Roll No.												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING

TEST - 1

Even Semester: 2018-19 | **Date**: 06 March 2019

Course Code: CSE 307 | **Time**: 1 Hour

Course Name: Data Mining | **Max Marks**: 40

Programme & Sem: B.Tech (DE) & VI Sem | **Weightage**: 20%

Instructions:

(i) ***Answer all questions sequentially***

(ii) ***Scientific calculators are allowed***

Part A

Answer **all** the Questions. **Each** question carries **three** marks. (5Qx3M=15)

1. Define the predictive and descriptive tasks of data mining. Also list the various tasks under each one of them.
2. Illustrate the various data mining steps involved in KDD with a neat diagram.
3. Differentiate noise and outliers with a suitable example for each.
4. Define equal width binning. Write the bin intervals used to discretize the following temperature values into 7 bins, using equal width binning.
   70,71,64,65,72,72,68,69,75,85,75,80, 83,81.
5. Write any two advantages of data warehouses.

Part B

Answer **both** the Questions. **Each** question carries **seven and half** marks. (2Qx7.5M=15)

6. The confusion matrix of a certain classifier that is trained to distinguish between positive and negative text statements is given below. Define and find each one of the following metrics namely, accuracy, error rate, TPR, FPR, TNR, FNR and Precision of this classifier.

	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	100	5
	-	10	50

7. Define the measures of node impurity namely, Gini index and classification error. Find the Gini Index and classification error of a subset with the class distribution given below.

Sports	20
Weather	40

Part C

Answer the Question. Question carries **ten** marks. (1Qx10M=10)

8. A decision tree based classification model is to be trained using the following data set to predict the factors affecting sunburn. Using multi-way split on the attributes, gain and entropy as the measure of node impurity, find the root node of the tree. Clearly show the detailed working with the gain of each candidate splitting attribute. **Draw the decision tree after the first iteration**.

Name	Hair	Height	Weight	Lotion	Sunburned
Sarah	Blonde	Average	Light	No	Yes
Dana	Blonde	Tall	Average	Yes	No
Alex	Brown	Short	Average	Yes	No
Annie	Blonde	Short	Average	No	Yes
Emily	Red	Average	Heavy	No	Yes
Pete	Brown	Tall	Heavy	No	No
John	Brown	Average	Heavy	No	No
Katie	Blonde	Short	Light	Yes	No

****

Roll No.												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING

TEST - 2

Even Semester: 2018-19
Course Code: CSE 307
Course Name: Data Mining
Program & Sem: B.Tech & VI Sem (DE)

Date: 16 April 2019
Time: 1 Hour
Max Marks: 40
Weightage: 20%

Instructions:

(i)Read the questions carefully and
(ii)State the assumptions, if any.

Part A

Answer **all** the Questions. **Each** question carries **five** marks. (3Qx5M=15)

1. Consider the two students S1 and S2, and their marks in two subjects "Data Mining" and "Compiler Design" are given below:

   S1=(90, 60) AND S2=(85, 70)

   a) Compute Manhattan Distance between S1 and S2
   b) Compute Euclidean Distance between S1 and S2
   c) Compute Supremum Distance between S1 and S2
   d) State the difference between Simple Matching Coefficient and Jaccard proximity measure

2. Consider the following Hostel related dataset ***DS*** and compute the proximity between the instances H1 and H3 as per the proximity measure used during the discussion in the class room.
   Note: Assume that the attribute "Hostel Facility" is nominal attribute and "Hostel Rating" is ordinal attribute.

Instance No.	Hostel Facility	Hostel Rating
H1	GOOD	1
H2	BAD	3
H3	MODERATE	2

3. a) Name at least two approaches for utilizing binary classifier for the multi class problem.
   b) Consider classes and its equivalent code word and classify the test instance ***T1*** with the code word " 1 0 1 1 1 1 1 " using Hamming Distance Measure.

Class	Code Word
Class1	1 1 1 1 1 1 1
Class2	1 1 0 0 1 1 1
Class3	1 1 1 0 0 0 1

Part B

Answer **both** the Questions. (2Q=15Marks)

4. Consider the following dataset "***DS_AdaBoost"*** and Sample data set ***"Sample_DS"*** for AdaBoost Algorithm with one iteration. (8 Marks)

DS_AdaBoost

X	0.2	0.4	0.6	0.8	1
Y	1	-1	1	-1	-1

Sample_DS

X	0.2	0.2	0.6	0.8	1
Y	1	1	1	-1	-1

i. Calculate error rate
ii. Calculate confidence factor
iii. State the importance of updating weight in the AdaBoost as compared to bagging approach.

