Total number of printed pages: 5 **PG/2nd/PCSE212**

2021

DATA MINING & DATA WAREHOUSING

Full Marks: 60

Time: Two hours

The figures in the margin indicate full marks for the questions.

A. Multiple Choice Questions 1 x 20=20

1. Which of the following is focusing on modeling and analysis of data?
   a. Subject-oriented
   b. Integrated
   c. Integrated
   d. None of these
2. Which one is used for the tasks of data warehouse?
   a. OLAP
   b. OLTP
   c. ER+
   d. None of these
3. Which type of approach is helpful in outlier detection?
   a. Clustering
   b. Classification
   c. Preprocessing
   d. Association
4. Outliers are discarded due to
   a. irrelevant
   b. Noise
   c. Missing data
   d. Not sufficient
5. A probability-based model uses the technique of
   a. Clustering

b. regression

c. prediction

d. Both ‘b’ & ‘c’

6. A good clustering approach produces

a. Cohesive within clusters

b. Distinctive within clusters

c. Non-cohesive within clusters

d. Non-Distinctive within clusters

7. Multiple level granularity structure is in

a. Density based approach

b. Grid based approach

c. Hierarchical based approach

d. Model based approach

8. Dissimilarity matrix represents

a. Proximity of pairs of objects

b. Non-proximity of pairs of objects

c. Highly similar pairs of objects

d. Simple data matrix

9. Nominal variables have

a. 2 states

b. More than 2 states

c. 2 states only

d. None of these

10. CLIQUE algorithm is considered as

a. Density-based clustering

b. Grid-based clustering

c. Subspace clustering

d. All of these

11. Prediction means

a. Mapping with finite number of classes

b. Mapping with infinite number of classes

c. Mapping with one classes

d. None of these

12. Clustering technique is used for

a. Data smoothing

b. Data integration

c. Data Normalization

d. None of these

13. Which one clustering technique is less influenced by outliers?

a. K-means

b. K-Medoids

c. Hierarchical

d. None of these

14. Precision considered as

a. measure of completeness

b. measure of exactness

c. measure of true negative

d. None of these

15. CART stands for

a. Classification and Random Forest Trees

b. Classification and Regression Trees

c. Clustering and Regression Trees

d. Clustering and Random Forest Trees

16. Recall is a

a. measure of completeness

b. measure of exactness

c. measure of true negative

d. None of these

17. Information gain is calculated over

a. Attributes

b. Subset of data

c. Both of these

d. None of these

18. Overfitting problem happens in decision tree due to
    a. Grow the tree just branch wise
    b. Grow the tree just deeply
    c. Grow the tree in both the way
    d. None of these
19. A multidimensional model exist in form of
    a. Star schema
    b. Snowflake schema
    c. Fact constellation schema
    d. All of these
20 HOLAP is benefitting from
    a. ROLAP
    b. MOLAP
    c. ROLAP and MOLAP
    d. None of these

B. Very Short Question 2*6=12

1. How is a data warehouse different from a database? How are they similar?
2. What are the value ranges of the following normalization methods?
   (a) min-max normalization
   (b) z-score normalization
   (c) normalization by decimal scaling
3. What is good clustering?
4. Why is naive Bayesian classification called “naïve”?
5. Why is tree pruning useful in decision tree induction?
6. What is confusion matrix?

C Short Question 4*7=28

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 avg_grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg_grade measure stores the

actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.
Draw a snowflake schema diagram for the data warehouse.

2. A database has five transactions. Let min sup = 60% and min conf = 80%. Find all frequent itemsets using Apriori.

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

3. Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm. Also, illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (such as AGNES).

4. Briefly outline how to compute the dissimilarity between objects described by the following types of variables:

   (a) Numerical (interval-scaled) variables

   (b) binary variables

   (c) Nominal variables

   (d) Ordinal variables

   (e) Ratio-scaled variables

5. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

   (a) Compute the Euclidean distance between the two objects.

   (b) Compute the Manhattan distance between the two objects.

   (c) Compute the Minkowski distance between the two objects, using q = 3.

6. Briefly outline the major ideas of classification and prediction.

7. Describe Information for a partition on an attribute with the formulation.