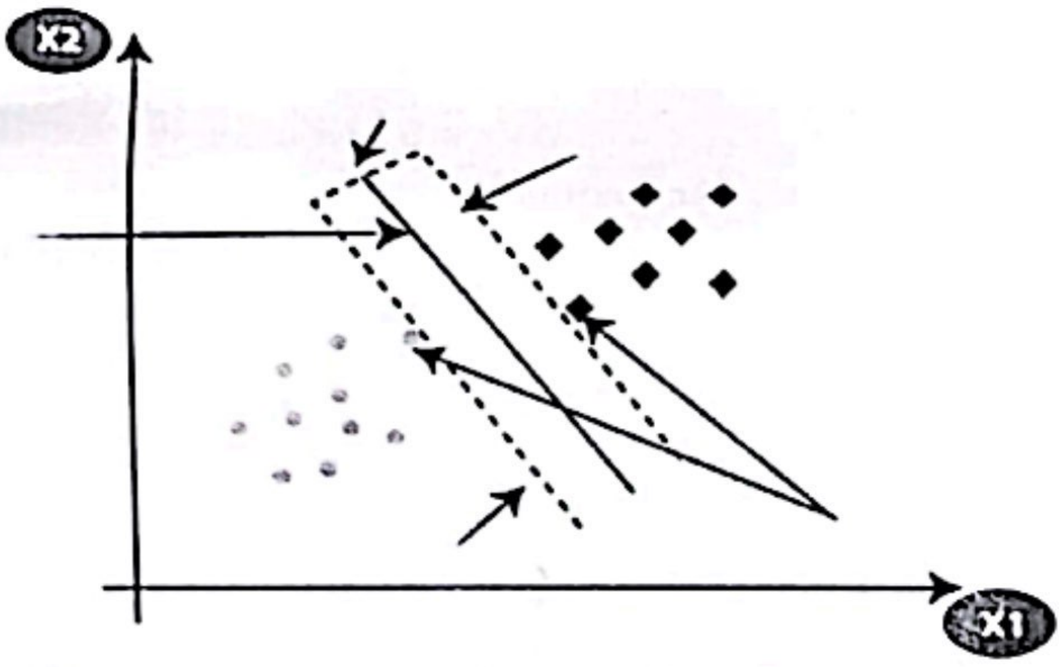
Jahangirnagar University
Department of Computer Science and Engineering
Fourth Year Second Semester B.Sc. (Hons.) Final Examination -2022

Course Title: **Machine Learning** Course No: **CSE-457**
Full Marks: **60** Time: **3 Hrs.**

[Answer each of the following questions. Each question carries equal marks. Figures in the right margin indicate marks.]

1. Answer **all** questions:
   a) Discuss the differences between the training set and the testing set. 2
   b) Write the differences between supervised and unsupervised machine learning techniques. 2
   c) Define hyperplane. Why do we use kernels in SVM? 2
   d) How do you handle missing data in a dataset? 2
   e) What is a class-imbalanced dataset in machine learning? 2
   f) What is the Leave-1-Out Cross-Validation accuracy? 2

2. Answer **Any Three** out of **Four** questions:
   a) SVM is a supervised machine learning algorithm that works on both classification and regression problem statements. Identify the indicated arrow sign using the SVM algorithm from the following diagram. 4



   b) Distinguish and compare between linear and logistic regression with proper examples. 4
   c) With the help of a diagram illustrate the relationship among artificial intelligence, machine learning, data science, and deep learning. 4
   d) Explain Markov Decision Process. 4

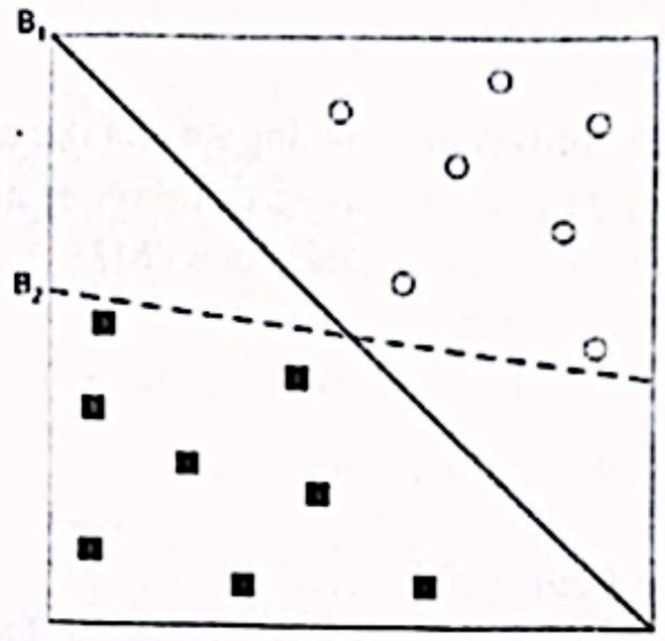
3. Answer **Any Three** out of **Four** questions:

