DOF
2814116.

SPRING END SEMESTER EXAMINATION-2016

2nd Semester M.Tech

DATA WAREHOUSING AND DATA MINING

CS-6301

(Regular-2015 & Back of Previous Admitted Batches)

Time: 3 Hours **Full Marks: 60**

Answer any SIX questions including Question No.1 which is compulsory.

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

1. a) Differentiate between OLTP and OLAP. [2 × 10

   b) What are the characteristics of data warehouse?

   c) Why is Business Intelligence (BI) important in data warehousing?

   d) What is Cube and Linked Cube with reference to data warehouse?

   e) What is the difference between "supervised" and unsupervised" learning scheme?

   f) What is meant by pruning in a decision tree induction?

   g) Find out the distance between two objects represented by attribute values (1, 6, 2, 5, 3) and (3, 5, 2, 6, 6) by using any two of the distance measures.

   h) Consider the one-dimensional data set shown in the following Table.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

   Classify the data point x = 5.0 according to its 1-, 3-, 5- and 9-nearest neighbors (using majority votes).

   i) Differentiate between agglomerative and divisive hierarchical clustering.

j) Mention the merits and demerits of hierarchical clustering.

2. a) Explain the components of a data warehousing system. [4

b) Discuss about the typical OLAP operations on multidimensional data with a suitable example. [4

3. a) Explain data reduction and data cube aggregation. [4

b) Consider the following data set. [4

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

i) Compute the support for itemsets {c}, {b, d}, {b, d, c} by treating each transaction ID as a market basket.

ii) Using the previous results, compute the confidence for the association rules {b, d} $\rightarrow$ {e} and {e} $\rightarrow$ {b, d}. Is confidence a symmetric measure?

4. Consider the training example shown in the table for binary classification problem. [8

Instance	A	B	C	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

What is the entropy of this collection of training examples with respect to the positive class?

Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Calculate the gain in Gini index when splitting on A and B. Which attribute would the decision tree algorithm choose?

5. Consider the data set shown in the following table.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

a) Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|-), P(B|-), P(C|-). [2

b) Use the estimate of conditional probabilities in (a) to predict the class label for a test sample (A=0, B=1, C=0) using naïve Bayes approach. [6

6. a) State the advantages of the decision tree approach over other approaches for performing classification. [4

b) Use the k-means algorithm and Euclidean distance the 10 given points into three clusters: $X_1(2, 10)$, $X_2(2, 5)$, $X_3(8, 4)$, $X_4(9, 4)$, $Y_1(5, 8)$, $Y_2(7, 5)$, $Y_3(6, 4)$, $Z_1(1, 2)$, $Z_2(4, 9)$, $Z_3(6, 10)$. Suppose that the initial cluster centers are X_1, X_4 [4

and Z_2. Run the k-means algorithm for 3-iterations. At the end of each iteration, show:

i) The new clusters. (i.e. the points belong to each cluster)

ii) The centers of new clusters.

7. a) What is outlier analysis? How to handle outlier? Explain with a suitable example. [4

b) Discuss constraint based cluster analysis with a relevant example. [4

8. Write short notes (Any four) [2 × 4

a) Data warehouse meta data

b) Data warehouse and data mart

c) Data transformation

d) Market basket analysis

e) Constraint based association mining

– *****