DWDM Class Test - I

Date 23-04-2021

...

Points: 17/20

1. Roll Number

20

2. Name

MD JIYAUDDIN

3. The process of viewing the cross-tab (Single dimensional) with a fixed value of one attribute is
(1/1 Point)

- a) Slicing ✓
- b) Dicing
- c) Pivoting
- d) Both Slicing and Dicing

4. The data Warehouse is_____
(1/1 Point)

- (●) read only. ✓
- ( ) write only.
- ( ) read write only.
- ( ) none.

5. starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?
(1/1 Point)

- (●) a Roll-up on time from day to year. ✓
- ( ) b. Drill-down on time from day to year.
- ( ) c. Roll-up on time from year to day.
- ( ) d. Drill-down on time from year to day.

6. The type of relationship in star schema is _____.
(1/1 Point)

- ( ) A. many-to-many.
- ( ) B. one-to-one.
- (●) C. one-to-many. ✓
- ( ) D. many-to-one.

7. The operation of moving from finer-granularity data to a coarser granularity is called as
(0/1 Point)

- Rollup ✓
- b) Drill down
- c) Dicing
- d) Pivoting

8. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor for all patients?
(1/1 Point)

- a Roll-up on patients from all to individual patient.
- b. Drill-down on patients from individual patient to all.
- c. Roll-up on patients from individual patient to all. ✓
- d. Drill-down on patients from all to individual patient.

9. The data is stored, retrieved & updated in _____.
(0/1 Point)

- A. OLTP. ✓
- B. OLAP.
- C. SMTP.
- D. FTP.

10. The data found within the data warehouse is_____.
(1/1 Point)

- subject-oriented.
- time-variant.
- integrated.
- All of the above ✓

11. Dimensionality reduction reduces the data set size by removing _____.
(1/1 Point)

- relevant attributes.
- irrelevant attributes. ✓
- derived attributes.
- composite attribute

12. ..... schema supports multiple fact tables
(1/1 Point)

- A. Star schema.
- B. Snowflake schema.
- C. Fact constellation. ✓
- D. Star-snowflake schema.

13. _____databases are owned by particular departments or business groups.
(1/1 Point)

- A. Informational.
- B. Operational. ✓
- C. Both informational and operational.
- D. Flat.

14. Normalize the given data 100, 300, 450, 650, 779 with decimal scaling . Value for 300 is..
(1/1 Point)

- -1.469
- 0.1
- -0.643
- 0.3 ✓

15. Data that can be modeled as dimension attributes and measure attributes are called _____data.
(1/1 Point)

- a) dimensional
- b) Single Dimensional
- c) Measured
- d) Multidimensional ✓

16. The process of handling missing value, smoothing noise is called-------
(1/1 Point)

- data cleaning ✓
- data integration
- data reduction
- data transformation

17. Normalize the given data 100, 300, 450, 650, 779 with z-score . Value for 100 is..
(1/1 Point)

- -1.469 ✓
- 0.1
- 0.643
- 0.3

18. For bin1: 4 ,8, 15 smoothing by bin median produce the result
(1/1 Point)

- 4,8,15
- 4,4,15
- 9,9,9
- 8,8,8 ✓

19. Data transformation includes _____.
(1/1 Point)

- A process to change data from a detailed level to a summary level. ✓
- A process to change data from a summary level to a detailed level.
- joining data from one source into various sources of data.
- separating data from one source into various sources of data.

20. Fact tables are _____
(-/1 Point)

- A. completely normalized.
- B. partially demoralized.
- C. completely denormalized.
- D. partially normalized

21. ..... is an essential process where intelligent methods are applied to extract data patterns.
(1/1 Point)

- Data warehousing
- Data mining ✓
- Text mining
- Data selection

22. The star schema is composed of _____ fact table.
(1/1 Point)

- A. two
- B. one ✓
- C. three.
- D. four.

Go back to thank you page

DWDM Class Test II

...

Points: 95%

1. Roll Number

20

2. Name

MD JIYAUDDIN

3. k-NN is based on ____distance

- Manhattan distance
- Minkowski distance
- Euclidean distance ✓
- Supremum distance

4. Post={assistant, associate, professor} attribute is of type-----

- Binary
- Categorical
- Numerical
- Ordinal ✓

5. In k-itemset, k denotes

- Number of iterations
- Number of items in the set ✓
- Number of joins
- None of

6. Interval -scaled and ratio-scaled are type of

- Continuous attribute
- Ordinal attribute
- Numeric attribute ✓
- Nominal attribute

7. 4 If total transactions are 4, ‘A’ appears in 2 transactions ‘B’ appears in 3 transactions , what is the support of A?

- 100%
- 75%
- 25%
- 50% ✓

8. Which of the following is true about Manhattan distance?

- (●) It can be used for continuous variables ✓
- ( ) It can be used for categorical variables
- ( ) It can be used for categorical as well as continuous
- ( ) None of these

9. Which of the following sentences are not correct in reference to Information gain?

- (●) It is biased towards single-valued attributes ✓
- ( ) It is biased towards multi-valued attributes
- ( ) ID3 makes use of information gain
- ( ) The approach used by ID3 is greedy

10. A binary variable is _____ if both of its states are equally valuable and carry the same weight

- (●) Symmetric ✓
- ( ) Asymmetric

11. ..... is a summarization of the general characteristics or features of a target class of data.

- (●) Data Characterization ✓
- ( ) Data Classification
- ( ) Data discrimination
- ( ) Data selection

12. The value that says that transactions in D that support X also support Y is called _____.

- (●) confidence. ✓
- ( ) support.
- ( ) support count.
- ( ) None of the above.

13. Find Euclidean distance between A(23,12), B(10,34)----

- ( ) 23
- (●) 25.55 ✓
- ( ) 0
- ( ) 29

14. Which of the following is/are the Data mining tasks?

- ( ) Regression
- ( ) Classification
- ( ) Clustering
- (●) All above ✓

15. Which algorithm requires fewer scans of data?

- ( ) Apriori
- (●) FP growth ✓
- ( ) Both a and b

None of the above

16. To detect fraudulent usage of credit cards, the following data mining task should be used Select one:

- Outlier analysis ✓
- prediction
- association analysis
- feature

17. What is the approach of basic algorithm for decision tree induction?

- Greedy ✓
- Top Down
- Procedural
- Step by step

18. What does FP growth algorithm do?

- It mines all frequent patterns through pruning rules with lesser support
- It mines all frequent patterns through pruning rules with higher support
- It mines all frequent patterns by constructing a FP tree ✓
- All of the above

19. Classification is

- A subdivision of a set of examples into a number of classes.
- A measure of the accuracy, of the classification of a concept that is given by a certain

theory.

- The task of assigning a classification to a set of examples ✓
- None of these

20. Color={red, white, blue} attribute is of type-----

- Binary
- Categorical ✓
- Ordered
- Numerical

21. How do you calculate Confidence(A -> B)?

- Support(AB) / Support (A) ✓
- Support(AB) / Support (B)
- Support(AB) / Support (A)
- Support(AB) / Support (B)

22. If a set is a frequent set and no superset of this set is a frequent set, then it is called _____.

- maximal frequent set. ✓
- border set.
- infrequent sets.
- lattice.

Go back to thank you page

1) Which of the following refers to the problem of finding abstracted patterns (or structures) in the unlabeled data?

a. Supervised learning
b. Unsupervised learning
c. Hybrid learning
d. Reinforcement learning

2) Which one of the following refers to querying the unstructured textual data?

a. Information access
b. Information update
c. Information retrieval
d. Information manipulation

3) Which of the following can be considered as the correct process of Data Mining?

a. Infrastructure, Exploration, Analysis, Interpretation, Exploitation
b. Exploration, Infrastructure, Analysis, Interpretation, Exploitation
c. Exploration, Infrastructure, Interpretation, Analysis, Exploitation
d. Exploration, Infrastructure, Analysis, Exploitation, Interpretation

4) Which of the following is an essential process in which the intelligent methods are applied to extract data patterns?

a. Warehousing
b. Data Mining

c. Text Mining
d. Data Selection

5) What is KDD in data mining?

a. Knowledge Discovery Database
b. Knowledge Discovery Data
c. Knowledge Data definition
d. Knowledge data house

6) The adaptive system management refers to:

a. Science of making machine performs the task that would require intelligence when performed by humans.
b. A computational procedure that takes some values as input and produces some values as the output.
c. It uses machine learning techniques, in which programs learn from their past experience and adapt themself to new conditions or situations.
d. All of the above.

