ID NO.	

PRESIDENCY UNIVERSITY, BENGALURU

SCHOOL OF ENGINEERING

Weightage: 40% Max Marks: 80 Max Time: 2 Hrs. 11 May 2018, Friday

END TERM FINAL EXAMINATION MAY 2018

Even Semester 2017-2018 Course: **CSE 307 Data Mining and WareHouse** VI Sem. CSE

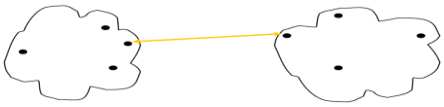
Instructions:

i. Assume missing data appropriately, if any
ii. Answers to all questions, use proper diagrams wherever necessary.
iii. Question paper consists of three parts. Part A, B and C are closed book type.

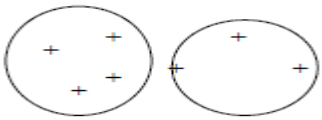
Part A

(5 Q x 4 M = 20 Marks)

1. Association rules with high confidence will be generally preferred in most of the applications.
   a. Often, we will not be interested in association rules that have a confidence of 100%. Why?
   b. Specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate) as compared to association rules with 100% confidence?
2. Compare the Pros and Cons of decision tree and artificial neural network classification methods.
3. a. What are ensemble Classifiers? Explain with a Diagram.
   b. State the advantage(s) of the ensemble classifier algorithm over decision-tree classifier.
4. a. Consider the figure given below with two clusters. Clusters are connected by a line which represents the distance used to determine inter-cluster similarity. Which inter-cluster similarity metric does this line represent?



   b. Are the two clusters shown below well separated? Justify your answer.



5. For a two-class classification problem, with a Positive class P and a negative class N, we can describe the performance of the algorithm using the following terms: TP, FP, TN, and FN.
   a. Place these four terms listed above into the appropriate slots in the table below.

		PREDICTED	
		POSITIVE	NEGATIVE
ACTUAL	POSITIVE		
	NEGATIVE		

   b. Provide the formula for precision and recall using TP, TN, FP, and FN.

Part B

(14+12+14M = 40 Marks)

6. Consider the following dataset with five objects. Assume that you are applying k-means clustering algorithm with k=3 and Euclidean distance measure to cluster examples into three clusters. Also, assume that the initial centroids (centers of each cluster) are A1, A4 and A7.

Object	Attribute1	Attribute2
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A7	1	2

a. Show the new clusters after first iteration with all the intermediate calculations.
b. Show the centers of the new clusters after the first iteration.

7. Consider the following Dataset DS.

OBJECT	Attribute1	Attribute2
A	1	3
B	2	6
C	5	3
D	3	2

Apply Agglomerative ***MAX or Complete Link*** Hierarchical Clustering algorithm on the dataset DS. Show the steps with calculations and dendrogram.
Note: Consider the Euclidean Distance measure as proximity metric.

8. Consider the following Dataset D.

Screen size	**Type**	**Company**	**Purchase?**
Medium	Laptop	DELL	Yes
Medium	Laptop	DELL	No
Medium	Laptop	DELL	Yes
Large	Laptop	DELL	No
Large	Laptop	HP	Yes
Large	PC	HP	No
Large	PC	HP	Yes
Large	PC	DELL	No
Medium	PC	HP	No
Medium	Laptop	DELL	No

Apply Naïve Bayes Classifier and classify the test record with the following values
"Medium, PC, DELL, ? "

Part C

(1 Q x 20 M = 20 Marks)

9. Consider the following transactional database where I1, I2, I3, I4, I5, I6, I7 are items. Assume the minimum support as 60%.

ID	**Items**
T1	I1, I2, I3, I5
T2	I1, I2, I3, I4, I5
T3	I1, I2, I3, I7
T4	I1, I3, I6
T5	I1, I2, I4, I5, I6

a. Find all frequent itemsets. Indicate each candidate set C_k, k = 1, 2, …, the candidates that are pruned by each pruning step, and the resulting frequent itemsets L_k.
b. Generate all possible association rules based on the frequent item set and list out the same.
Note: No need to compute and show the confidence for the rules

ID NO:	

PRESIDENCY UNIVERSITY, BENGALURU

SCHOOL OF ENGINEERING

Weightage: 20% Max Marks: 40 Max Time: 1 hr. 28 March Wednesday 2018

TEST – 2 **SET A**

Even Semester 2017-18 Course: **CSE 307 DATA MINING AND WAREHOUSE** VI Sem. CSE

Instruction:

(i) Read the questions properly and answer accordingly.

(ii) Question paper consists of 3 parts.