a) Suppose, we use a logistic regression model to predict whether or not 400 different college basketball players get drafted into the NBA. From the predicted and actual test outcome the confusion matrix formed as the following. Calculate Precision, Recall, Accuracy, and F1-measure.

	Actual "Yes"	Actual "No"
Predicted "Yes"	120	40
Predicted "No"	70	170

0.75
0.632
0.725
0.6864

b) Consider the data points in the figure below:



Let us assume that the square boxes represent the positive class whereas the circles represent the negative class. Which line B_1 / B_2 is the maximum-margin hyperplane and why?

c) Explain what are the various metrics used to evaluate the machine learning model performance 4

d) With a suitable numeric example, illustrate the k-means clustering algorithm. 4

4. Answer **Any Three** out of **Four** questions:

a) Given that k = 3, use the k-means algorithm to cluster the following 6 records in Table 1 into 3 clusters. Suppose that the initial seeds (centers of each cluster) are E02, E04, and E06. Run the k-means algorithm for 1 epoch only and show the following: 4

EmployeeID	YearService	Income (K)
E01	4	9
E02	8	4
E03	2	10
E04	5	8
E05	6	4
E06	7	5

(3.67, 9)
(6.5, 4.5)

Table 1

i. Calculate the Euclidean distances between each point and the cluster centers.
ii. Determine the new clusters (i.e. the examples/points belonging to each cluster)
iii. Determine the centers of the new clusters and show the clusters after the first epoch and the new centroids.

b) "While clustering analysis in Machine Learning, Inter-cluster distances are minimized and Intra-cluster distances are maximized,"- agree or disagree? Justify your answer. 4

c) Find the problem solution using the Naïve Bayes Algorithm: 4

Given all the previous patients I've seen (below are their symptoms and diagnosis)...

chills	runny nose	headache	fever	flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	N	Mild	Y	Y

Do I believe that a patient with the following symptoms has the flu?

chills	runny nose	headache	fever	flu?
Y	N	Mild	Y	?

d) Explain the concepts of clustering approaches. How does it differ from classification? 4

5. Answer **Any Two** out of **Three** questions:

a) An HMM model is provided in Figure 1. States are E, M, and C. Observation symbols are H and L. Transition probabilities are shown in Figure 1 using arcs, the emission probabilities for each state are given in the rectangles underneath the states. The initial state probabilities, A, are provided on the left. 6