For what purpose, the analysis tools pre-compute the summaries of the huge amount of data?

a. In order to maintain consistency
b. For authentication
c. For data access
d. To obtain the queries response

8) What are the functions of Data Mining?

a. Association and correctional analysis classification
b. Prediction and characterization
c. Cluster analysis and Evolution analysis
d. All of the above

9) In the following given diagram, which type of clustering is used?

a. Hierarchal
b. Naive Bayes
c. Partitional
d. None of the above

10) Which of the following statements is incorrect about the hierarchal clustering?

a. The hierarchal type of clustering is also known as the HCA
b. The choice of an appropriate metric can influence the shape of the cluster
c. In general, the splits and merges both are determined in a greedy manner
d. All of the above

11) Which one of the following can be considered as the final output of the hierarchal type of clustering?

a. A tree which displays how the close thing are to each other
b. Assignment of each point to clusters
c. Finalize estimation of cluster centroids

d. None of the above

12) Which one of the following statements about the K-means clustering is incorrect?

a. The goal of the k-means clustering is to partition (n) observation into (k) clusters

b. K-means clustering can be defined as the method of quantization

c. The nearest neighbor is the same as the K-means

d. All of the above

13) Which of the following statements about hierarchal clustering is incorrect?

a. The hierarchal clustering can primarily be used for the aim of exploration

b. The hierarchal clustering should not be primarily used for the aim of exploration

c. Both A and B

d. None of the above

14) Which one of the clustering technique needs the merging approach?

a. Partitioned

b. Naïve Bayes

c. Hierarchical

d. Both A and C

15) The self-organizing maps can also be considered as the instance of _____ type of learning.

a. Supervised learning
b. Unsupervised learning
c. Missing data imputation
d. Both A & C

16) The following given statement can be considered as the examples of_____

Suppose one wants to predict the number of newborns according to the size of storks' population by performing supervised learning

a. Structural equation modeling
b. Clustering
c. Regression
d. Classification

17) In the example predicting the number of newborns, the final number of total newborns can be considered as the _____

a. Features
b. Observation
c. Attribute
d. Outcome

18) Which of the following statement is true about the classification?

a. It is a measure of accuracy

b. It is a subdivision of a set

c. It is the task of assigning a classification

d. None of the above

19) Which of the following statements is correct about data mining?

a. It can be referred to as the procedure of mining knowledge from data

b. Data mining can be defined as the procedure of extracting information from a set of the data

c. The procedure of data mining also involves several other processes like data cleaning, data transformation, and data integration

d. All of the above

20) In data mining, how many categories of functions are included?

a. 5

b. 4

c. 2

d. 3

21) Which of the following can be considered as the classification or mapping of a set or class with some predefined group or classes?

a. Data set

b. Data Characterization

c. Data Sub Structure

d. Data Discrimination

22) The analysis performed to uncover the interesting statistical correlation between associated -attributes value pairs are known as the _____.

a. Mining of association
b. Mining of correlation
c. Mining of clusters
d. All of the above

23) Which one of the following can be defined as the data object which does not comply with the general behavior (or the model of available data)?

a. Evaluation Analysis
b. Outliner Analysis
c. Classification
d. Prediction

24) Which one of the following statements is not correct about the data cleaning?

a. It refers to the process of data cleaning
b. It refers to the transformation of wrong data into correct data
c. It refers to correcting inconsistent data
d. All of the above

25) The classification of the data mining system involves:

a. Database technology
b. Information Science
c. Machine learning
d. All of the above

26) In order to integrate heterogeneous databases, how many types of approaches are there in the data warehousing?

a. 3
b. 4
c. 5
d. 2

27) The issues like efficiency, scalability of data mining algorithms comes under_____

a. Performance issues
b. Diverse data type issues
c. Mining methodology and user interaction
d. All of the above

28) Which of the following is the correct advantage of the Update-Driven Approach?

a. This approach provides high performance.
b. The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.
c. Both A and B

d. None of the above

28) Which of the following is the correct advantage of the Update-Driven Approach?

a. This approach provides high performance.
b. The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.
c. Both A and B
d. None of the above

Hide Answer Workspace

Answer: c

Explanation: The statements given in both A and B are the advantage of the Update-Driven Approach in Data Warehousing. So the correct answer is C.

---

29) Which of the following statements about the query tools is correct?

a. Tools developed to query the database
b. Attributes of a database table that can take only numerical values
c. Both and B
d. None of the above

Hide Answer Workspace

Answer: a

Explanation: The query tools are used to query the database. Or we can also say that these tools are generally used to get only the necessary information from the entire database.

---

30) Which one of the following correctly defines the term cluster?

a. Group of similar objects that differ significantly from other objects
b. Symbolic representation of facts or ideas from which information can potentially be extracted
c. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
d. All of the above

Hide Answer Workspace

Answer: a

Explanation: The term "cluster" refers to the set of similar objects or items that differ significantly from the other available objects. In other words, we can understand clusters as making groups of objects that contain similar characteristics form all available objects. Therefore the correct answer is A.

---

31) Which one of the following refers to the binary attribute?

a. This takes only two values. In general, these values will be 0 and 1, and they can be coded as one bit
b. The natural environment of a certain species
c. Systems that can be used without knowledge of internal operations
d. All of the above

Hide Answer Workspace

Answer: a

Explanation: In general, the binary attribute takes only two types of values, that are 0 and 1and these values can be coded as one bit. So the correct answer will be A.

---

32) Which of the following correctly refers the data selection?

a. A subject-oriented integrated time-variant non-volatile collection of data in support of management

b. The actual discovery phase of a knowledge discovery process

c. The stage of selecting the right data for a KDD process

d. All of the above

Hide Answer Workspace

Answer: c

Explanation: Data selection can be defined as the stage in which the correct data is selected for the phase of a knowledge discovery process (or KKD process). Therefore the correct answer C.

---

33) Which one of the following correctly refers to the task of the classification?

a. A measure of the accuracy, of the classification of a concept that is given by a certain theory

b. The task of assigning a classification to a set of examples

c. A subdivision of a set of examples into a number of classes

d. None of the above

Hide Answer Workspace

Answer: b

Explanation: The task of classification refers to dividing the set into subsets or in the numbers of the classes. Therefore the correct answer is C.

---

34) Which of the following correctly defines the term "Hybrid"?

a. Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.

b. Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.

c. Combining different types of method or information

d. None of these

Hide Answer Workspace

Answer: c

Explanation: The term "hybrid" refers to merging two objects and forms individual object that contains features of the combined objects.

---

35) Which of the following correctly defines the term "Discovery"?

a. It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).

b. An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.

c. It is a kind of process of executing implicit, previously unknown and potentially useful information from data

d. None of the above

Hide Answer Workspace

Answer: c

Explanation: The term "discovery" means to discover something new that has not yet been discovered. It can also be interpreted as a process of executing underlying, previously unknown and potentially useful information from data.

---

36) Euclidean distance measure is can also defined as _____

a. The process of finding a solution for a problem simply by enumerating all possible solutions according to some predefined order and then testing them

b. The distance between two points as calculated using the Pythagoras theorem

c. A stage of the KDD process in which new data is added to the existing selection.

d. All of the above

Hide Answer Workspace

Answer: c

Explanation: Euclidean distance measure can be defined as the calculating distance between two points in either in-plane or three-dimensional space measures the length of the segments connecting two points. It can also define as the distance between two points as calculated using the Pythagoras theorem.

---

37) Which one of the following can be considered as the correct application of the data mining?

a. Fraud detection

b. Corporate Analysis & Risk management
c. Management and market analysis
d. All of the above

Hide Answer Workspace

Answer: d

Explanation: Data mining is highly useful in a variety of areas such as fraud detection, corporate analysis, and risk management, and market analysis, etc., so the correct option is D.