Part A

(3 Q x 3 M = 9 Marks)

1. a. Define Overfitting and Underfitting.
   b. Mention the application/criteria where Overfitting is not at all a problem.
2. State the advantage of the RIPPER in terms of Instance Elimination approach.
3. Consider a tree with 15 leaf nodes and 30 errors on training (out of 1000 instances).
   a) Calculate the training error.
   b) Calculate Generalization error based on Pessimistic approach.

Part B

(2 Q x 8 M = 16 Marks)

4. Consider a training set that contains 200 positive data instances (class = "+") and 500 negative data instances (class = "-"). Consider the following rules R1 and R2 with the following scenarios:

   R1: (Alpha=10)→ class = "+" (covers 40 positive and 12 negative data instances)

   R2: (Beta=8) → class = "+" (covers 35 positive and 15 negative data instances)

   Calculate FOIL's information gain (as done by the RIPPER algorithm) for each rule and state which rule will be selected as a best rule by FOIL's information gain metric. Show your work.

5. Consider the following data set DS.

S.No.	Attribute1	Attribute2	Attribute3	Attribute4	Class
1	C1	1	3	no	YES
2	C1	2	2	yes	NO
3	C2	0	2	yes	NO
4	C1	0	2	no	YES
5	C3	1	1	no	YES
6	C2	2	1	no	NO
7	C2	1	1	no	NO
8	C1	0	3	yes	NO

Consider a decision tree construction using ID3 algorithm [Note: Use entropy calculations for feature/attribute selection].

a. Identify the root attribute and show the calculation.
b. Show the child nodes records as per the root attribute.

Part C

(1Q x 15 M = 15 Marks)

6. Consider the following general sequential covering algorithm used to construct classification rules as per the discussion in the class room and answer the following questions:

Sequential Covering Algorithm:

1. Let D be a dataset of training data instances with *n* predictive attributes *A1,…An*, and a target attribute C with possible values *c1,…,ck.*
2. Let RuleSet = {} be the initial rule list.
3. **for** each class *ci* in C **do**
4. **while** stopping criterion is not met **do**
5. R ← Learn-One-Rule(D,*ci*)
6. D ← D – data instances covered by R (i.e., remove training data instances from D that are covered by R)
7. RuleSet ← RuleSet U R (i.e., add R at the bottom of the rule list in RuleSet)
8. **end-while**
9. **end-for**

a. In what order does the RIPPER algorithm consider the class values *c1,…,ck* in line 3 while constructing rules? Explain.

b. Consider the ***LearnOneRule*** function in line 5 of the algorithm.

i. What rule growing approach is used in RIPPER?

ii. Explain about the metric used in RIPPER to select the best candidate condition among the candidate conditions to add to the rule?

c. What stopping criterion does RIPPER use in line 4 of the algorithm above?

d. Briefly explain the rule set optimization process in RIPPER algorithm.

ID NO:	

PRESIDENCY UNIVERSITY, BENGALURU

SCHOOL OF ENGINEERING

Weightage: 20 % Max Marks: 40 Max Time: 1 hr. 20 Feb Tuesday 2018

TEST – 1

Even Semester 2017-18 Course: **CSE 307 DATA MINING AND WAREHOUSE** VI Sem. CSE

Instruction:

(i) Read the questions properly and answer accordingly.

(ii) Question paper consists of 3 parts.

Part A

(3 Q x 3 M = 9 Marks)

1. a) Distinguish between Classification and Regression.
   b) Distinguish between Data Mining, DBMS and OLAP.
2. Suppose a group of persons with the sorted medical store credit points listed as follows: 0, 400, 1200, 1600, 1600, 1800, 2400, 2600, 2800
   a) Partition them by Equi-width binning – for bin width of 1000
   b) Partition them by Equi-frequency binning – for bin density of 3
3. Compute the SMC similarity, the Jaccard similarity and L1 distance between the following two binary vectors x and y:
   x = 0101010001
   y = 0100011000

Part B

(2 Q x 8 M = 16 Marks)

4. Explain the steps of KDD process with a diagram.
5. a. What is feature subset selection?
   b. Explain the different approaches for feature subset selection?
   c. Illustrate a scenario where feature subset selection takes care of the Curse of Dimensionality

Part C

(1Q x 15 M = 15 Marks)

6. Consider the training set shown below for a binary decision tree classification problem.
   (a) Compute the Gini index (Gini(t) = 1 $-\sum_{i=0}^{c-1}[P(i|t)]^2$) for the overall collection of training examples.
   (b) Compute the Gini (split) for the Gender attribute.
   (c) Compute the Gini (split) for the Type attribute using multi-way split.
   (d) Compute the Gini (split) for the Size attribute using multi-way split.
   (e) Which attribute is better: Gender, Type, or Size?
   (f) Justify multi-way split leads to less impurity as compared to two-way/Binary split.

TRAINING SET:

Record No.	GENDER	TYPE	SIZE	CLASS/TARGET
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1