

RAMAKRISHNA MISSION VIDYAMANDIRA

(Residential Autonomous College affiliated to University of Calcutta)

B.A./B.Sc. SIXTH SEMESTER EXAMINATION, MAY 2024

THIRD YEAR [BATCH 2021-24]

COMPUTER SCIENCE [Honours]

Paper : DSE 3

Date : 17/05/2024

Time : 11 am – 1 pm

Full Marks : 50

Answer **any five** questions of the following :

[5×10]

1. a) The following dataset contains loan information and can be used to try to predict whether a borrower will default (the last column is the classification). Use the Naïve Bayes method to determine whether a loan X=(Home Owner = No, Marital Status=Married, Income=High) should be classified as a Defaulted Borrower or not.

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Low	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Low	Yes
9	No	Married	Low	No
10	No	Single	Low	Yes

- b) What do you mean by feature importance in ML?

(8+2)

2. a) Apply K nearest neighbour classifier to predict the class.

Height (CM)	Weight (Kg)	Class
160	50	Underweight
170	59	Normal
182	62	Normal
161	45	Underweight
170	80	Normal
165	70	Normal
168	48	?

- b) What is precision and recall?

(8+2)

3. a) Use the following dataset to explain the Principal Component Analysis (PCA) step-by- step. Find out the Principal Component from the given information.

Systolic BP	Diastolic BP
128	89
125	71
137	92
116	84
142	90
129	75

- b) Discuss the one-hot encoding using proper example.

(8+2)

4. a) Perform the K-means clustering on the following dataset, considering k= 2. Show all steps.

Item	Quantity	Price
Books	150	5500
Pen	200	1000
Color	100	8000
Bag	50	50,000

- b) How do you mean by online learning?

(8+2)

5. a) At the beginning of an exam, you try to predict whether each problem is easy or difficult (D = + if it is difficult and – if it is easy). Let us assume that you use two observable problem attributes (predicates): The text length L of the problem (L = 1 if it is long, 0 otherwise). The amount M of math in the text (M = 1 if there is a lot of math, 0 otherwise) For training data, assume that you have examined 12 previous problems from the homeworks, and have collected the following data:

L	M	D	#
0	0	-	4
0	0	+	1
0	1	-	0
0	1	+	3
1	0	-	1
1	0	+	2
1	1	-	1
1	1	+	0

The first line of this table reads as follows: 4 problems for which $L = 0$ and $M = 0$ were not difficult ($D = -$). The second line says: 1 problem for which $L = 0$ and $M = 0$ was difficult ($D = +$). Etc... Note that you observed no problem for which $L = 0$ and $M = 1$, or $L = 1$ and $M = 1$. Based on this training data, you want to compute a representation of a difficult problem (D) in the form of a decision tree using the two binary attributes L and M . Construct the best decision tree you can for the training data.

- b) What do you mean by Association Rule Mining? (8+2)
6. a) Discuss SVM in detail with an example. (8+2)
 b) What is Confusion Matrix, underfitting?
7. a) Given the following dataset find out the predicted infection of the persons using Linear Discriminant Analysis (LDA). Find out the specificity and sensitivity value. (Hints. We, know that Linear Discriminant (LD) = $\alpha_1 X_1 + \alpha_2 X_2$. Let the weight values $\alpha_1=0.11$, $\alpha_2=0.70$. These values give us the optimal separation between the groups).

Infection	CRP(mg/L)	Temp (C)
Viral	38.0	38.0
Viral	17.1	41.2
Viral	35.0	33.5
Bacterial	47.0	34.6
Bacterial	32.1	49.2
Bacterial	59.0	31.5

- b) What is the need of Gradient descent technique in ML. (8+2)

_____ × _____