Department of Computer Science & Engineering
Question Bank
IV Year I Semester – 2024-25
DATA MINING

UNIT-1

Multiple Choice Questions

1. Smoothing techniques are _____

A. binning

B. aggregation

C. Normalization

D. None

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

A) Data warehousing

B) Data mining

C) Text mining

D) Data selection

3. 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

4. Which of the following is not a data mining functionality?

A) Characterization and Discrimination

B) Classification and regression

C) Selection and interpretation

D) Clustering and Analysis

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

A) Data Characterization

B) Data Classification

C) Data discrimination

D) Data selection

6. …….. is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

A) Data Characterization

B) Data Classification

C) Data discrimination

D) Data selection

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

A) Data Characterization

B) Data Classification

C) Data discrimination

D) Data selection

8. The various aspects of data mining methodologies is/are …….

i) Mining various and new kinds of knowledge

ii) Mining knowledge in multidimensional space

iii) Pattern evaluation and pattern or constraint-guided mining.

iv) Handling uncertainty, noise, or incompleteness of data

A) i, ii and iv only

B) ii, iii and iv only

C) i, ii and iii only

D) All i, ii, iii and iv

9. The full form of KDD is ……

A) Knowledge Database

B) Knowledge Discovery Database

C) Knowledge Data House

D) Knowledge Data Definition

10. The out put of KDD is …….

A) Data

B) Information

C) Query

D) Useful information

Fill in the Blanks

11. Extracting knowledge from multiple, heterogeneous and external sources refers to_____

12. _____perform a linear transformation on the original data?

13. Many people treat data mining as synonym for another popularly used term----------

14. The discrete wavelet transformation is closely related to the ----transform

15. _____ is task of discovering interesting patterns from large amounts of data.

16. _____ data marts are sources directly from enterprise data warehouse.

17. _____contains a subset of corporate –wide data that is of value to a specific group of users.

18. _____ converts data from legacy or host format to warehouse format.

19. _____ is the estimate of the strength of the implication of the rule.

20.Binning Method is used for----------

UNIT-2

Multiple Choice Questions

1. What does Apriori algorithm do?

a. It mines all frequent patterns through pruning rules with lesser support

b. It mines all frequent patterns through pruning rules with higher support

c. Both a and b

d. None of the above

2. What does FP growth algorithm do?

a. It mines all frequent patterns through pruning rules with lesser support

b. It mines all frequent patterns through pruning rules with higher support

c. It mines all frequent patterns by constructing a FP tree

d. All of the above

3. What techniques can be used to improve the efficiency of apriori algorithm?

a. Hash-based techniques

b. Transaction Reduction

c. Partitioning

d. All of the above

4. What do you mean by support(A)?

a. Total number of transactions containing A

b. Total Number of transactions not containing A

c. Number of transactions containing A / Total number of transactions

d. Number of transactions not containing A / Total number of transactions

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

a. Support(A B) / Support (A)

b. Support(A B) / Support (B)

c. Support(A B) / Support (A)

d. Support(A B) / Support (B)

6. Which of the following is direct application of frequent itemset mining?

a. Social Network Analysis

b. Market Basket Analysis

c. Outlier Detection

d. Intrusion Detection

7. What is not true about FP growth algorithms?

a)It mines frequent itemsets without candidate generation.

b. There are chances that FP trees may not fit in the memory

c. FP trees are very expensive to build

d. It expands the original database to build FP trees.

8. When do you consider an association rule interesting?

a. If it only satisfies min_support

b. If it only satisfies min_confidence

c. If it satisfies both min_support and min_confidence

d. There are other measures to check so

9. What is the difference between absolute and relative support?

a. Absolute - Minimum support count threshold and Relative - Minimum support threshold

b. Absolute - Minimum support threshold and Relative - Minimum support count threshold

c. Both mean same

d.No relation between the two

10. What is the relation between candidate and frequent itemsets?

a. A candidate itemset is always a frequent itemset

b. A frequent itemset must be a candidate itemset

c. No relation between the two

d. Both are same

Fill in the Blanks

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

12.If T consist of 500000 transactions, 20000 transaction contain bread, 30000 transaction contain jam, 10000 transaction contain both bread and jam. Then the support of bread and jam is _____.

13. The left hand side of an association rule is called_____

14. The right hand side of an association rule is called____

15. All set of items whose support is greater than the user-specified minimum support are called as_____.

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

17.A priori algorithm is otherwise called as _____.

18.A rule concerns associations between the presence or absence of items, then it referred as

---

19.----------association rule consist of atleast two dimensions.

20.---------- algorithm is used to mine frequent item sets using candidate generation.

UNIT-3

Multiple Choice Questions

1. The basic idea of the apriori algorithm is to generate_____ item sets of a particular size &scans the database.
A. candidate.
B. primary.
C. secondary.
D. superkey.
2. _____is the most well known association rule algorithm and is used in most commercial products.
A. Apriori algorithm.
B. Partition algorithm.
C. Distributed algorithm.
D. Pincer-search algorithm.
3. An algorithm called_____is used to generate the candidate item sets for each pass after thefirst.
A. apriori.
B. apriori-gen.
C. sampling.
D. partition.

4. The basic partition algorithm reduces the number of database scans to _____ & divides it intopartitions.
A. one.
B. two.
C. three.
D. four.
5. _____and prediction may be viewed as types of classification.
A. Decision.
B. Verification.
C. Estimation.
D. Illustration.

