NagpurStudents

www.nagpurstudents.org

B.E. (Computer Science & Engineering) Seventh Semester (C.B.S.)

Data Warehousing & Mining

P. Pages : 3 **NIR/KW/18/3571**

Time : Three Hours * 1 8 5 4 * Max. Marks : 80

Notes :
1. All questions carry marks as indicated.
2. Solve Question 1 OR Questions No. 2.
3. Solve Question 3 OR Questions No. 4.
4. Solve Question 5 OR Questions No. 6.
5. Solve Question 7 OR Questions No. 8.
6. Solve Question 9 OR Questions No. 10.
7. Solve Question 11 OR Questions No. 12.
8. Assume suitable data whenever necessary.
9. Illustrate your answers whenever necessary with the help of neat sketches.
10. Use of non programmable calculator is permitted.

1. a) Explain following data mining functionalities:- **6**

i) Association and correlation Analysis.

ii) Clustering

iii) Classification.

Give example of each data mining functionality.

b) Describe why concept hierarchies are useful in data mining. Draw concept hierarchy for location dimension. **7**

OR

2. a) Suppose that the data for analysis include the attribute height. The height values for the data tuples are 14, 16, 17, 17, 20, 21, 21, 22, 23, 23, 26, 26, 26, 26, 31, 34, 34, 36, 36, 36, 36, 37, 41, 46, 47, 53, 71. **7**

i) What is mean and median of data.

ii) What is mode of data. Comment on Modality.

iii) What is Midrange, Quartile (Q_1) and Quartile (Q_3) of the data.

iv) Give five . Number summary of the data.

v) Show boxplot of a data.

b) In real-world data, tuples with missing values for some attributes are a common occurrence. Give various methods for handling this problem. **6**

3. a) Bring out the difference between OLAP and OLTP. **5**

b) Suppose data Warehouse consists of three dimensions time, doctor and patient and the two measures count and charge, where charge is the fee that doctor charges patient for a visit. **8**

i) Draw star schema for above data warehouse

ii) Convert star schema into snowflake schema

OR

4. a) Consider data warehouse for university consists of the following four dimensions: Student, course, semester and instructor and two measures count and Avg-grade. **6**

a) Draw a snowflake schema diagram for the data warehouse.

b) If each dimension has four levels (Excluding All), how many Cuboids will this cube contains?

b) Draw and explain data warehouse architecture along with neat sketch. **7**

5. a) Given the following six transaction on items {A, B, C, D, E} **8**

T_{ID}	Items
1	A, B, C
2	A, B, C
3	B, C
4	B, D
5	B, C, D, E
6	E

Use the Apriori Algorithm to compute frequent itemsets and their support.
Generate association rules from frequent itemset Minimum support count is 2.

b) Write short notes on constraint – based association Mining. **6**

OR

6. a) Explain following terms. **8**

i) Frequent itemsets ii) Closed itemsets
iii) Confidence iv) Support.

b) Given the following six transaction on items. **6**

T_{ID}	Items
T_1	A, B, E
T_2	B, C, D
T_3	B, D, E
T_4	C, D, E
T_5	B, C, D, E
T_6	B, C, E

Use FP – growth algorithm to compute frequent itemsets. Draw FP – tree.
Minimum support is 20%.

7. a) Briefly outline major steps of decision tree classification. **6**

b) Compare the advantages and disadvantages of eager classification versus lazy classification. **7**

OR

8. a) Why is Naive Bayesian classification called Naive? Briefly outline major ideas of Naive Bayesian classification. **6**

b) Write short Notes on:- **7**

i) Rule Based classification ii) Support vector Machine.

9. a) Point out the difference between classification and clustering. **5**

b) The distance between five pair of cases given below: **9**

	A	B	C	D	E
A	0				
B	8	0			
C	2	6	0		
D	5	4	8	0	
E	10	9	3	7	0

Cluster the five cases using below procedure and Draw Dendrogram structure.

i) Single Linkage Hierarchical procedure
ii) Complete linkage Hierarchical procedure
iii) Average linkage Hierarchical procedure.

OR

10. a) Consider the following cluster **6**

$A_1(2,10)$ $A_2(2,5)$ $A_3(8,4)$ $B_1(5,8)$
$B_2(7,5)$ $B_3(6,4)$ $C_1(1,2)$ $C_2(4,0)$

The distance function is Euclidean distance. Suppose initially A_1, B_1 and C_1 as the center of each cluster, respectively. Use k-means algorithm to show only.

i) Three cluster center after first round of execution
ii) Final three clusters after second round of execution.

b) Write short notes on:- **any two.** **8**

i) Density based clustering.
ii) Outlier Analysis
iii) Constraint Based cluster analysis.

11. a) Describe the following Methodologies for stream Data processing: **6**

i) Random sampling
ii) Histograms

b) Write short notes on social Network Analysis. Give real life example to support your answer. **7**

OR

12. a) What is multi-relational data mining? Explain various approaches for multi-relational classification. **6**

b) Illustrate how sequence pattern can be mined in biological data. **7**

****

NagpurStudents

www.nagpurstudents.org

The best time to plant a tree was 20 years ago. The second best time is now.

~ Chinese Proverb