Anekant Education Society's

Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati

(**Autonomous**)

Department of Computer Science

Class : M.Sc.(Computer Science) – I **Semester : II**

Paper Title : Data Mining and Data Warehousing (C) **Paper Code : COMP4202**

Question Bank

Type of Question : Objective Questions

1. Which of the following is not a tree pruning method?
    A. Reduced error pruning
    B. Cost complexity pruning
    C. Alpha Beta Pruning
    D. None

2. Basic approaches to graph mining are
    A. Apriori based approach
    B. Pattern growth approach
    C. All
    D. None

3. The type of relationship in star schema is _____.
    A. many-to-many.
    B. one-to-one.
    C. one-to-many
    D. many-to-one.

4. Record cannot be updated in _____.
    A. OLTP
    B. File
    C. RDBMS
    D. data warehouse

5. Information is
    A. Data
    B. Processed Data
    C. Manipulated input
    D. Computer output

6. For taking decisions data must be
    A. Very accurate
    B. Massive
    C. Processed correctly
    D. Collected from diverse sources

7. FP Growth algorithm uses
    A. 4 passes over dataset
    B. 2 passes over dataset
    C. n passes over dataset
    D. n X m passes over dataset

8. Market Basket Analysis is an example of
    A. FP Tree
    B. Frequent Item Set
    C. FP - Growth
    D. None of the Above

9. _____ are responsible for running queries and reports against data warehouse tables.

    A. Hardware.
    B. Software
    C. End users.
    D. Middle ware.

10. The process of checking some sequence of tokens for presence of constituents of some pattern is called as
    A. Pattern Matching
    B. Machine Learning
    C. Machine Matching
    D. Pattern Learning

11. _____ is a three or higher dimensional array of values, commonly used to describe a time series of image data.
    A. Data Cube
    B. Cluster
    C. Data Warehouse
    D. Data Mart

12. _____ is the input to KDD.
    A. Data.
    B. Information.
    C. Query.
    D. Process.

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

14. OLAP stands for
    A. Online Analysis Process
    B. Open Analytical Process
    C. Online Analytical Processing
    D. Open Access Process

15. OLAP server is
    A. Top tier of DW architecture
    B. Bottom tier of DW architecture
    C. Middle tier of DW architecture
    D. None of above

16. _____ maps data into subsets and the applies a compact description for that subset.
    A. Description
    B. Summarization
    C. Clustering
    D. Classification

17. Data Mining is a _____ step of knowledge discovery in database process.
    A. Design
    B. Analysis
    C. Testing
    D. Coding

18. Classification is
    A. A subdivision of a set of examples into a number of classes
    B. A measure of accuracy of classification of concept i.e. given by certain theory
    C. The task of assigning a classification to a set of examples
    D. None

19. The star schema is composed of _____ fact table
    A. One
    B. Two
    C. Three
    D. Four

20. Data mining is
    A. The actual discovery phase of a knowledge discovery process
    B. The stage of selecting the right data for a KDD process
    C. A subject-oriented integrated time variant non-volatile collection of data in support of management
    D. None

21. Fact tables are _____
    A. Completely Demoralized
    B. Partially Demoralized
    C. Completely Normalized
    D. Partially Normalized

22. Classification rules are extracted from _____
    A. Root node
    B. Decision tree
    C. Nodes
    D. Branches

23. Data warehouse architecture is based on _____
    A. DBMS
    B. RDBMS
    C. Sybase
    D. SQL Server

24. The full form of KDD is _____
    A. Knowledge database
    B. Knowledge discovery in database
    C. Knowledge data house
    D. Knowledge data definition

25. Various visualization techniques are used in _____ step of KDD.
    A. Selection
    B. Transformation
    C. Data mining
    D. Interpretation

26. Incorrect or invalid data is known as _____
    A. Changing data
    B. Noisy data
    C. Outliers
    D. Missing data

27. ROI is an acronym of _____
    A. Return on Investment
    B. Return on Information
    C. Repetition of Information
    D. Runtime of Instruction

28. The left hand side of an association rule is called _____.
    A. Consequent
    B. Onset
    C. Antecedent
    D. Precedent

29. The FP-growth algorithm has _____ phases
    A. One
    B. Two
    C. Three
    D. Four

30. _____ is the heart of the warehouse.
    A. Data mining database servers.
    B. Data warehouse database servers.
    C. Data mart database servers.
    D. Relational data base servers.

31. Rule based classification algorithms generate ____ rule to perform classification.
    A. if-then
    B. while
    C. do while
    D. switch

32. A _____ is necessary condition for KDDs effective implementation.
    A. Data set
    B. Database
    C. Data warehouse
    D. Data

33. Metadata is used by the end users for _____
    A. Managing database
    B. Structuring database
    C. Querying purposes
    D. Making decisions

34. The partition of overall data warehouse is _____
    A. Database
    B. Data cube
    C. Data mart
    D. Operational data

35. The _____ operation is used for reducing data cube by one or more dimensions.
    A. Drilling
    B. Rolling
    C. Dicing
    D. Slicing

36. Removing duplicate records is a process called_____
    A. Recovery
    B. Data cleaning
    C. Data reduction
    D. Data pruning

37. The data Warehouse is_____
    A. read only.
    B. write only.
    C. read write only.
    D. none.

38. Data mining is an integral part of _____
    A. SE
    B. DBMS
    C. KDD
    D. OS

39. _____ 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.