5. Consider the following plant dataset DS1 and apply K-NN Classifier with K=2 and classify the instance ***Test= (100, 200, ?) (i.e. if plant length=100 and plant width=200, find the plant type attribute value).*** (7 Marks)

Note: Euclidean Measure will be used as a proximity measure

Instance No.	Plant length	Plant Width	Plant type
1	100	150	Type1
2	110	140	Type1
3	130	200	Type2
4	110	205	Type2

Part C

Answer the Question. The Question carries **ten** marks. (1Qx10M=10)

6. Consider the following training Dataset D.

Screen size	Type	Company	Purchase?
A	C	A	Yes
C	A	C	Yes
A	A	C	No
B	C	A	No
C	C	B	No

Apply Naïve Bayes Classifier and classify the test record with the following values ***"A, C, B, ? ".***

Roll No												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU
SCHOOL OF ENGINEERING

SUMMER TERM / MAKE UP ENDTERM EXAMINATION

Semester: Summer Term 2019 | **Date**: 24 July 2019

Course Code: CSE 307 | **Time**: 2 Hours

Course Name: Data Mining and Warehouse | **Max Marks**: 80

Program: B.Tech (CSE) & VI Sem (2015 Batch) | **Weightage**: 40%

Instructions:

(i) ***Answer all the questions and state the assumptions if any***

(ii) ***Scientific calculators are allowed***

Part A

Answer **all** the Questions with suitable answer from the given options. **Each** question carries **1** mark. (20Qx1M=20M)

1.

i. The earliest step in the data mining process is usually?
A. Visualization **B.** Modelling **C.** Preprocessing **D.** Deployment

ii. Which of the following operations can be performed on nominal attributes?
A. Distinctness **B.** Order **C.** Addition **D.** Multiplication

iii. Friendship structure of users in a social networking site can be considered as an example of:
A. Record data **B.** Ordered data **C.** Graph data **D.** None of these

iv. Leaf nodes of a decision tree correspond to:
A. Attributes **B.** classes **C.** Data instances **D.** None of these

v. Which of the following applied on warehouses?
A. Write only **B.** Read only **C.** Both A & B **D.** None of these

vi. Which of the following statement is NOT true about clustering?
A. It is a supervised learning technique
B. It is an unsupervised learning technique
C. It is also known as exploratory data analysis
D. It groups data into homogeneous groups

vii. _____ is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.
A. Data mining **B.** Data warehousing **C.** Web mining **D.** Text mining

viii. Apriori algorithm is otherwise called as _____.
A. Width wise algorithm **B.** Level wise algorithm **C.** Pincer search **D.** FP growth

ix. Which of the following method is not a Normalization method?
A. Equal width Binning **B.** Min-Max **C.** Z-score **D.** Decimal scaling

x. Match the following: (4Qx1M=4M)

A. ID numbers	-	Ordinal
B. Ratings (1 to 5)	-	Nominal
C. Calendar Dates	-	Ratio
D. Age	-	Interval

xi. The output of KDD is _____.
A. Data **B.** Information **C.** Query **D.** Useful information

xii. K-means is not deterministic and it also consist of number of iterations. **True** or **false**?

xiii. _____ is the goal of data mining.
A. To explain some observed event or condition. **B.** To confirm that data exists.
C. To analyze data for expected relationships. **D.** To create a new data warehouse.

xiv. Which of the following is required by K-means clustering ?
A. Defined distance metric **B.** Number of clusters
C. Initial guess as to cluster centroids **D.** All of these

xv. Which is not a common property of a Distance metric?
A. Symmetry **B.** Positive definiteness **C.** Triangle inequality **D.** Dissimilarity

xvi. Distance between two clusters in single linkage clustering is defined as:
A. Distance between the closest pair of points between the clusters
B. Distance between the furthest pair of points between the clusters
C. Distance between the most centrally located pair of points in the clusters
D. None of these

xvii. How do you calculate Confidence(A -> B)?
A. Support(A∪B) / Support (A) **B.** Support(A∪B) / Support (B)
C. Support(A∩B) / Support (A) **D.** Support(A∩B) / Support (B)

Part B

Answer **all** the Questions. **Each** question carries **twelve** marks. (3Qx12M=36M)

2. Why cluster validation is required? Explain any two different measures of cluster validation in detail.

3. Consider the following data set which describes the weather condition for playing some unspecified game.

Outlook	Temp	Humidity	Windy	Play
sunny	hot	High	false	no
sunny	hot	High	true	no
overcast	hot	High	false	yes
rainy	mild	High	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

Apply Naïve Bayes classifier and classify the following test record.

Outlook	Temp	Humidity	Windy	Play
Sunny	cool	High	true	?

4. For the following given transaction data set, generate frequent item sets using Apriori algorithm by considering support=22%.

Transaction ID	1	2	3	4	5	6	7	8	9
Items purchased	I1,I2,I5	I2,I4	I2,I3	I1,I2,I4	I1,I3	I2,I3	I1,I3	I1,I2,I3,I5	I1,I2,I3

Part C

Answer the Questions. **Each** question carries **twenty four** marks. (1Qx24M=24M)

5. Consider the data set given below and apply agglomerative clustering using Single Linkage. Draw the Dendogram for the same.

Item	A	B	C	D	E	F
A	0					
B	0.71	0				
C	5.66	4.95	0			
D	3.61	2.92	2.24	0		
E	4.24	3.54	1.41	1.00	0	
F	3.20	2.50	2.50	0.50	1.12	0