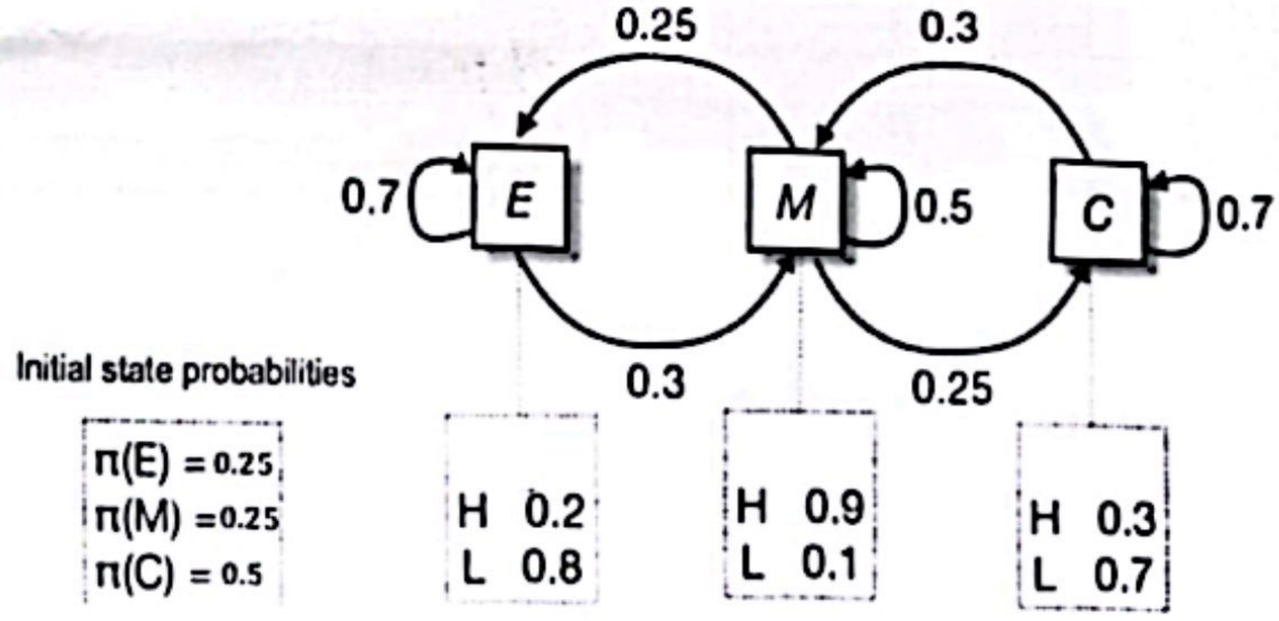
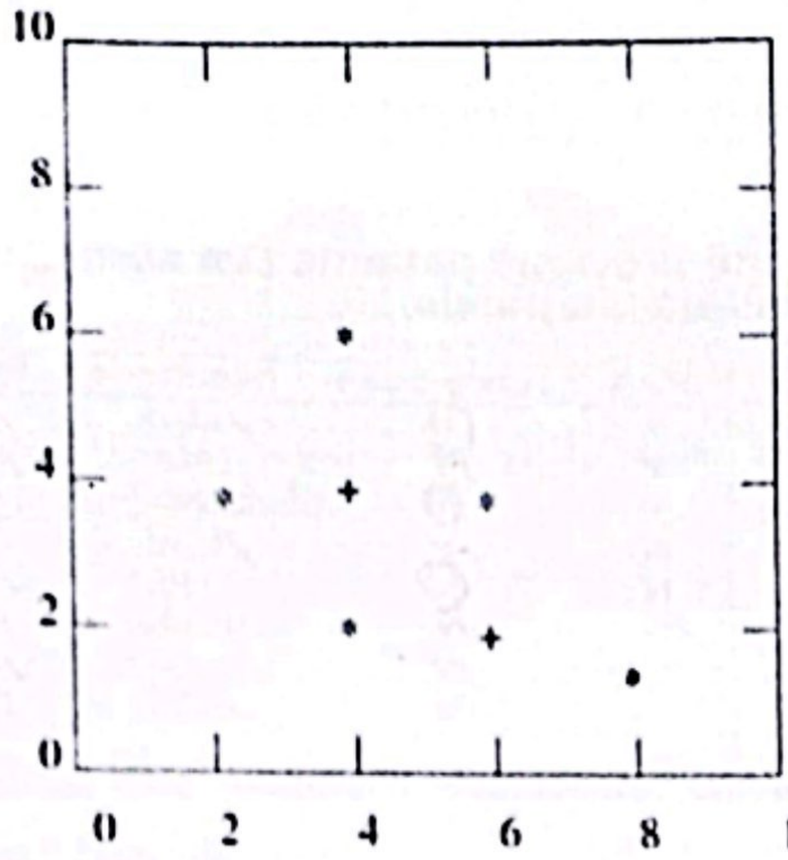


Figure 1: Emission Model

Use this model to compute the probability of the observed protein sequence HLHH is generated from the state sequence EMCC in the given diagram of the Hidden Markov Model. Show your formulation and calculations using the forward algorithm.

b) In the following questions, you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. Here, a plus indicates a positive example and a star a negative example. 6



In the above figure, how would the point (8,1) be classified using 2-nn? Show the calculation.

9) Consider the training dataset given in the following table. Use Weighted k-NN and determine the class. Test instance (7.6, 60, 8) and K=3. 6

Sl. No.	CGPA	Assessment	Project submitted	Results
1.	9.2	85	8	Pass
2.	8	80	7	Pass
3.	8.5	81	8	Pass
4.	6	45	5	Fail
5.	6.5	50	4	Fail
6.	8.2	72	7	Pass
7.	5.8	38	5	Fail
8.	8.9	91	9	Pass