Question Paper

Exam Date & Time: 04-Dec-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

FIFTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, DEC 2023

DATA MINING AND PREDICTIVE ANAYSIS [ICT 3171]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) Assume the following different observed values of age: 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. (5)

A)

i. Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].

ii. Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 year.

iii. Use normalization by decimal scaling to transform age value 35.

iv. Comment on which method you would prefer to use for the given data, giving reasons as to why.

B) Apply the K-Means algorithm's one iteration for a Two-Dimensional data set given as follows: {(2,2), (2.5,3),(4,5.5),(6,8),(4.5,6),(5,5),(3.5,4),(5.5,2.5)} with K=2 and (2,2) and (6,8) as initial cluster centers. Use Eucledian distance measure. (3)

C) Identify and discuss any two data mining challenges. (2)

2) Construct a FP-tree for the following transactional data set: T1: {1,2,3,5}, T2:{1,2,3,4,5}, T3:{2,4,1,5}, T4:{2,1,4}, T5:{4}, T6:{2,4}, T7:{1,4,5}, T8:{3,2}. Assume the minimum support count as 4. From FP tree, find out all frequent sets. (5)

A)

B) Considering the Two-Dimensional data set: {(8,7)(3,7),(4,9),(9,6),(8,5),(5,8),(7,3),(8,4),(7,5),(4,5)}. Find out 2 clusters with initial representatives as (4,5) and (8,5) by Applying K-Medoids. Use Manhattan distance. Also Find total cost incurred. (3)

C) List any two conditions under which density based clustering is more suitable than partition-based clustering and hierarchical clustering. (2)

3) Trading decisions can be made with a decision tree model using the Gini Index. Find the root node of the Decision Tree for the following example data by calculating the Gini Index for past trend, open interest, and trading volume (5)

A)

Past Trend	Open Interest	Trading Volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up

Positive	High	High	Up
Negative	Low	High	Down
Positive	Low	Low	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up

B) Find simple linear regression equation for the following 2 data sets: x=(2,4,6,8), y=(3,7,5,10) (3)

C) Determine the precision, recall, specificity, accuracy, F1-score, and error rate for a binary classification task whose confusion matrix is given below (2)

< caption>TABLE Q3C< /caption>

	PREDICTED POSITIVE	PREDICTED NEGATIVE
ACTUAL POSITIVE	400	70
ACTUAL NEGATIVE	50	80

4) Apply the Euclidean distance to the dataset using DBSCAN algorithm to identify outliers, core points, and border points. (Consider epsilon=1.9 units and MinPts= 4) (5)

A)

Points	X	Y
P1	7	4
P2	6	4
P3	5	6
P4	4	2
P5	6	3
P6	5	2
P7	3	3
P8	4	5
P9	6	5
P10	3	6
P11	4	4
P12	8	2

B) Consider a population data set D with a total of 30 individuals, which is being used to create a dataset that predicts whether a person will go to the gym or not. Out of these 30 individuals, 16 people go to the gym while 14 do not. The task is to calculate the entropy or info(D) and information gain for this dataset. (3)

	Go for Gym	Not go for gym	Total
Energy High	12	1	13
Energy low	4	13	17
Total	16	14	30

C) How is Post-pruning performed in Decision tree? Explain with an example. (2)

5) Consider a positive set of data points {(4,1), (4, -1), (6,1), (6, -1), (5,2), (5, -2)} and negative set of data point {(2,0), (1,1), (1 ,-1), (-1,0), (-2, 2), (-2, -1)}. Divide the data sets into 2 classes by using linear support vector machine. Assume (2,0), (4,1), (4,-1) as support vectors. (5)

A)

B) Estimate the conditional probabilities of each attribute (Font color, # paragraphs, Font size, Technical) for the document type classes (WORD, PPT) using the Naive Bayes' method for the data given in the following table. Using these probabilities estimate values for the new instance (Font Color = Blue, # paragraphs = 2, Font size = Big, Technical = No) (3)

Font Color	# paragraphs	Font size	Technical	Document type
Black	3	Medium	Yes	PPT
Blue	2	Big	No	PPT
Blue	3	Medium	Yes	PPT
Black	3	Medium	Yes	PPT
Blue	2	Medium	No	WORD
Black	2	Big	No	WORD
Black	2	Big	No	WORD
Black	2	Medium	Yes	WORD

C) Determine the sequences of transactions for all 5 customers for the transaction database given below. Find whether the following subsequences are frequent or not by considering minimum support count as 2. Write the sequence number that supports these sub sequences. C -> D, AB, D->G (2)

Customer	A	B	C	D	E	F	G	J

C4	0	0	1	0	0	0	0	0
C1	0	0	1	0	0	0	0	0
C3	0	0	1	0	1	0	1	0
C2	0	0	0	1	0	1	1	0
C2	0	0	1	0	0	0	0	0
C5	0	0	0	0	0	0	0	1
C1	0	0	0	0	0	0	0	1
C4	0	0	0	1	0	0	1	0
C2	1	1	0	0	0	0	0	0
C4	0	0	0	0	0	0	0	1
C1	0	0	1	1	0	0	0	0
C3	0	0	0	0	1	0	0	1
C5	1	1	0	0	0	0	0	0

-----End-----