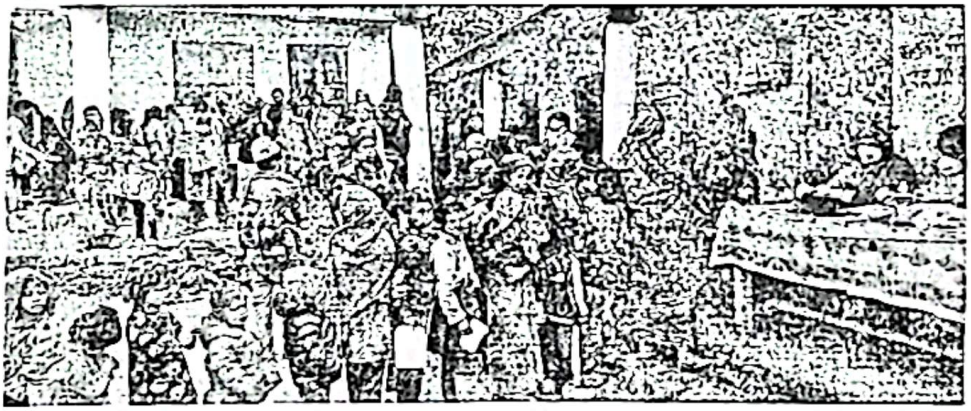
b You are given a dataset of three-dimensional data points with x, y, and z coordinates. Your task is to perform DBSCAN clustering on this dataset and identify the clusters. The dataset is as follows: [7]

Points	X-coordinate	Y-coordinate	Z-coordinate
P1	2	3	4
P2	2.5	3.2	4.1
P3	2.2	3.1	4.2
P4	11.9	12.0	12.6
P5	12	12.5	12.3
P6	11.8	12	12.6

For eps=1.5 and MinPts=3, Calculate the following:

- The number of clusters found.
- List the points in each cluster.
- Identify any noise points if present.

2. a Imagine a scenario in a small town where a rare disease, "Lumina Fever," has been spreading. The local clinic developed a diagnostic test to identify the presence of Lumina Fever in patients. The test was administered to a group of 100 individuals who were either healthy or infected. [10]



Out of those tested:

- ~~Many~~ 35 individuals who were truly infected with Lumina Fever tested positive for the disease. These patients were relieved to receive correct diagnoses.
- ~~Some~~ 15 healthy individuals, however, tested positive incorrectly. They were alarmed by the false-positive results and sought further medical advice.
- Unfortunately, ~~a few~~ 20 individuals who were infected with Lumina Fever tested negative. They were initially relieved by the negative result but later discovered the oversight.
- A significant number of healthy individuals (30) correctly tested negative, providing them with peace of mind.

In this scenario, you have the information to construct a confusion matrix for the Lumina Fever diagnostic test. Thus, construct precision, recall, and f1-score performance metrics. In addition, construct a confusion matrix.

70 64 66.87

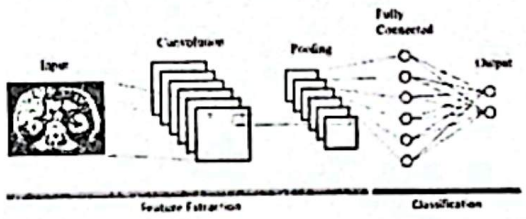
OR

Consider a dataset named "photonics.csv". Assume this is a numerical dataset with 110 attributes and 1 million data points. Now write a code to do the following tasks: [10] [CO5

- Load the dataset.
- Preprocess it.
- Train a machine learning model
- Evaluate the model.

For every task write a justification as comment of your code.

3. In a medical research project, a team of scientists is working on developing an image classifier to detect the presence of a rare disease, "Luminexia," from medical images. Luminexia is known to affect certain tissues in the human body, and early detection is crucial for effective treatment. The research team collected a dataset of medical images representing tissue samples from patients, both healthy individuals and those diagnosed with Luminexia. These images are preprocessed and converted into numerical matrices. To detect disease patterns, the team employs convolutional neural networks (CNNs). One of the key steps in this process is the use of **convolutional layers, max-pooling, and flattening** to create a feature vector for classification. In this context, you are given a 5x5 numerical matrix representing a section of a tissue image [ref (i)], along with a convolutional filter [ref (ii)]. Additionally, max-pooling and flattening operations are applied as part of the feature extraction process. [10] [CO



Your task is to apply the above mentioned (in BOLD) operations to the input matrix and filter and explain the step-by-step process, culminating in a feature vector that can be used for image classification. Understanding these operations is critical for the development of the Luminexia detection model.

Matrix = [[1, -2, 0, 1, 0]. ——— (i)

[3, 1, 2, -2, 1].

[0, 1, 3, 0, -2].

[1, 2, -2, 1, 3].

[0, -1, 1, 2, -1]]

Filter = [[0, 1, 0]. ——— (ii)

[2, 1, 2].

[1, 0, 1]]

4 a. Explain why selecting the appropriate 'k' value is crucial in KNN and how it affects the classification results. [3] [CO

b. You are given a dataset containing information about various fruits. The dataset includes the fruit's weight in grams, its color (encoded as 0 for "Red," 1 for "Yellow," and 2 for "Green"), and whether the fruit is labeled as "Sweet" (encoded as 0 for "No" and 1 for "Yes"). The dataset is as follows: [7] [CO

Weight (g)	Color	Sweet
120	0	1
150	1	1
80	2	0
10	0	1
70	2	1
130	1	1
90	2	1
50	2	0
5	0	1

You are tasked with classifying a new fruit based on its weight and color using the K-Nearest Neighbors (KNN) algorithm with k=3. The new fruit weighs 100 grams and is red color.

i) Calculate the Euclidean distance between the new fruit and each of the fruits in the dataset.

ii) Determine the 3 nearest neighbors to the new fruit based on the calculated distances.

iii) Based on the majority class among the 3 nearest neighbors, predict whether the new fruit is "Sweet" or "Not Sweet."

a. Describe the fundamental distinction between linear and logistic regression techniques, including the nature of the target variable and the type of problem each is suited for. [3]

b. You are provided with a dataset containing information about students' exam performance and whether they were admitted to a university. The dataset includes two features: the score of Exam 1 and the score of Exam 2. The target variable, labeled "Admitted," indicates whether a student was admitted (1 for "Yes" and 0 for "No"). The dataset is as follows: [7]

Exam 1 Score	Exam 2 Score	Admitted
80	90	?
60	70	?
75	80	?
55	65	?
90	85	?
65	75	?
70	70	?
50	60	?
85	95	?
80	70	?

Your task is to build a logistic regression classifier to predict whether a student will be admitted based on their exam scores. Assume that the logistic regression model has been trained and the following coefficients have been learned:

- Intercept (Bias): -6.0
- Coefficient for Exam 1: 0.05
- Coefficient for Exam 2: 0.08
- Threshold: 0.55

i) Calculate the logistic regression model's output for each student in the dataset based on their exam scores.

ii) Based on the model's output, make predictions for each student, indicating whether they are predicted to be "Admitted" or "Not Admitted."