---

38) Which one of the following correctly refers to the Class study in the data cauterization?

a. Final class
b. Study class
c. Target class
d. Both A and C

Hide Answer Workspace

Answer: c

Explanation: In the data cauterization, generally, the study class refers to the target class, and the study class is the class that is under the process of summarizing data.

---

39) Which of the following refers to the sequence of pattern that occurs frequently?

a. Frequent sub-sequence
b. Frequent sub-structure

c. Frequent sub-items
d. All of the above

Hide Answer Workspace

Answer: a

Explanation: In data mining, the frequent sub-sequence refers to a certain sequence of patterns that occurs frequently, for example, buying a camera followed by the memory card. So the correct answer will be A.

---

40) Which one of the following refers to the model regularities or to the objects that trends or not consistent with the change in time?

a. Prediction
b. Evolution analysis
c. Classification
d. Both A and B

Hide Answer Workspace

Answer: b

Explanation: In general, the evolution analysis refers to the model regularities or the object trends that vary with change in time.

---

41) The issues like "handling the rational and complex types of data" comes under which of the following category?

a. Diverse Data Type
b. Mining methodology and user interaction Issues
c. Performance issues

d. All of the above

Hide Answer Workspace

Answer: a

Explanation: It is quite often that a database can contain multiple types of data, complex objects, and temporary data, etc., so it is not possible that only one type of system can filter all data. Therefore this type of issue comes under the category Diverse Data type. So the correct answer is A.

---

42) Which of the following also used as the first step in the knowledge discovery process?

a. Data selection
b. Data cleaning
c. Data transformation
d. Data integration

Hide Answer Workspace

Answer: b

Explanation: Data cleaning is included as one of the first steps of the knowledge discovery process. So the correct answer is B.

---

43) Which of the following refers to the steps of the knowledge discovery process, in which the several data sources are combined?

a. Data selection
b. Data cleaning
c. Data transformation
d. Data integration

Hide Answer Workspace

Answer: d

Explanation: The step "data integration" of the knowledge discovery process refers to combining several data sources. Therefore the correct answer is D.

---

44) Which of the following can be considered as the drawback of the query-Driven approach in data warehousing?

a. This approach is expensive for queries that require aggregations
b. This approach is expensive insufficient, and very frequent queries
c. This approach requires a very complex integration and filtering process
d. All of the above

Hide Answer Workspace

Answer: d

Explanation: All statements given in the above question are drawbacks of the query-driven approach. Therefore the correct answer is D.

---

45) Which of the following correctly refers to the term "Data Independence"?

a. It means that the programs are not dependent on the logical attributes
b. It refers to that data that is defined separately, not included in the program

c. It means that the programs are totally dependent on the physical attributes of data
d. Both A and C

Hide Answer Workspace

Answer: d

Explanation: The term "Data Independence" refers that the programs are not dependent on the physical attributes of data and neither on the logical attributes of data.

---

46) Which of the following is generally used by the E-R model to represent the weak entities?

a. Diamond
b. Doubly outlined rectangle
c. Dotted rectangle
d. Both B & C

Hide Answer Workspace

Answer: b

Explanation: Generally, the double outline rectangle is used in the E-R model to represent the weak entities.

---

47) Which one of the following refers to the Black Box?

a. It can be referred as the system that can be used without the knowledge of the internal operations
b. It referrers the natural environment of the specific species
c. It takes only two values at most that are 0 and 1

d. All of the above

Hide Answer Workspace

Answer: a

Explanation: Black Box is referred to as the system which takes only two values at most are zero and one.

---

48) Which one of the following issues must be considered before investing in data mining?

a. Compatibility
b. Functionality
c. Vendor consideration
d. All of the above

Hide Answer Workspace

Answer: d

Explanation: The common but important issues like functionality and compatibility must always be discussed before investing in data mining. Therefore the correct answer is D.

---

49) The term "DMQL" stands for _____

a. Data Marts Query Language
b. DBMiner Query Language
c. Data Mining Query Language
d. None of the above

Hide Answer Workspace

Answer: c

Explanation: The term "DMQL" refers to the Data Mining Query Language. Therefore the correct answer is C.

---

50) In certain cases, it is not clear what kind of pattern need to find, data mining should_____:

a. Try to perform all possible tasks
b. Perform both predictive and descriptive task
c. It may allow interaction with the user so that he can guide the mining process
d. All of the above

Hide Answer Workspace

Answer: c

7. To integrate heterogeneous databases, how many approaches are there in Data Warehousing?

1. 1
2. 2
3. 3
4. 4

Show Answer
2

8. Which of the following is correct advantage of Update-Driven Approach in Data Warehousing?

A. This approach provides high performance.
B. The data can be copied, processed, integrated, annotated, summarized and

restructured in the semantic data store in advance.
C. Both A and B
D. None Of the above

Show Answer
Both A and B

9. What is the use of data cleaning?
A. to remove the noisy data
B. correct the inconsistencies in data
C. transformations to correct the wrong data.
D. All of the above

Show Answer
All of the above

10. Data Mining System Classification consists of?
A. Database Technology
B. Machine Learning
C. Information Science
D. All of the above

Show Answer
All of the above

Data mining and warehousing mcq sppu
11. Which of the following is a good alternative to the star schema?

1. snow flake schema
2. star schema
3. star snow flake schema
4. fact constellation

Show Answer
fact constellation

12. Patterns that can be discovered from a given database are which type…

1. More than one type
2. Multiple type always
3. One type only
4. No specific type

Show Answer

More than one type

13. Background knowledge is…

1. It is a form of automatic learning.
2. A neural network that makes use of a hidden layer
3. The additional acquaintance used by a learning algorithm to facilitate the learning process
4. None of these

Show Answer

The additional acquaintance used by a learning algorithm to facilitate the learning process

14. Which of the following is true for Classification?

1. subdivision of a set
2. A measure of the accuracy
3. The task of assigning a classification
4. All of these

Show Answer

subdivision of a set

Data mining and Warehousing mcq

15. Data mining is?

1. time variant non-volatile collection of data
2. The actual discovery phase of a knowledge
3. The stage of selecting the right data
4. None of these

Show Answer
The actual discovery phase of a knowledge

16. ——-- is not a data mining functionality?
A) Clustering and Analysis
B) Selection and interpretation
C) Classification and regression
D) Characterization and Discrimination

Show Answer
Selection and interpretation

17. Which of the following can also applied to other forms?
a) Data streams & Sequence data
b) Networked data
c) Text & Spatial data
d) All of these

Show Answer
All of these

18. ——— is the out put of KDD
a) Query
b) Useful Information
c) Data
d) information

Show Answer
Useful Information

19. What is noise?
a) component of a network
b) context of KDD and data mining

c) aspects of a data warehouse
d) None of these

Show Answer
context of KDD and data mining

data mining and warehousing mcq sppu
20. Firms that are engaged in sentiment mining are analyzing data collected from?
A. social media sites.
B. in-depth interviews.
C. focus groups.
D. experiments.

Show Answer
social media sites.

21. Which of the following forms of data mining assigns records to one of a
predefined set of classes?
(A). Classification
(B). Clustering
(C). Both A and B
(D). None

Show Answer
Clustering

22. The learning which is used to find the hidden pattern in unlabeled data is called?
(A). Unsupervised learning
(B). Supervised learning
(C). Reinforcement learning

Show Answer

Unsupervised learning

23. The learning which is the example of Self-organizing maps?

(A). Reinforcement learning
(B). Supervised learning
(C). Unsupervised learning
(D). Missing data imputation

Show Answer
Unsupervised learning

24. According to storks' population size, find the total number of babies from the following example of predicting the number of babies.

