Total No. of printed pages = 3

CSE 1816 PE 21   16/6/23

Roll No. of candidate ☐☐☐☐☐☐☐☐☐

2023

B.Tech. 6th Semester End-Term Examination

DATA MINING

New regulation (w.e.f. 2017-18)

New Syllabus (w.e.f. 2018-19)

Full Marks – 70   Time – Three hours

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

Answer question No. 1 and any *four* from the rest.

1.   Answer the following : (MCQ)   (10 × 1 = 10)

   (i)   The goal of ————— is to discover both the dense and sparse regions of a data set.

     (a)   Association rule .

     (b)   Classification.

     (c)   Clustering.

     (d)   Genetic Algorithm.

   (ii) Clustering is

     (a)   Predictive and unsupervised

     (b)   Predictive and supervised

     (c)   Descriptive and supervised

     (d)   Descriptive and unsupervised

   (iii) The most important part of ————— is selecting the variables on which clustering is based.

     (a)   interpreting and profiling clusters

     (b)   selecting a clustering procedure

     (c)   assessing the validity of clustering

     (d)   formulating the clustering problem

[Turn over

(iv) The most commonly used measure of similarity is the ——————— or its square.

    (a) euclidean distance

    (b) city-block distance

    (c) Chebychev's distance

    (d) Manhattan distance

(v) Which method of analysis does not classify variables as dependent or independent?

    (a) regression analysis

    (b) discriminant analysis

    (c) analysis of variance

    (d) cluster analysis

(vi) Which statement is not true about cluster analysis?

    (a) Objects in each cluster tend to be similar to each other and dissimilar to objects in the other clusters.

    (b) Cluster analysis is also called classification analysis or numerical taxonomy.

    (c) Groups or clusters are suggested by the data, not defined a priori.

    (d) Cluster analysis is a technique for analyzing data when the criterion or dependent variable is categorical and the independent variables are interval in nature.

(vii) Which one of the clustering technique needs the merging approach?

    (a) Partitioned   (b) Naïve Bayes

    (c) Hierarchical   (d) Both (a) and (c)

(viii) 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

(ix) Identify the type of relationship between fact and dimension table in a star schema.

    (a) One to one   (b) Many to many

    (c) One to many   (d) Many to one

(x) Identify the method for incremental conceptual clustering.

    (a) STING   (b) COBWEB

    (c) COBRA   (d) OLAP

2. (a) Explain the components of the architecture of a typical data mining system with the help of a diagram. (7)

   (b) Explain the need and steps involved in data preprocessing. (8)

3. (a) What is data mining? How is it different from KDD? Show the different stages of KDD. (5)

   (b) Explain the architecture of a data warehouse with the help of a diagram. (10)

4. (a) Discuss the various types of meta data. (5)

   (b) Differentiate between OLTP and OLAP systems. (5)

   (c) Explain the multidimensional data model. (5)

5. (a) What are the advantages and the disadvantages of decision tree over classification techniques. (5)

   (b) A database has 4 transactions. Let min_sup = 60% and min_conf = 80%

TID	DATE	ITEMS
T1	11/01/2018	{K,A,D,B}
T2	02/03/2018	{D,A,C,E,B}
T3	25/07/2018	{C,A,B,E}
T4	22/09/2018	{B,A,D}

Find all the frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the 2 mining processes. (5 + 5 = 10)

6. (a) What is Clustering? Explain anyone method to find dissimilarity of objects in a given dataset. (5)

   (b) What are the different types of Clustering method? Explain anyone of them. (5)

   (c) What is Outlier Analysis? Explain anyone method for detecting outlier. (5)

7. (a) Describe statistical measures in large databases. (5)

   (b) Explain mining multilevel association rules from transactional databases. (5)

   (c) Define and describe the basic similarities and differences among ROLAP, MOLAP and HOLAP. (5)

8. Write short notes: (any three) (3 × 5 = 15)

   (a) Data-Mart.

   (b) k-Medoids Method

   (c) DBSCAN

   (d) ROCK