13-06-19

Total No. of printed pages = 6

CS 131603

Roll No. of candidate □□□□□□□□□□

2019

B.Tech. (CSE) 6th Semester End-Term Examination

DATA MINING

Full Marks – 100 Time – Three hours

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

Question No.1 is compulsory. Answer any six from the following.

1. Answer the following Multiple Choice questions. (10 × 1 = 10)

(i) Data mining is

(a) The actual discovery phase of a knowledge discover process.

(b) The stage of selecting the right data for a KDD process.

(c) A subject-oriented integrated time variant non-volatile collection of data in support of management

(d) None of the above

[Turn over

(ii) Classification is

(a) A subdivision of a set of examples into a number of classes.

(b) A measure of the accuracy of the classification of a concept that us given by a certain theory.

(c) The task of assigning a classification to a set of examples.

(d) None of the above

(iii) Data selection is

(a) The actual discovery phase of a knowledge discovery process.

(b) The stage of selecting the right data for a KDD process

(c) A subject-oriented integrated time variant non-volatile collection of data in support of management.

(d) None of the above

(iv) Euclidean distance measure is

(a) A stage of the KDD process in which new data is added to the existing selection.

(b) The process of funding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them

(c) The distance between two points as calculated using the Pythagoras theorem

(d) None of the above

(v) Heterogeneous databases referred to

(a) A set of databases from different vendors, possibly using different database paradigms

(b) An approach to a problem that is not guaranteed

(c) Information that is hidden in a dațabase and that cannot he recovered by a simple SQL query.

(d) None of the above

(vi) KDD (Knowledge Discovery in Databases) is referred to

(a) Non-trivial extraction of implicit previously unknown arid potentially useful information from data

(b) Set of columns in a database table that can he used to identify each record within this table uniquely

(c) Collection of interesting and useful patterns in a database.

(d) None of the above

(vii) Machine learning is

(a) An algorithm that can learn

(b) A sub-discipline of computer science that deals with the design and implementation of learning algorithms.

(c) An approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.

(d) None of these

(viii) Patterns that can be discovered from a given database, can be of

(a) One type only

(b) No specific type

(c) More than one type

(d) Multiple type always

(ix) Supervised learning and unsupervised clustering both require at least one

(a) Hidden attribute.

(b) Output attribute

(c) Input attribute

(d) Categorical attribute

(x) A nearest neighbour approach is best used

(a) With large-sized datasets

(b) When irrelevant attributes have been removed from the data

(c) When a generalized model of the data is desirable

(d) When an explanation of What has been Fund is of primary importance.

Answer any SIX

(6 × 15 = 90)

2. Answer the following (5 × 3 = 15)

(a) Define: Data Mining

(b) What are the different types of data repositories?

(c) What are the issues in data mining?

(d) Differentiate classification and prediction

(e) How do we handle missing data?

3. Answer the following (5 × 3 = 15)

   (a) What is a data warehouse?

   (b) What is data transformation? Give example.

   (c) Define data cube.

   (d) What is metadata?

   (e) Differentiate OLTP and OLAP.

4. Answer the following (5 × 3 = 15)

   (a) What is a frequent item set?

   (b) What do you mean by Association rules?

   (c) Define the terms support and confidence.

   (d) Give two limitations of Apriori algorithm.

   (e) What is correlation Analysis?

5. Answer the following (5 × 3 = 15)

   (a) Compare clustering and classification

   (b) Differentiate between agglomerative and divisive hierarchical clustering.

   (c) What is cluster evaluation?

   (d) What is outlier detection?

   (e) Explain Graph-based clustering.

6. Answer the following (3 × 5 = 15)

   (a) What is data mining functionality? Explain different types of data mining functionality with examples.

   (b) With a neat diagram explain the architecture of data mining

   (c) Explain FP tree algorithm with an example.

7. Answer the following (3 × 5 = 15)

(a) Discuss the activities of data cleaning with the process associated with it.

(b) What is data classification? Give any one data classification method with example.

(c) Describe ROC curve

8. Answer the following (3 × 5 = 15)

(a) How to perform cross validation?

(b) Discuss Decision tree

(c) How to evaluate the performance of a classifier?

9. Answer the following (3 × 5 = 15)

(a) Write a short note about text mining.

(b) What is web usage mining?

(c) Describe the application of Weka.