(A). feature
(B). outcome
(C). attribute
(D). observation

Show Answer
outcome

25. Which of the following is not belong to data mining?

(A). Knowledge extraction
(B). Data transformation
(C). Data exploration
(D). Data archaeology

Show Answer
Data archaeology

26. The learning which is used for inferring a model from labeled training data is called?

(A). Unsupervised learning
(B). Reinforcement learning
(C). Supervised learning
(D). Missing data imputation

Show Answer
Supervised learning

27. Which of the following is the right approach to Data Mining?

(A). Infrastructure, exploration, analysis, exploitation, interpretation
(B). Infrastructure, exploration, analysis, interpretation, exploitation
(C). Infrastructure, analysis, exploration, interpretation, exploitation
(D). None of these

Show Answer
Infrastructure, exploration, analysis, interpretation, exploitation

28. Which of the following terms is used as a synonym for data mining?

(A). knowledge discovery in databases
(B). data warehousing
(C). regression analysis
(D). parallel processing in databases

Show Answer
knowledge discovery in databases

29. ……..is an essential process where intelligent methods are applied to extract data patterns

A) Data Warehousing
B) Data Mining
C) Data Base
D) Data Structure

Show Answer
Data Mining

30. Data mining requires

1. Large quantities of operational data stored over a period of time
2. Lots of tactical data
3. Several tape drives to store archival data
4. Large mainframe computers

Show Answer
Large quantities of operational data stored over a period of time

data mining and warehousing mcq questions

31. Data by itself is not useful unless

1. It is massive
2. It is processed to obtain information
3. It is collected as a raw data from diverse sources
4. It is properly stated

Show Answer
It is processed to obtain information

32. Which of the following is NOT example of ordinal attributes?

1. Zip codes
2. Ordered numbers
3. Ascending or descending names
4. Military ranks

Show Answer
Zip codes

33. In asymmetric attribute

1. Order of values is important

2. All values are equals
3. Only non-zero value is important
4. Range of values is important

Show Answer

Only non-zero value is important

34. Identify the example of Nominal attribute

1. Temperature
2. Mass
3. Salary
4. Gender

Show Answer

Gender

35. Which of the following is not a data pre-processing methods?

1. Data Visualization
2. Data Discretization
3. Data Cleaning
4. Data Reduction

Show Answer

Data Visualization

36. Correlation analysis is used for __

1. Handling missing values
2. Identifying redundant attributes
3. Handling different data formats
4. Eliminating noise

Show Answer

Identifying redundant attributes

37. _____combines data from multiple sources into a coherent store

1. Data Characterization
2. Data Classification
3. Data Integration

4. Data Selection

Show Answer

Data Integration

38. Which of the following is / are attribute subset selection criterion(s) ?

1. Forward selection
2. Backward elimination
3. Decision tree induction
4. All of the above

Show Answer

All of the above

39. Data mining can also applied to other forms such as……..

i) Data streams
ii) Sequence data
iii) Networked data
iv) Text data
v) Spatial data
A)i, ii, iii and v only
B) ii, iii, iv and v only
C) i, iii, iv and v only
D) All i, ii, iii, iv and v

Show Answer

All i, ii, iii, iv and v

40. ____ normalization is not very well efficient in handling the outliers Min max

1. Min max
2. Z Score
3. Decimal Scaling
4. None of the above

Show Answer

Min max

Data mining and warehousing mcq with answers

41. The full form of KDD is……

A) Knowledge Database
B) Knowledge Discovery Database
C) Knowledge Data House
D) Knowledge Data Definition

Show Answer
Knowledge Discovery Database

Data Analytics sppu mcq

42. A collection of interesting and useful patterns in database is called __

A. knowledge.
B. information.
C. data.
D. algorithm

Show Answer
knowledge.

43. Data ……. is the process of finding a model that describes and
distinguishes data classes or concepts.

a)Characterization
b)Mining
c) clustering
d)Classification

Show Answer
Classification

44. To remove noise and inconsistent data __ is needed

1. Data Transformation
2. Data Reduction
3. Data Integration
4. Data Cleaning

Show Answer

Data Cleaning

45.The terms equality and roll up are associated with _

1. OLTP
2. Visualization
3. Data mart
4. Decision Tree

Show Answer

Data mart

46. An operational system is which of the following?

A. A system that is used to run the business in real time and is based on historical data.

B. A system that is used to run the business in real time and is based on current
data.

C. A system that is used to support decision making and is based on current data.

D. A system that is used to support decision making and is based on historical data.

Show Answer

A system that is used to run the business in real time and is based on current data.

47. Data warehouse is which of the following?

A. Can be updated by end users.

B. Contains numerous naming conventions and formats.

C. Organized around important subject areas.
D. Contains only current data.

Show Answer
Organized around important subject areas.

48. Data transformation includes which of the following?
A. A process to change data from a detailed level to a summary level
B. A process to change data from a summary level to a detailed level
C. Joining data from one source into various sources of data
D. Separating data from one source into various sources of data

Show Answer
A process to change data from a detailed level to a summary level

49. The ..... allows the selection of the relevant information necessary for
the data warehouse.
A top-down view
B data warehouse view
C data source view
D business query view

Show Answer
A top-down view

50. Which of the following is not a component of a data warehouse?
A Metadata
B Current detail data
C Lightly summarized data
D Component Key

Show Answer
Component Key

51. Which of the following is not a kind of data warehouse application?

A Information processing
B Analytical processing
C Data mining
D Transaction processing

Show Answer
Transaction processing

52. ___ is not associated with data cleaning process.

1. Deduplication
2. Domain consistency
3. Segmentation
4. Disambiguation

Show Answer
Segmentation

53. Dimensionality refers to

1. Cardinality of key values in a star schema
2. The data that describes the transactions in the fact table
3. The level of detail of data that is held in the fact table
4. The level of detail of data that is held in the dimension table

Show Answer
The data that describes the transactions in the fact table

54. Expansion for DSS in DW is

1. Decisive Strategic System
2. Data Support System
3. Data Store System
4. Decision Support system

Show Answer
Decision Support system

55. Data in a data warehouse

1. in a flat file format
2. can be normalised but often is not
3. must be in normalised form to at least 3NF
4. must be in normalised form to at least 2NF

Show Answer

can be normalised but often is not

56. Friendship structure of users in a social networking site can be considered as
an example of ____

1. Record data
2. Ordered data
3. Graph data
4. None of the above

Show Answer

Graph data

57. A café owner wanted to compare how much revenue he gained from lattes
across different months of the year. What type of variable is 'month'?

1. Continuous
2. Categorical
3. Discrete
4. Nominal

Show Answer

Categorical

58. An outlier is a _

1. Description of records in the data
2. Data point which is considered different from other data points
3. Record with missing attributes
4. Duplicate record

Show Answer

Data point which is considered different from other data points

59. Which of the following operations can be performed on ordinal attributes?

1. Distictness
2. Documents
3. Both of the above
4. None of the above

Show Answer

Both of the above

60. Height of a person, can be considered as an attribute of _____type?

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Show Answer

Ratio

61. The cosine similarity measure counts for _

1. The Euclidian distance between vectors
2. The Manhattan distance between vectors
3. The similarity of documents
4. The dissimilarity of vectors

Show Answer

62. Formula for dissimilarity computation between two objects for categorical
variable is – here p is categorical variable and m denotes number of matches

1. D (i, j) = p – m / p
2. D (i, j) = p – m / m
3. D (i, j) = m – p / p
4. D (i, j) = m – p / m

Show Answer

D (i, j) = p – m / p

63. Euclidean and Manhattan distances between the objects P, Q and R (1, 2, 3) and
(2, 1, 0) are _

1. 3.32, 4 respectively
2. 3.32, 5 respectively
3. 5, 3.32 respectively
4. 3.30, 3 respectively

Show Answer
3.32, 5 respectively

64. The main organisational justification for implementing a data warehouse is to
provide

1. ETL from operation systems to strategic systems
2. Large scale transaction processing
3. Storing large volumes of data
4. Decision support

Show Answer
Decision support

65. A data warehouse
a. must import data from transactional systems whenever significant changes occur in the
transactional data
b. works on live transactional data to provide up to date and valid results
c. takes regular copies of transaction data
d. takes preprocessed transaction data and stores in a way that is optimised for
analysis

Show Answer
takes preprocessed transaction data and stores in a way that is optimised for analysis

