A2–2367/SCSJC62 APRIL 2022

DATA MINING

Time : Three hours Maximum : 75 marks

PART A — (10 × 1 = 10 marks)

Answer ALL questions.

Choose the correct answer :

1. The FP-growth algorithm has _____ phases.

   (a) one (b) two

   (c) three (d) four

2. A priori algorithm is otherwise called as _____

   (a) width-wise algorithm

   (b) level-wise algorithm

   (c) pincer-search algorithm

   (d) FP growth algorithm

3. The 'Slice' operation deals with _____

   (a) Selecting all but one dimension of the data cube

   (b) Merging the cells along one dimension

   (c) Merging cells of all but one dimension

   (d) Selecting the cells of any one dimension of the data cube.

4. _____ is data about data.

(a) Metadata (b) Microdata

(c) Minidata (d) Multidata

5. Attributes in the database schema that will be used to label nodes in the tree and around which the decisions will take place are called_____

(a) ordering (b) decision

(c) splitting (d) Algorithm

6. Overfitting occurs when a model _____

(a) does fit in future states

(b) does not fit in future states

(c) does fit in current state

(d) does not fit in current state.

7. In _____ algorithm each cluster is represented by the center of gravity of the cluster.

(a) k-medoid (b) k-means

(c) STIRR (d) ROCK

8. _____ distance metric is based on the maximum attribute difference.

(a) Categorical data (b) Chebychev

(c) Euclidean (d) Manhattan

9. _____ mining is concerned with discovering the model underlying the link structures of the web.

(a) Data structure (b) Web structure

(c) Text structure (d) Image structure

10. In web mining, _____ is used to know the order in which URLs tend to be accessed.

(a) clustering

(b) associations

(c) sequential analysis

(d) classification

PART B — (5 × 7 = 35 marks)

Answer ALL questions, choosing either (a) or (b)

11. (a) Explain the salient differences between the major data mining techniques.

Or

(b) Describe the performance evaluation of algorithms.

12. (a) What is OLAP? Explain.

Or

(b) Write down the guidelines for data warehouse implementation.

13. (a) What kind of data is the decision tree method most suitable for? Explain.

Or

(b) Explain the terms for bootstrapping, bagging and boosting for improving the accuracy of a classification method.

14. (a) What are the types of data? Give examples.

Or

(b) Write down the desired features of cluster analysis.

15. (a) Explain about the web structure mining.

Or

(b) Discuss the advantages and disadvantages of the HITS algorithm.

PART C — (3 × 10 = 30 marks)

Answer any THREE questions.

16. Explain the six groups of data mining applications.

17. What are the main purposes of metadata? What properties should metadata have? Describe.

18. Discuss about the Naïve Bayes method

19. What are the various methods for computing distances between clusters? Explain.

20. What is web page? Explain the classification of web pages.