6. _____can be thought of as classifying an attribute value into one of a set of possibleclasses.
A. Estimation.
B. Prediction.
C. Identification.
D. Clarification.
7. Prediction can be viewed as forecasting a_____value.
A. non-continuous.
B. constant.
C. continuous.
D. variable.
8. _____data consists of sample input data as well as the classification assignment for the data.
A. Missing.
B. Measuring.
C. Non-training.
D. Training.
9 .Rule based classification algorithms generate _____ rule to perform the classification.
A. if-then.
B. while.
C. do while.
D. switch.
10. A _____ is a flowchart-like tree structure
A. Navie Bayesian
B. Rule Based
C. Decision tree
D. Binary tree.

Fill in the Blanks

1. ----------is the process of scaling data such that it falls with in a specified range.

2. A-----represent the test on an attribute

3. --------represents the outcome of the test

4. The ------- represents the class label

5. --------gives the impurity of the data partition

6. A categorical variable is generalized form of ---------with more than two states

7. Baye's Theorem___

8. _____ are the methods to remove model over fitting

9.classification is ____

10.over fitting means _____

UNIT-4

Multiple Choice Questions

1._____ clustering technique start with as many clusters as there are records, with each cluster having only one record.
A Agglomerative.
B. divisive.
C. Partition.
D. Numeric.

2.In _____ algorithm each cluster is represented by the center of gravity of the cluster.
A. k-medoid.
B. k-means.
C. STIRR.
D. ROCK.

3. In _____ each cluster is represented by one of the objects of the cluster located near thecenter.
A. k-medoid.
B. k-means.
C. STIRR.
D. ROCK.

4. Pick out a hierarchical clustering algorithm.
A. DBSCAN
B. BIRCH.
C. PAM.
D. CURE.

5. CLARANS stands for _____.

A. CLARA Net Server.

B. Clustering Large Application RAnge Network Search.

C. Clustering Large Applications based on RANdomized Search.

D. CLustering Application Randomized Search.

6. Interval scaled variables are----measurements of a linear scale.

A. numeric

B.continuous

C.differentiable

D. non-continuous

7. Clustering large applications can be shortened as ----

A. DBSCAN

B.OPTICS

C.STING

D.CLARA

8. Most of the partitioning methods cluster objects are based on ----

A. number of clusters

B.distance between objects

C.number of objects in each class

D.learning rate

9. DBSCAN is a -----clustering algorithm

A. partitioning methods

B.hierarchical methods

C.density based methods

D.grid based methods

10. Which of the following is required by K-means clustering ?
a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) all of the Mentioned

Fill in the Blanks

1. dissimilarity matrix is also called as -------structure.

2. Data matrix is also called as ---------structure.

3. A categorical variable is generalized form of ---------with more than two states.

4. Rock stands for ----------

5. ----------is a statistical information grid approach

6.Data which are inconsistent with the remaining set of data is called as _____

7.A hierarchical method can be classified as being either _____

8._____methods quantize the object space into a finite number of cells that form a grid structure.

9._____ methods hypothesize a model for each of the clusters and find the best fit of the data to the given model.

10._____clustering approach that performs clustering by incorporation of user-specified or application-oriented constraints.

Unit-5

Multiple Choice Questions

1._____sequence data consist of long sequences of numeric data, recorded at equal time intervals

a.time-series data

b. Symbolic sequence data

c. Biological sequences

d.sequence data

2._____consist of long sequences of event or nominal data, which typically are not observed at equal time intervals.

a. time-series data

b. Symbolic sequence data

c. Biological sequences

d.sequence data

3._____ sequences include DNA and protein sequences.

a. time-series data

b. Symbolic sequence data

c. Biological sequences

d.sequence data

4. For effective trend analysis, the data often need to be “deseasonalized” based on a _____computed by autocorrelation

a. seasonal index
b. long-term movements
c. Cyclic movements
d. Random movements

5. _____ are used to represent the probabilities of substitutions of nucleotides or amino acids and probabilities of insertions and deletion

a. Substitution matrices

b. multiple sequence

c. pairwise sequence

d. a phylogenetic tree.

6. _____ data are data that relate to both space and time

a. Time-series

b. Spatiotemporal

c. Multimedia

d. Text and web

7. A _____typically consists of a large number of interacting physical and information components

a. cyber-physical system
b. cyber-crime system
c. crime-physical system

d. cyber-data system

8 _____data mining is an interdisciplinary field that integrates image processing and understanding, computer vision, data mining, and pattern recognition

a. Data stream
b. Text
c. Spatial
d. Multimedia

9. _____mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics.

a. Data stream
b. Text
c. Spatial
d. Multimedia

10. _____mining analyzes web content such as text, multimedia data, and structured data

a. Web content

b. Spatial

c. Multimedia

d. Data stream

Fill in the Blanks

1._____mining is the process of using graph and network mining theory and methods to analyze the nodes and connection structures on the Web

2._____refer to data that flow into a system in vast volumes, change dynamically, are possibly infinite, and contain multidimensional features.

3."High quality" in text mining usually refers to a combination of _____

4.CPS systems may be interconnected so as to form large _____cyber-physical networks.

5.Spatial data, in many cases, refer to _____data stored in geospatial data repositories

6. A _____data set consists of sequences of numeric values obtained over repeated measurements of time.

7.Many time-series similarity queries require _____matching

8.An _____is the process of lining up sequences to achieve a maximal identity level

9. The _____is a synopsis data structure that can be used to approximate the frequency distribution of element values in a data stream.

10.A pattern is considered frequent if its count satisfies a _____support