66. Data warehouse contains _____data that is seldom found in the operational environment

1. informational
2. normalized
3. denormalized
4. summary

Show Answer
summary

67. In a snowflake schema which of the following types of tables is considered?

1. Fact
2. Dimension
3. Both (a) and (b)
4. None of the above

Show Answer
Both (a) and (b)

68. Which of the following statements about data warehouse is true?

1. A data warehouse is necessary to all those organisations that are using relational OLTP
2. A data warehouse is useful to all organisations that currently use OLTP
3. A data warehouse is valuable to the organisations that need to keep an audit trail of their activities
4. A data warehouse is valuable only if the organisation has an interest in analysing historical data

Show Answer
A data warehouse is valuable only if the organisation has an interest in analysing historical data

69. When you ____ the data, you are aggregating the data to a higher level

1. Slice

2. Roll Up
3. Roll Down
4. Drill Down

Show Answer
Roll Up

70. The process of viewing the cross-tab (Single dimensional) with a fixed value of one attribute is _

1. Slicing
2. Dicing
3. Pivoting
4. Both Slicing and Dicing

Show Answer
Slicing

71. What do data warehouses support?

1. OLAP
2. OLTP
3. OLAP and OLTP
4. Operational databases

Show Answer
OLAP

72. A data cube consist of _

1. Dimensional data
2. Multidimensional data
3. No dimensional data
4. 1 dimensional data

Show Answer
Multidimensional data

73. Which type of data storage architecture gives fastest performance?

1. ROLAP

2. MOLAP
3. HOLAP
4. DOLAP

Show Answer
MOLAP

74. Dissimilarity can be defined as __

1. How much certain objects differ from each other
2. How much certain objects simillar from each other
3. Dissimilarities are non negative numbers d(i,j) that are small when i and j are close to each other and that become large when i and j are very different
4. Both (a) and (c)

Show Answer
Both (a) and (c)

75. _____supports basic OLAP operations, including slice and dice, drill-down,
roll-up and pivoting

1. Information processing
2. Analytical processing
3. Data processing
4. Transaction processing

Show Answer
Analytical processing

1. What is true about data mining?

A. Data Mining is defined as the procedure of extracting information from huge sets of data
B. Data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation
C. Data mining is the procedure of mining knowledge from data.
D. All of the above

View Answer

Ans : D

Explanation: Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so that it can be used.

2. How many categories of functions involved in Data Mining?

A. 2
B. 3
C. 4
D. 5

View Answer

Ans : A

Explanation: there are two categories of functions involved in Data Mining : 1. Descriptive, 2. Classification and Prediction

3. The mapping or classification of a class with some predefined group or class is known as?

A. Data Characterization
B. Data Discrimination
C. Data Set
D. Data Sub Structure

View Answer

Ans : B

Explanation: Data Discrimination : It refers to the mapping or classification of a class with some predefined group or class

4. The analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs is called?

A. Mining of Association
B. Mining of Clusters
C. Mining of Correlations
D. None of the above

View Answer

Ans : C

Explanation: Mining of Correlations : It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

5. _____ may be defined as the data objects that do not comply with the general behavior or model of the data available.

A. Outlier Analysis
B. Evolution Analysis
C. Prediction
D. Classification

View Answer

Ans : A

Explanation: Outlier Analysis : Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.

6. "Efficiency and scalability of data mining algorithms" issues comes under?

A. Mining Methodology and User Interaction Issues
B. Performance Issues
C. Diverse Data Types Issues
D. None of the above

View Answer

Ans : B

Explanation: In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

7. To integrate heterogeneous databases, how many approaches are there in Data Warehousing?

A. 2
B. 3
C. 4
D. 5

View Answer

Ans : A

Explanation: Data warehousing involves data cleaning, data integration, and data consolidations. To integrate heterogeneous databases, we have the following two approaches : Query Driven Approach, Update Driven Approach

8. Which of the following is correct advantage of Update-Driven Approach in Data Warehousing?

A. This approach provides high performance.
B. The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.
C. Both A and B
D. None Of the above

View Answer

Ans : C

Explanation: Both A and B are advantage of Update-Driven Approach in Data Warehousing.

9. What is the use of data cleaning?

A. to remove the noisy data
B. correct the inconsistencies in data
C. transformations to correct the wrong data.
D. All of the above

View Answer

Ans : D

Explanation: Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data preprocessing step while preparing the data for a data warehouse.

10. Data Mining System Classification consists of?

A. Database Technology
B. Machine Learning
C. Information Science
D. All of the above

View Answer

Ans : D

11. Which of the following is correct application of data mining?

A. Market Analysis and Management
B. Corporate Analysis & Risk Management
C. Fraud Detection
D. All of the above

View Answer

Ans : D

Explanation: Data mining is highly useful in the following domains : Market Analysis and Management, Corporate Analysis & Risk Management, Fraud Detection

12. In Data Characterization, class under study is called as?

A. Study Class
B. Intial Class
C. Target Class
D. Final Class

View Answer

Ans : C

Explanation: Data Characterization : This refers to summarizing data of class under study. This class under study is called as Target Class.

13. A sequence of patterns that occur frequently is known as?

A. Frequent Item Set
B. Frequent Subsequence
C. Frequent Sub Structure
D. All of the above

View Answer

Ans : B

Explanation: Frequent Subsequence : A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.

14. _____ refers to the description and model regularities or trends for objects whose behavior changes over time.

A. Outlier Analysis
B. Evolution Analysis
C. Prediction
D. Classification

View Answer

Ans : B

Explanation: Evolution Analysis : Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

15. Pattern evaluation issue comes under?

A. Mining Methodology and User Interaction Issues
B. Performance Issues
C. Diverse Data Types Issues
D. None of the above

View Answer

Ans : A

Explanation: Pattern evaluation : The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

16. "Handling of relational and complex types of data" issue comes under?

A. Mining Methodology and User Interaction Issues
B. Performance Issues
C. Diverse Data Types Issues
D. None of the above

View Answer

Ans : C

Explanation: The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.

17. Which of the following is correct disadvantage of Query-Driven Approach in Data Warehousing?

A. The Query Driven Approach needs complex integration and filtering processes.
B. It is very inefficient and very expensive for frequent queries.
C. This approach is expensive for queries that require aggregations.
D. All of the above

View Answer

Ans : D

Explanation: All statement are disadvantage of Query-Driven Approach in Data Warehousing.

18. The first steps involved in the knowledge discovery is?

A. Data Integration
B. Data Selection
C. Data Transformation
D. Data Cleaning

View Answer

Ans : D

Explanation: The first steps involved in the knowledge discovery is Data Integration.

19. In which step of Knowledge Discovery, multiple data sources are combined?

A. Data Cleaning
B. Data Integration
C. Data Selection
D. Data Transformation

View Answer

Ans : B

Explanation: Data Integration : multiple data sources are combined.

20. DMQL stands for?

A. Data Mining Query Language
B. Dataset Mining Query Language
C. DBMiner Query Language
D. Data Marts Query Language

View Answer
Ans : A

1. A priori algorithm operates in ___ method
a. Bottom-up search method
b. Breadth-first search method
c. None of above
d. Both a & b

2. A bi-directional search takes advantage of ___ process
a. Bottom-up process
b. Top-down process
c. None
d. Both a & b

3. The pincer-search has an advantage over a priori algorithm when the largest frequent item set is long.
a. True
b. false

4. MCFS stand for
a. Maximum Frequent Candidate Set
b. Minimal Frequent Candidate Set
c. None of above

5. MFCS helps in pruning the candidate set
a. True
b. False

6. DIC algorithm stands for ___
a. Dynamic itemset counting algorithm
b. Dynamic itself counting algorithm
c. Dynamic item set countless algorithms
d. None of above

