



Roll Number												
-------------	--	--	--	--	--	--	--	--	--	--	--	--

Narula Institute of Technology
An Autonomous Institute under MAKAUT
2024
END SEMESTER EXAMINATION - EVEN 2024
CS802B - MACHINE LEARNING

TIME ALLOTTED: 3Hours**FULL MARKS: 70**

Instructions to the candidate:

Figures to the right indicate full marks.

Draw neat sketches and diagram wherever is necessary.

Candidates are required to give their answers in their own words as far as practicable

Group A

(Multiple Choice Type Questions)

Answer any ten from the following, choosing the correct alternative of each question: 10×1=10

- 1.i) The training examples closest to the separating hyperplane in SVM (1) CO3 BL1
are called _____.
a) Training vectors
b) Test vectors
c) Support vectors
d) None of these
- 1.ii) The average squared difference between classifier predicted (1) CO2 BL1
output and actual output is called
a) mean squared error
b) root mean squared error
c) mean absolute error
d) mean relative error
- 1.iii) The clustering algorithm which initially assumes that each data (1) CO4 BL3
instance represents a single cluster is called
a) agglomerative clustering
b) conceptual clustering
c) K-Means clustering
d) expectation maximization
- 1.iv) Which of the following is an example of feature extraction? (1) CO2 BL2
a) Constructing bag of words vector from an email
b) Applying PCA projects to a large high-dimensional data
c) Removing stopwords in a sentence
d) All of the above
- 1.v) Choose that following statement is true or false: True error is (1) CO1 BL1
defined over the entire instance space, and not just over training
data
a) True

b) False

1.vi) Which type of kernels are commonly used in kernel methods for support vector machines? (1) CO5 BL2

- a) Separable kernels
- b) Unseparable kernels
- c) Positive definite symmetric kernels
- d) Negative definite symmetric kernels

1.vii) Which method is primarily used for classifying data points based on their proximity to existing data points? (1) CO2 BL1

- a) Linear Models
- b) Probabilistic Models
- c) Distance-based Models
- d) Tree Models

1.viii) Which of the following machine learning algorithm is based upon the idea of bagging? (1) CO1 BL1

- a) Regression
- b) RF
- c) Classification

1.ix) Identify whether true or false: In PCA the number of input dimensions is equal to principal components. (1) CO1 BL1

- a) True
- b) False

1.x) Which type of model is specifically associated with learning ordered rule lists and subgroup discovery? (1) CO3 BL2

- a) Decision Trees
- b) Linear Models
- c) Rule Models
- d) Distance-based Models

1.xi) Which technique is commonly used to handle incomplete features in machine learning? (1) CO3 BL1

- a) Thresholding and discretization
- b) Normalization and calibration
- c) Feature selection using matrix transformations
- d) Structured features analysis

1.xii) What concept in model ensembles addresses the balance between the bias and variance of individual models within the ensemble? (1) CO4 BL2

- a) Bagging and random forests
- b) Boosted rule learning
- c) Mapping the ensemble landscape
- d) Meta learning

Group B
(Short Answer Type Questions)

(Answer any three of the following) 3x5=15

2. Explain briefly the working principle of Logistic Regression algorithm. (5) CO4 BL4
3. Consider the following confusion matrix: (5) CO1 BL4

		Actual values	
		Yes	No
Predicted values	Yes	12	3
	No	1	9

Compute the accuracy, precision, recall, f1-score, and specificity for the above problem.

4. Explain the concept of meta-learning in the context of machine learning experiments and discuss how meta-learning techniques can be utilized to improve the efficiency and effectiveness of machine learning experiments. (5) CO4 BL2
5. Short Answer Type Question. Explain in brief. (5)
- a) Explain the role of Mean Squared Error (MSE) in Machine Learning in brief. (3) CO2 BL4
- b) Highlight the role of gradient descent in linear regression. (2) CO2 BL1
6. Describe how decision trees are utilized in ranking and probability estimation tasks. Provide an example scenario where decision trees could be effectively applied for ranking or probability estimation, and discuss the advantages and limitations of using decision trees in such contexts. (5) CO1 BL1

Group C**(Long Answer Type Questions)****(Answer any three of the following) 3x15=45**

7. Long Answer Type Question. Explain in detail. (15)
- a) Define Concept Learning. (2) CO5 BL1
- b) Write down the tasks of concept learning. (4) CO5 BL6
- c) Explain how the concept learning can be viewed as a task of searching. (3) CO5 BL4
- d) Explain the following algorithms with proper examples: (6) CO5 BL4
- i) Find-S algorithm
- ii) Candidate Elimination algorithm

8. Long Answer Type Question. Explain in detail. (15)

a) Draw a comparative study of clustering and classification with examples.

(3) CO4 BL2

b)

(8) CO3 BL4

Weight	Height	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

On the basis of the above-given data, classify the below set of data as "Underweight" or "Normal" using KNN algorithm.

57	170	?
----	-----	---

c) Explain the following distance metrics with proper examples:

(4) CO3 BL4

- i) Euclidean distance ii) Manhattan distance iii) Minkowski distance
- iv) Hamming distance

9. Long Answer Type Question. Explain in detail.

(15)

a) Explain the concepts of entropy and information gain with examples.

(5) CO3 BL4

b) Apply ID3 algorithm for constructing decision tree for the following training example:

(10) CO4 BL2

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
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	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

10. Long Answer Type Question. Explain in detail.

(15)

a)

(8) CO4 BL4

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Using the K-Means Clustering algorithm, form two clusters (k1, k2) and assign the above data points to the appropriate clusters.

b)

(7) CO3 BL5

No.	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

If the weather is sunny, can a player play golf? Find out by implementing the Naive Bayes classifier algorithm.

- 11a. Describe the process of feature normalization and calibration in machine learning. How do these techniques improve model performance? Discuss any potential challenges or considerations in implementing them. (5) CO4 BL3
- 11b. Given six data points as (1,1), (2,1), (3,5), (4,3), (4,6), (6,4). Apply Hierarchical clustering algorithm to develop the dendrogram using these points. (5) CO4 BL2
- 11c. Compare and contrast the naïve Bayes classification model with other classification algorithms such as logistic regression and decision trees. Discuss the assumptions underlying the naïve Bayes model and how they might impact its performance in practice. (5) CO1 BL1