40. The star schema is composed of _____ fact table.
    A. one
    B. Two
    C. Three
    D. Four

41. Leave one out is a special case of _____
    A. Confusion matrix
    B. Cross validation
    C. Recall
    D. Precision

42. Page rank is used for _____
    A. Create page
    B. Delete page
    C. Prioritize page
    D. Update page

43. Agglomerative clustering is _____
    A. Top down strategy
    B. Bottom up strategy
    C. Both
    D. None

44. Weka is fully implemented in _____
    A. Dot net
    B. C
    C. Java
    D. PHP

45. Which of the following is not a clustering algorithm?
    A. Apriori
    B. K-means
    C. EM
    D. Hierarchical

46. Correlation clustering is also called as _____
    A. Cluster editing
    B. Cluster updating
    C. Cluster booking
    D. Cluster printing

47. Active learning is a form of _____
    A. Supervised machine learning
    B. Semi-supervised machine learning
    C. None
    D. Both

48. Input node of perceptron is called as _____
    A. Associators
    B. Photo receptors
    C. Responder
    D. All

49. Output node of perceptron is called as _____
    A. Associators
    B. Photo receptors
    C. Responder
    D. All

50. Which panel displays scatter plot matrix for the current data set.
    A. Visualize panel
    B. Associate panel
    C. Classifier panel
    D. Preprocess panel

Type of Questions : Answer in One / Two Sentence

1. Define Data Mining?
2. Define Data warehouse?
3. Define Machine learning.
4. Define Pattern Matching.
5. Define Data Cube.
6. What is fact constellation schema?
7. What is association rule?
8. What is Apriori property/downward closure property?
9. What is predictor variable?
10. What is response variable?
11. What is accuracy?
12. What is precision?
13. What is low precision?
14. What is high precision?
15. What is sensitivity?
16. What is specificity?
17. What is R?
18. What is weka?
19. What is active learning?
20. What is reinforcement learning?

Type of Questions : Short Notes

1. Write a note on data mining tasks.
2. Write a short note on knowledge discovery in databases.
3. Write a short note on Data Mart.
4. Write a short note on star schema.
5. Write a short note on snowflake schema.
6. Write a note on Sequence Mining.
7. Write a note on Tree Mining.
8. Write a note on Graph Mining.
9. Write a note on FP-Growth algorithm.
10. Write a note on Decision Tree Algorithm.
11. Write a note on attribute selection.
12. Write a note on Bayesian network.
13. Write a note on Tree Pruning.
14. Write a note on linear regression.
15. Write a short note on CART.
16. Write a note on Least Square.
17. Write a note on Perceptron.
18. Write a note on Support Vector Machine.
19. Write a note on Text mining approaches
20. Write a note on web mining applications.

Type of Questions : Short Answer Questions

1. What are the social implications of data mining?
2. Compare and contrast data warehouse and data mart.
3. State the difference between logical design and physical design.
4. What are the different categories of OLAP?
5. What are the applications of frequent item sets?
6. What are the advantages and disadvantages of FP-Growth algorithm?
7. What is joint probability distribution?
8. What is inference? What are the three main inference tasks of Bayesian network?
9. Write a note on Recall.
10. Write a note on F-measure.
11. Write a note on Confusion matrix.
12. Write a note on Cross validation.
13. Write a note on Bootstrap.
14. What are the features of R?
15. What are the advantages of weka?
16. What are the different categories in which clustering algorithms are divided?
17. Write a note on hierarchical clustering algorithm.
18. Write difference between agglomerative and divisive clustering.
19. Write a note on correlation clustering?
20. Write a note on Graphical model?

Type of Questions: Long Answer Questions

1. Explain in detail all attribute types.
2. Explain data mining applications in detail.
3. Explain data warehouse architecture in detail.
4. Compare OLTP with OLAP systems.
5. Explain OLAP operations in detail.
6. Write a note on datawarehouse model.
7. Write a note on data pre-processing.
8. Explain Apriori algorithm in detail.
9. Write a FP tree algorithm.
10. Write algorithm for Apriori-based approach of graph mining.
11. Write algorithm for pattern growth approach of graph mining.
12. Write a note on WEKA.
13. Write a note on k-means algorithm.
14. Write a note on Text mining?
15. Write a note on Web mining?

Type of Questions: Examples and case studies

1. Suppose that a data warehouse for Big University consists of the following four dimensions: Student, Course, Semester and Instructor and two measures count and average grade. When the lower conceptual level (e.g. for a given student, course, and semester instructor combination. The average grade measure stores the actual course grade of the student. At higher conceptual levels, average grade stores the average grade for the given combination. Draw snowflake and star diagram for the data warehouse.