7. If the item set is in a dashed circle while completing a full pass it moves towards
a. Dashed circle
b. Dashed box

c. Solid Box
d. Solid circle

8. If the item set is in the dashed box then it moves into a solid box after completing a full pass
a. True
b. False

9. The dashed arrow indicates the movement of the item set
a. True
b. False

10. The vertical arrow indicates the movement of the item set after reaching the frequency threshold
a. True
b. False

11. Frequent set properties are:
a. Downward closure property
b. Upward closure property
c. A & B
d. None of these

12. Any subset of a frequent set is a frequent set is
A. Downward closure property
B. Upward closure property
C. A and b

13. Periodic maintenance of a data mart means
a. Loading
b. Refreshing
c. Purging
d. All are true

14. The Fp-tree Growth algorithm was proposed by
a. Srikant
b. Aggrawal
c. Hanetal
d. None of these

15. The main idea of the algorithm is to maintain a frequent pattern tree of the date set. An extended prefix tree structure starting crucial and quantitative information about frequent sets
a. Priori Algorithm
b. Pinchers Algorithm
c. FP- Tree Growth algo.
d. All of these

16. The data warehousing and data mining technologies have extensive potential applications in the govt in various central govt sectors such as :
a. Agriculture
b. Rural Development
c. Health and Energy
d. all of the true

17. ODS Stands for
a. External operational data sources
b. operational data source
c. output data source
d. none of the above

18. Good performance can be achieved in a data mart environment by extensive use of
a. Indexes
b. creating profile records
c. volumes of data
d. all of the above

19. Features of Fp tree are
(i). It is dependent on the support threshold
(ii). It depends on the ordering of the items
(iii). It depends on the different values of trees
(iv). It depends on frequent itemsets with respect to give information
a. (i) & (ii)
b. (iii) & (iv)
c. (i) & (iii)
d. (ii) only

20. For a list T, we denote head_t as its first element and body-t as the remaining part of the list (the portion of the list T often removal of head_t) thus t is
a. {head} {body}

b. {head_t} {body_t}
c. {t_head}{t_body}
d. None of these

21. Partition Algorithm executes in
a. One phase
b. Two-Phase
c. Three phase
d. None of these

22. In the First Phase of the Partition Algorithm
a. Logically divides into a number of non-overlapping partitions
b. Logically divides into a number of overlapping Partitions
c. Not divides into partitions
d. Divides into non-logically and non-overlapping Partitions

23. Functions of the second phase of the partition algorithm are
a. Actual support of item sets are generated
b. Frequent itemsets are identified
c. Both (a) & (b)
d. None of these

24. Partition algorithm is based on the
a. Size of the global Candidate set
b. Size of the local Candidate set
c. Size of frequent itemsets
d. No. Of item sets

25. Pincer search algorithm based on the principle of
a. Bottom-up
b. Top-Down
c. Directional
d. Bi-Directional

26. Pincer-Search Method Algorithm contains
(i) Frequent item set in a bottom-up manner
(ii) Recovery procedure to recover candidates
(iii) List of maximal frequent itemsets
(iv) Generate a number of partitions
a. (i) only

b. (i) & (iii) only
c. (i),(iii) & (iv)
d. (i),(ii)&(iii)

27. Is a full-breadth search, where no background knowledge of frequent itemsets is used for pruning?
a. Level-crises filtering by the single item
b. Level-by-level independent
c. Multi-level mining with uniform support
d. Multi-level mining with reduced support

28. Disadvantage of uniform support is
a. Items at lower levels of abstraction will occur as frequently.
b. If the minimum support threshold is set too high, I could miss several meaningful associations
c. Both (a) & (b)
d. None of these

29.Warehouse administrator responsible for
a. Administrator
b. maintenance
c. both a and b
d. none of the above

30. The pincer-search has an advantage over a priori algorithm when the largest frequent itemset is long
a. True
b. false

31. What are the common approaches to tree pruning?
a. Prepruning and Postpruning approach.
b. Prepruning.
c. Postpruning.
d. None of the above.

32. Tree pruning methods address this problem of ___?
a. Overfitting the branches
b. Overfitting the data
c. a and b both
d. None of the above

33. What is the Full Form of MDL.
a. Maximum Description Length
b. Minimum Description Length
c. Mean Described Length
d. Minimum Described Length

34. State that the Statements are True / False:
a. Post pruning approach Removes Branches from a 'Fully Grown' Tree.
a. True
b. False

b. The "Best Pruned Tree is the one that maximizes the number of encoding bits.
a. True
b. False

35. Upon halting, the node becomes a ___
A. Heap
B. Subset
C. Leaf
D. Superset

36. demographic and neural clustering are methods of clustering based on
a. data types
b. methodology of calculation
c. Inter record distance
d. all of the above

37. POS stands for
a. Peer of sale
b. Point of sale
c. part of the sale
d. none of the above

38. Classification and Prediction are two forms of
a. Data analysis
b. Decision Tree
c. A and B
d. None of these

39. Classification predicts
a. Categorical labels

b. Prediction models continued valued function
c. A and B
d. None of these

40. True / False
a. Each Tuple is assumed to belong to a predefined class as determined by one of the attributes, called the class label attribute.
b. The individual tuples making up the training set are referred to as the training data set.
c. Classification and Regression are the two major type of data analysis.
Ans. A-True, B-True, C-False

41.True / False
a. Classification and Regression are the two major type of data analysis.
b. Classification is used to predict discrete or nominal values.
c. Regression is used to predict continuous or ordered values.
d. All are true

42. Classification and Prediction have numerous applications:
a. Credit approval
b. Medical diagnosis
c. Performance prediction & selective marketing
d. All of these

43. Class label of each training sample is provided with this step is known as
a. Unsupervised learning
b. Supervised learning
c. Training samples
d. Clustering

44. Decision tree is based on
a. Bottom-down technique
b. Top-down technique
c. Divide-and-conquer manner
d. Top-down recursive divide-and-conquer manner

45. Recursive Partitioning stops in Decision Tree when
a. All samples for a given node belong to the same class.
b. There are no remaining attributes on which samples may be further partitioned.

c. There are no samples for the branch test.
d. All the above.

46. To select the test attribute of each node in a decision tree we use
a. Entity Selection Measure
b. Data Selection Measure
c. Information Gain Measure
d. None of these

47.Test attribute for the current node in the decision tree is chosen on the basis of
a. Lowest entity gain
b. Highest data gain
c. Highest Information Gain
d. Lowest Attribute Gain

48. Advantage of the Information-theoretic approach of the decision tree is
a. Minimizes the expected number of tests needed
b. Minimizes the number of Nodes
c. Maximizes the number of nodes
d. Maximizes the number of tests

49. Let us be the no. of samples of S in class Ci then expected information to classify a given sample is given by
a. L(s1,s2,.....sm)=_log2(pi)
b. L(s1,s2,.....sm)=-_pilog2(pi)
c. L(s1,s2,.....sm)=_pilog2x
d. L(s1,s2,.....sm)=_pilog2(pi)

50. Steps applied to the data in order to improve the accuracy, efficiency, and scalability are:-
a. Data cleaning
b. Relevance analysis
c. Data transformation
d. All of the above

51. The process used to remove or reduce noise and the treatment of missing values
a. Data cleaning
b. Relevance analysis
c. Data transformation
d. None of above

52. Relevance analysis may be performed on the data by removing any irrelevant attribute from the process.

a. True

b. False

53. Classification and prediction method can be affected by:-

a. Accuracy & Speed

b. Robustness & Scalability

c. Interpretability

d. All of the above

54. In a decision tree internal node denotes a test on an attribute and Leaf nodes represent classes or class distributions

a. True

b. false

55. ___ attempts to identify and remove branches, with Improving accuracy

a. decision tree

b. tree pruning

c. both of them

d. none of above

56. To deal with larger data sets, a sampling method, called ___

a. Clara

b. Dara

c. Pam

d. None

57. What is the Full Form of CLARA.

a. Clustering Large Applicant

b. Close Large Applicant

c. Clustering Large Applications

d. None of the above

58. What is the Full Form of CLARANS.

a. Clustering Large Applications Based Upon Randomized Search

b. Close Large Applicant Based Upon Role Search

c. Clustering Large Applicant Based Upon Randomized Search

d. None of the above

59. Which Algorithm was proposed that combines the Sapling Technique with PAM.
a. CLARA
b. CLARANS
c. Both a and b
d. None of these.

