GURU NANAK INSTITUTE OF TECHNOLOGY

An Autonomous Institute under MAKAUT

2021

MACHINE LEARNING

CS802B

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

The figures in the margin indicate full marks.
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**

Marks **CO No**

1. (i) Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging? 1 CO1
a. Decision Tree
b. Regression
c. Classification
d. Random forest

(ii) SVM can be used to solve ___________ problems.
a. Classification
b. Regression 1 CO2
c. Clustering
d. Both Classification and Regression

(iii) SVM is a ___________ learning algorithm.
a. Supervised
b. Unsupervised 1 CO3
c. Both of these
d. None of these

(iv) SVM is termed as ________ classifier. 1 CO2
a. Minimum margin b. Maximum margin

(v) The training examples closest to the separating hyperplane are called as _______.
a. Training vectors 1 CO3
b. Test vectors
c. Support vectors
d. None of these

(vi) Which of the following is a type of SVM?
a. Maximum margin classifier b. Soft margin classifier c. Support vector regression d. All 1 CO3

(vii) How can you prevent a clustering algorithm from getting stuck in bad local optima? 1 CO3
a. Set the same seed value for each run
b. Use multiple random initializations
c. Both A and B

(viii) Which of the following is a disadvantage of decision trees? 1 CO3
a. Factor analysis
b. Decision trees are robust to outliers
c. Decision trees are prone to be overfit
d. None of the above

Data used to build a data mining model

(ix) a. validation data 1 CO1
b. training data
c. test data
d. hidden data

(x) Supervised learning and unsupervised clustering both require at least one 1 CO3
a. hidden attribute
b. output attribute
c. input attribute
d. categorical attribute

(xi) Supervised learning differs from unsupervised clustering in that supervised learning requires 1 CO3
a. at least one input attribute
b. input attributes to be categorical
c. at least one output attribute
d. output attributes to be categorical

(xii) Which of the following is true about Naive Bayes? 1 CO4
a. Assumes that all the features in a dataset are equally important
b. Assumes that all the features in a dataset are independent
c. Both A and B
d. None of the above

GROUP – B
(Short Answer Type Questions)

Answer any ***three*** from the following: **3×5=15**

Marks **CO No**

2. (a) What is Feed Forward Neural Network? 1 CO1

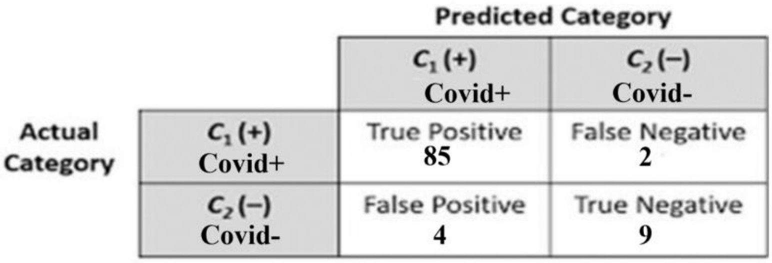
(b) Explain Backpropagation in details. 4 CO1

3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 5 CO3
13,15,16,16,19,20,20,21,22,22,25,25,25,25, 30,33,33,35,35,35,35,36, 40,45,46,52,70.
(i) Find the mean and median of the data.
(ii) What is the mode of the data? Comment on the data's modality.
(iii) Find first quartile(Q1) and third quartile(Q3) of the data.
(iv) Give the five-number summary of the data.
(v) Show the boxplot of the data

4. Suppose we have the following two-dimensional data set. Consider the data as 2d data points. Given a new point x = (1.4, 1.6) as a query, rank the database point based on similarity with the query using Euclidean distance and Cosine similarity. 5 CO4

	A1	A2
x1	1.5	1.7
x2	2	1.9
x3	1.6	1.8
x4	1.2	1.5
x5	1.5	1.0

5. (a) Compute accuracy, precision, recall, F-measure, sensitivity and specificity in respect of following classification model's outcome. 2 CO2



		Predicted Category	
		C_1 (+) Covid+	C_2 (−) Covid-
Actual Category	C_1 (+) Covid+	True Positive 85	False Negative 2
	C_2 (−) Covid-	False Positive 4	True Negative 9

(b) Illustrate Boxplot with respect to the given data: {199, 201, 236, 269,271,278,283,291, 301, 303, 341}. 3 CO2

6. (a) Apply the concept of regression model for the following dataset to determine the glucose level of a person having age 55. 3 CO2

AGE (X)	GLUCODE LEVEL (Y)
43	99
21	65
25	79
42	75
57	87
59	81

(b) Compute R-square value of the regression model with respect to above dataset. 2 CO2

GROUP – C
(Long Answer Type Questions)

Answer any ***three*** from the following: **3×15=45**

			Marks	**CO No**
7.	(a)	Find the value of the Lagrange multipliers and the extreme values of the following problem. $f(x, y) = x^3 + y^2$ $g(x, y) = x^2 - 1 >= 0$	5	CO2,3
	(b)	Derive the Wolfe-dual Lagrangian function of SVM.	5	CO2
	(c)	Calculates the Kth Principal Minors of a 3×3 Matrix.	5	CO3

$$A = \begin{bmatrix} 3 & -4 & 1 \\ 7 & 2 & 6 \\ -2 & 8 & 9 \end{bmatrix}$$

8.	(a)	Find the minimum value of the Rosenbrock's banana function $f(x, y) = (2-x)^2+100(y-x^2)^2$	5	CO2
	(b)	What is Lagrange Multiplier?	5	CO3
	(c)	Explain Hessian Matrix.	5	CO2

9 (a) Explain the working principle of Hierarchical Divisive Clustering strategy. 3 CO3

(b) Apply the concept of Hierarchical Agglomerative Clustering strategy on the following data to construct the dendrogram as diagrammatic representation of the entire clustering process. 10 CO3

Point	**X**	**Y**
P1	1	1
P2	1.5	1.5
P3	5	5
P4	3	4
P5	4	4
P6	3	3.5

(c) How can we measure the quality of a cluster? 2 CO3

10. (a) Discuss briefly about the feasibility of supervised learning strategy for predicting a specific disease based on a definite set of symptoms like loss-of-smell(yes/no), fever(yes/no), loss-of-appetite(yes/no), diarrhea (yes/no), runny-nose(yes/no), body-ache(yes/no), oxygen-saturation-level (very low/low/medium/high). 3 CO1

(b) Illustrate the concept of Naïve Bayes Classification strategy based on the following dataset and determine whether one should play tennis given outlook=sunny, temp=cool, humidity=high, windy=strong. 5 CO2

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**Play Tennis**
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

(c) Apply the principle of apriori algorithm to generate association rules with reference to the following transaction dataset of items: A, B, C, D, E where it is assumed that the minimum support value is 3 and confidence is 75%. 7 CO3,4

Transaction Id	Items
1	A, B, D, E
2	A, B, C, D, E
3	A, B, C, E
4	A, B, D
5	D
6	B, D
7	A, D, E
8	B, C

11. Write short notes on any three of the following: 3×5=15

(a) K-N-N algorithm. CO2

(b) Random Forest algorithm. CO1

(c) Support Vector Machine. CO1

(d) Over fitting problem CO3

(e) logistic Regression CO4