Total No. of printed pages = 3

CSE 1816 PE 21

3742

Roll No. of candidate | | | | | | | | | | |

2022

B.Tech. 6th Semester End-Term Examination

Computer Science and Engineering

DATA MINING

(New Regulation & New Syllabus)

Full Marks – 70

Time – Three hours

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

Answer Question No. 1 is compulsory and any *four* from the rest.

1. Answer the following :(MCQ/Fill in the blanks) (10 × 1 = 10)

(i) Finding of hidden structure in unlabeled data is called

(a) Supervised learning (b) Unsupervised learning

(c) Reinforcement learning (d) none of the above

(ii) Which one of the following refers to the binary attribute?

(a) This takes only two values: 0 and 1

(b) The natural environment of a certain species

(c) Systems that can be used without knowledge of internal operations

(d) All of the above

(iii) 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

(iv) ____________ is data about data

(a) Minidata (b) Microdata

(c) Metadata (d) Multidata

(v) Removing duplicate records is a process called

(a) Pruning (b) Cleansing

(c) Cleaning (d) Recovery

[Turn over

(vi) 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

(vii) What does OLTP stand for:

(a) Offline Transaction Processing

(b) Online Transaction Processing

(c) Outline Traffic Processing

(d) None of the above

(viii) Which is needed by K-means clustering?

(a) Defined distance metric

(b) Number of clusters

(c) Initial guess as to cluster centroids

(d) All of the above

(ix) A ———————— allows data to be modeled and viewed in multiple Dimensions.

(x) Web data is ————————

(a) Structured data (b) Un-structured data

(c) Only text data (d) Binary data

2. (a) What is data mining? Briefly explain about various data mining tasks. Also mention the key challenges of data mining. (3 + 3 + 4 = 10)

(b) What do you mean by data repository? What are the different types of data repositories? (2 + 3 = 5)

3. (a) What do you mean by similarity measure? Briefly explain about at least two measures. (3 + 3 = 6)

(b) Given two objects, x (22, 1,42) and y (20,0, 36), in d-dimensional space (3 × 3 = 9)

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

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

(iii) Compute the Minkowski distance between the two objects, using $p = 3$.

4. (a) What do you mean by association rule mining? (3)

(b) Define the following: (3 + 2 = 5)

(i) Support and confidence

(ii) Frequent itemset

(c) Explain Apriori algorithms for generating frequent item sets using candidate generation for the following transaction dataset: (7)

Transaction	List of Items
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
I6	I1, I2, I3, I4

Where support = 50% and Confidence = 60%

5. (a) What do you mean by cluster Analysis? What are the different approaches for cluster analysis? (3 + 3 = 6)

(b) Discuss any one of the following clustering algorithms with a suitable example: (9)

(i) K-Means

(ii) BIRCH

(iii) DBSCAN

6. (a) What is an outlier? Mention about the various schemes for handling outliers. (3+3=6)

(b) How classification is performed in data mining? Explain them with suitable examples in brief. (3+6=9)