60. Which are the two type of Hierarchical Clustering?
a. Agglomerative Hierarchical Clustering and Density Hierarchical Clustering
b. Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering
c. Divisive Hierarchical Clustering and Density Hierarchical Clustering
d. None of the above

61. Cluster is a :
a. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.
b. A cluster of data objects can be treated collectively as one group in many applications
c. Cluster analysis is an important human activity.
d. All of the above

62. Cluster analysis tools based on
a. K-means
b. K-medosis
c. A and B
d. None of these

63. S-Plus, SPSS, SAS software packages use for
a. Data Mining
b. Classification
c. Clustering
d. Prediction

64. Unsupervised learning is an example of
a. Classification and prediction
b. Classification and Regression
c. clustering
d. Data Mining

65. Requirement of Clustering in Data Mining
a. Scalability
b. Ability to deal with different types of attributes

c. Ability to deal with noisy data
d. Discovery of clusters with arbitrary shape
e. Minimal requirement for domain knowledge to determine input parameters
f. Insensitivity to the order of input records
g. High dimensionality
h. Constraint-based clustering
(a). a, c, d, f
(b). g, h
(c). All of these
(d.) None of these

66. Clustering method can be classified
a. Partitioning Methods
b. Hierarchical methods
c. Density-based methods
d. All of these

67. Hierarchical methods can be classified
a. Agglomerative Approach
b. Divisive Approach
c. A and B
d. None of these

68. Agglomerative approach is called as
a. Bottom-up Approach
b. Top-Down Approach
c. A and B
d. None of these

69. Top-Down Approach is
a. Agglomerative Approach
b. Divisive Approach

70. Drawback of Hierarchical Methods
a. Suffer from the fact that once a step is done, it can never be undone.
b. A technique is that they cannot correct erroneous decision.
c. Both a & b
d. None of these

71. Two approaches to improving the quality of hierarchical clustering:
a. Perform careful analysis of object "linkages" at each hierarchical partitioning, such as in CURE and Chameleon
b. Integrate Hierarchical agglomeration and iterative relocation by first using a hierarchical agglomerative algorithm and refining the result using an iterative relocation
c. Both a & b
d. None of these

72. Classical Portioning methods are
a. k-means and k-median
b. k-means and k-medoids
c. k-modes only
d. none of these

73. K-means technique is based on
a. Centroid Object
b. Reference object
c. Representative object
d. Partition Object

74. K-medoids technique is based on
a. Centroid Object
b. Representative object
c. Partition Object
d. None of these

75. The k-means and the k-modes methods can be integrated to cluster data with mixed numeric and categorical values, resulting in
a. k-median method
b. k-partition method
c. k-prototypes method
d. k-medoids method

76. The squared-error criterion is used in a k-means method defined as
a. E=_I=1tok _pεci [p-m_i]
b. E=_I=1tok _pεci [m_i]$_2$
c. E=_I=1tok _pεci [p]$_2$
d. E=_I=1tok _pεci [p-mi]$_2$

77. The Computational Complexity of the k-means method algorithm is
a. O(log x)
b. Θ(nkt)
c. O(nkt)
d. Θ(log x)

78. Which Method is more Robust-k-means or k-medoids?
a. The k means is more robust in the presence of noise
b. The k-medoids method is more robust in the presence of noise and outliers
c. The k-medoids method is more robust due to no. of partitions
d. The k means is more robust due to its less complexity

79. First k-medoids algorithm introduced is
a. Prototype Above Medoids
b. Partition Below Medoids
c. Prototype Around Medoids
d. Partitioning Around Medoids

80. PAM stands for
a. Prototype Above Medoids
b. Prototype Around Means
c. Partitioning Around Medoids
d. Partitioning Above Means

81.Which statements are true fork-means
(i). It can apply only when the mean of the cluster is defined.
(ii). It is not suitable for discovering clusters with non-convex shapes
(iii). This method is relatively efficient in processing only small data.
a. (i) only
b. (i) & (ii) only
c. (iii) only
d. All the above

82 DBSCAN stands for:
a. Divisive Based Clustering Method
b. Density-Based Clustering Method
c. Both a & b
d. None of above

83: DBSCAN defines a cluster as a maximal set of density –
Connected points
a. True
b. False

84: For a non-negative value $\varepsilon, N_e(O_i)=\{ O_j \in D \mid d(O_i,O_j) \leq \varepsilon\}$
a. True
b. false

85. The ___ client is a desktop that relies on the server to which it is connected for the majority of its computing power.
a. thin
b. none
c. thick
d. web server

86. An object is said to be the Core Object if
a._ $N_e(O)$_ ≥ MinPts
b. _ N (O)_ ≥ MaxPts
c. none of above
d. both a & b

87. The density-reachability relation is transitive but not symmetric
a. True
b. False

88. Non-core objects are:-
a. border object
b. noise object
c. non-object
d. both a & b

89. DBSCAN algorithm can classify into:
a. classified
b. unclassified
c. noise
d. all of above

90. Unsupervised learning is an example of
a. Classification and prediction
b. Classification and Regression

c. clustering
d. Data Mining

91. Data can be classified as
a. reference data
b. transaction data and derived data
c. derived data
d. all of the above

92. Reference and transaction data originates from
a. operational system
b. Unnormalized data
c. data marts
d. all are true

93. Derived data is derived from
a. reference data
b. transaction data
c. reference and transaction data
d. none of the above

94. Unnormalized data, which is the basis for online analytical processing tools are prepared periodically but is directly based on detailed ___.
a. reference data
b. transaction data
c. reference and transaction data
d. none of the above

95. The data mart is loaded with data from a data warehouse by means of a ___
a. load program
b. process
c. project
d. all is valid

96. The chief considerations for a Load program are:
a. frequency and schedule
b. total or partial refreshment
c. customization and re-sequencing
d. all are true

97. Periodic maintenance of a data mart means
a. all are true
b. loading
c. refreshing
d. purging

98. Detailed level data, summary level, preprocessed and Adhoc data are data in
a. data warehouse
b. data mart
c. both
d. none of the above

99. Data sources in the data warehouse are referred to as
a. External data source
b. Operational data source
c. External operational data source
d. none of the above

100. ___ Table help and enable the end-users of the data mart to relate the data to its expanded version.
a. data
b. reference
c. both a and b
d. none of the above

DWDM End Term

1. Roll Number

29

2. Name

Yashdeep Prakash Vaitage

3. In given confusion Matrix, With respect to apple class, calculate true positive?

	True Class		
	Apple	Orange	Mango
Apple	7	8	9
Orange	1	2	3
Mango	3	2	1

(1 Point)

Oa

Ci) 7

Ü9

U2

4. In given confusion Matrix (Q.3), With respect to apple class, calculate true negative?
(1 Point)

@s

07

09

02

S. In given confusion Matrix (Q.3), With respect to apple class, calculate False negative?
(1 Point)

Os

07

0 17

@4

6. In given confusion Matrix (Q.3), With respect to apple class, calculate False Positive?
(1 Point)

Os

07

@ 17

◯ 4

7. In_____ un-labeled training data is used for the learning
(1 Point)

Q Supervised

Ü unsupervised

@ semi-supervised

Ü reinforcement

8. Which one of the folowing is the Method to solve Noisy Data problem?
(1 Point)

Ü Ignore the tuple

Q Binning

Ü Data integration

@ All of the above

9. Which one of the following is a required for Regression?
(1 Point)

Q Apriori

Q Bayseian

Ü Decision Tree

@ Linear Regression

10. Which of the following classifications would best suit the student performance classification systems?
(1 Point)

@ If...then ... analysis

Q Market-basket analysis

- O Regression analysis
- O Cluster analysis

11. The Apriori algorithm is a _____.
(1 Point)

- O top-down search
- O breadth first search
- O depth first search
- @ bottom-up search

12. Incremental learning referred to_
(1 Point)

- Q Machine-learning involving different techniques
- @ The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned
- Q Learning by generalizing from examples
- O None of these

13. ..... is an essential process where intelligent methods are applied to extract data patterns.
(1 Point)

- O Data warehousing
- @ Data mining
- O Text mining
- O Data selection

14. In___data is partitioned in k-folds
(1 Point)

O a. holdout

@ b. cross validation

O c. Random subsampling

O d. bootstrap

15. Which of the following process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evolution and knowledge presentation?
(1 Point)

@ KDD process

O Related Database

Ü OLAP

O MDX process

16. Market-basket problem was formulated by _____.
(1 Point)

@ Agrawal et al

O Steve et al

O Toda et al

O Simon et al

17. Learning algorithm referrers to
(1 Point)

O A) An algorithm that can learn using training data

O B) A sub-discipline of computer science that deals with the design and implementation of learning algorithms.