2. Suppose that a data warehouse of a match consists of four dimensions, date, spectator, location and game and the two measures count and charge where the charge is the fare that spectator pays when watching a game on a given date, spectators may be students, adults or seniors , with each category having its own charge rate.
   Draw a star and snowflake schema for the data warehouse.

3. Suppose that a data warehouse consists of the three dimensions time, doctor and patient and the two measures count and charge where charge is the fee that a doctor charges a patient for a visit. Enumerate and draw snowflake schema.

4. Consider the following transaction table and generate the candidate itemsets and frequent Itemsets, where the minimum support count is 2.

TID	List of items
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

   Apply Apriori algorithm to find the candidate Itemset and frequent item set.

5. Consider following transaction table and generate the candidate itemsets and frequent itemsets using Apriori algorithm where minimum support count is 2.

TID	List of items
1	Bread,Butter,Sugar
2	Bread,Butter,Milk,Sugar
3	Bread,Butter,Milk
4	Bread,Butter,Sugar
5	Butter,Sugar
6	Butter,Sugar
7	Bread,Milk
8	Butter,Milk
9	Bread,Milk

6. Consider the following transaction table and generate the candidate itemsets and frequent itemsets, where the minimum support count is 2.

TID	List of items
1	a,b,c
2	b,d
3	b,e
4	a,b,d
5	a,e,
6	b,e
7	a,e
8	a,b,e,c
9	a,b,e

7. Construct an FP tree for the following data.

TID	List of items
1	A,B,C
2	D,A,C,B
3	C,A,B
4	B,A,D
5	D
6	D,B
7	A,D,B
8	B,C

8. Construct FP tree for the following set of transactions with support count =2

TID	ITEMS
1	A,B
2	B,C,D
3	A,C,D,E
4	A,D,E
5	A,B,C
6	A,B,C,D
7	B,C
8	A,B,C
9	A,B,D
10	B,C,E

9. The following table consists of training data. Construct a decision tree based on this data using the basic algorithm of decision tree induction. Classify the records by "status" attribute.

Department	Status	Age	Salary	Count
Sales	Senior	31-35	46K-50K	30
Sales	Junior	26-30	26K-30K	40
Sales	Junior	31-35	31K-35K	40
Systems	Junior	21-25	46K-50K	20
Systems	Senior	31-35	66K-70K	5
Systems	Junior	26-30	46K-50K	3
Systems	Senior	41-45	66K-70K	3
Marketing	Senior	36-40	46K-50K	10
Marketing	Junior	31-35	41K-45K	4
Secretary	Senior	46-50	36K-40K	4
Secretary	Junior	26-30	26K-30K	6

10. The following table consists of training data. Construct a decision tree based on this data using the basic algorithm of decision tree induction. Classify the records by "Play-golf" attribute.

Outlook	Temp	humidity	Windy	Play golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes

11. Consider the database given below

Patient age	Disease	Sugar level	Survival chances
Small	Serious	High	Yes
Medium	Normal	Low	Yes
Senior	Lifetime	Normal	Yes
Small	Lifetime	High	No
Small	Normal	High	Yes
Senior	Serious	Normal	No
Medium	Serious	Low	Yes
Senior	Normal	Low	No
Medium	Lifetime	Normal	Yes
Medium	Serious	High	No
Senior	Normal	Low	No

Find out class label of the given tuple using Baysian Classification.
<Patient age: senior, Disease : Normal, Sugar level : Normal>

12. Consider the following training data set with class label buys-computer, having two distinct values {yes,no}.

TID	Age	Income	Student	Buys-computer
1	Youth	Medium	Yes	Yes
2	Youth	Low	Yes	No
3	Middle-Aged	High	No	Yes
4	Senior	High	No	No
5	Senior	Low	No	No
6	Senior	High	Yes	No
7	Youth	High	Yes	Yes
8	Middle-Aged	Medium	Yes	Yes
9	Youth	Medium	No	No
10	Youth	High	No	No

Predicate the class label for the following tuple X using naïve Bayesian classification

X=(youth, medium,yes)

13. The following data is collected about students in a class

No. of lectures attended	Marks Obtained
30	56
18	42
40	75
20	49
36	65
32	60
43	80
09	36
25	25
22	48

Using straight line regression analysis predict how many marks a student, who has attended 28 lectures will sure?

14. Cluster the following eight points (with (x,y) representing locations) into three clusters
A1(2,10), A2(2,5), A3(8,4), A4(5,8), A5(7,5), A6(6,4), A7(1,2), A8(4,9).
Initial Cluster centers are : A1(2,10),A4(5,8) and A7(1,2)
The distance between two points a=(x1,y1) and b=(x2,y2) is defined as :
ρ(a,b)=I x2-x1 I + I y2-y1 I

Use the K- mean algorithm to find

1. The three cluster centers after the first round of execution and
2. The final three clusters

15. Suppose that the data mining task is to cluster the following eight points (with (x;y) representing location) into three clusters.
A1(2;10), A2(2;5), A3(8;4), B1(5;8), B2(7;5), B3(6;4), C1(1;2), C2(4;9)
The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. Use the K-mean algorithm to show only.

1. The three cluster centers after the first round of execution and
2. The final three clusters.