@ C) A machine-learning approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.

Ü D) None of these

18. Cost complexity pruning algorithm is used in?
(1 Point)

@ CART

Ü (4.5

Ü ID3

Ü All

19. Neural network is the.....network
(1 Point)

@ Multilayer Perceptron

Ü Multiused

Ü Multipal

Ü None of the above

20. Classification is
(1 Point)

@ A) A subdivision of a set of examples into a number of classes.

Ü B) A measure of the accuracy, of the classification of a concept that is given by a certain theory.

Ü C) The task of assigning a classification to a set of examples

U D) None of these

21. Euclidean distance measure is
(1 Point)

@ A stage of the KDD process in which new data is added to the existing selection

O The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them

O The process of executing implicit previously unknown and potentially useful information from data

O The distance between two points as calculated using the Pythagoras theorem

22. Attribute selection measures are also known as splitting rules.
(1 Point)

@ True

O False

23. Which statement is true about the K-Means algorithm?
(1 Point)

O All attribute values must be categorical

O The output attribute must be cateogrical

O Attribute values may be either categorical or numeric

@ All attributes must be numeric

24. In ___semi supervised learning , model is built with training data and same one is used for

training with non labeled data
(1 Point)

Ü Self training

(i) cotraining

25. Classification accuracy is
(1 Point)

Q A) A subdivision of a set of examples into a number of classes

(i) B) Measure of the accuracy, of the classification of a concept that is given by a certain theory.

Q C) The task of assigning a classification to a set of examples

Ü D) None of these

26. Classification is
(1 Point)

(i) subdivision of a set of examples into a number of classes

Q measure of the accuracy, of the classification of a concept that is given by a certain theory

Q The task of assigning a classification to a set of examples

Ü None of these

27. _____ are needed to identify training data and desired results.
(1 Point)

Ü Programmers

Ü Designers

@ Users

Ü Administrators

28. Cross validation is a .... procedure
(1 Point)

@ Resampling

Ü Split

Ü Shuffle dataset

Ü All of the above

29. What is the input for K-Means Clustering Algorithm?
(1 Point)

Ü only a data table

Q only number of clusters (k)

Q only data set containing n objects (n)

@ Both number of Clusters (k) and a data set containing objects (n)

30. Which one of the following is Multidimentional Data Model type?
(1 Point)

@ Star Schema Model

Ü KDD

Q Market Basket Analysis

Ü CLARA

31. The first phase of A Priori algorithm is __ .
(1 Point)

Q Candidate generation

@ Itemset generation

Q Pruning

Q Partitioning

32. Decision tree Algorithm originaly intended for ....
(1 Point)

Ü Association

@ Classification

Q Clustering

Ü sequentitial pattern

33. Which one of these is a tree based learner?
(1 Point)

Ü Rule based

Q Bayesian Belief Network

Ü Bayesian classifier

@ Random

34. Classification training referred to
(1 Point)

Q A) A subdivision of a set of examples into a number of classes

◉ B) A measure of the accuracy, of the classification of a concept that is given by a certain

theory.

O C) The task of building a classification model by set of examples

Ü D) None of these

35. Which of the following is the collection of data objects that are similar to one another within the same group?
(1 Point)

O Partitioning

@ Association

O Cluster

Ü Classification

36. Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the other.
(1 Point)

@ True

O False

37. Which one of the following is a classification algorithm?
(1 Point)

O Apriori

Ü Bayseian

@ K-Mean's

Ü K-Medoid

38. _____ data are noisy and have many missing attribute values.
(1 Point)

@ Preprocessed

Q Cleaned

Ü Real-world

Ü Transformed

39. Learning algorithm referrers to_
(1 Point)

@ An algorithm that can learn

Ü A sub-discipline of computer science that deals with the design and implementation of learning algorithms

Ü A machine-learning approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.

Ü None of above

40. Which of the following are the related technologies of Data Mining
(1 Point)

@ Machine Learning

Ü DBMS

Q Statistics

Q All of the above

41. Treating incorrect or missing data is called as _____.
(1 Point)

U selection

@ preprocessing

Ü transformation

Ü interpretation

42. After the pruning of Apriori algorithm, _____ will remain.
(1 Point)

@ Only candidate set

Ü No candidate set

Ü Only border set

Ü No border set

43. _____ is efficient for large data sets but sensitive to outliers
(1 Point)

@ K-Medoids

Q K-Means

Q Apriori

Ü CLARA

44. 103 uses ___ and Information Gain to construct a decision tree.
(1 Point)

Q Entropy

Q Data set

@ Knowledge Discovery

Ü KNN Classifier

45. Learning is
(1 Point)

Ü A) The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system

@ B) It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks.

O C) A process where an individual learns how to carry out a certain task

Ü D) None of these

46. In___data is partitioned into training and testing
(1 Point)

Ü holdout

@ cross validation

Ü Random subsampling

Ü bootstrap

47. Most frequency data item is computed in terms of
(1 Point)

O mean

@ median

O mode

Ü Standard

48. Treating incorrect or missing data is called as _____.
(1 Point)

@ selection

O preprocessi ng

O transformation

O interpretation

49. Capability of data mining is to build _____ models.
(1 Point)

Ü retrospective

Ü interrogative

Ü imperative

@ predictive

50. Machine learning is_
(1 Point)

Ü An algorithm that can learn

@ A sub-discipline of computer science that deals with the design and implementation of learning algorithms

O An approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.

O None of above

51. Which of the following sentences are not correct in reference to Information gain?
(1 Point)

@ It is biased towards single-valued attributes

O It is biased towards multi-valued attributes

U ID3 makes use of information gain

Ü The approach used by ID3 is greedy

52. Inductive learning is_
(1 Point)

@ a. The learning algorithmic analyzes the examples on a systematic basis and makes incremental b.

O b. adjustments to the theory that is learned

O c. Learning by generalizing from examples

Ü d. None of these

53. CLARA stands for
(1 Point)

@ Clustering Large Applications

O Clustering Algorithm for Random Sample

O Clustering Algirithm for Large Sample

O Clusteing Algorithm for Random Applications

54. Which one of the following is a clustering algorithm?
(1 Point)

O Apriori

O Bayseian

@ Decision Tree

O K-Medoid

55. Multivariate split is where the partitioning of tuples is based on a combination of attributes rather than on a single attribute.
(1 Point)

@ True

O False

56. An OLAP tool provides:
(1 Point)

@ Multidimensional analysis

Ü Roll-up and drill-down

Ü Slicing and dicing

Ü Rotation

57. What is the approach of basic algorithm for decision tree induction?
(1 Point)

O Greedy

@ Top Down

Ü Procedural

Ü Step by step

58. Which of the Rule is closely allied with Market Basket Analysis?
(1 Point)

O Classification Rule

Ü Cluster Analysis Rule

@ Association Rule

O Relational Databases

59. What is gini index?
(1 Point)

@ It is a type of index structure

O It is a measure of purity

O Both options except none

O None of the options

60. ML is a field of AI consisting of learning algorithms that?
(1 Point)

O Improve their performance

O At executing some task

O Over time with experience

@ All of the above

61. Binary attribute are
(1 Point)

@ A) This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit

O B) The natural environment of a certain species.

O C) Systems that can be used without knowledge of internal operations.

O D) None of these

62. _____ is efficient for large data sets but sensitive to outliers
(1 Point)

- ○ K-Medoids
- ◉ K-Means
- ○ Apriori
- ○ CLARA

Submit

Never give out your password. Report abuse