Code : 051822

B.Tech 8th Semester Exam., 2019

DATA MINING

Time : 3 hours *Full Marks* : 70

Instructions :

(i) The marks are indicated in the right-hand margin.

(ii) There are ***NINE*** *questions in this paper.*

(iii) Attempt ***FIVE*** *questions in all.*

(iv) Question No. ***1*** *is compulsory.*

1. Choose the correct answer of the following (any *seven*) : 2×7=14

(a) Data scrubbing is which of the following?

(i) A process to reject data from the data warehouse and to create the necessary indexes.

(ii) A process to load the data in the data warehouse and to create the necessary indexes.

(iii) A process to upgrade the quality of data after it is moved into a data warehouse.

(iv) A process to upgrade the quality of data before it is moved into a data warehouse.

(b) The active data warehouse architecture includes which of the following?

(i) At least one data mart.

(ii) Data that can extracted from numerous internal and external sources.

(iii) Near real-time updates

(iv) All of the above

(c) A goal of data mining includes which of the following?

(i) To explain some observed event or condition.

(ii) To confirm that data exists.

(iii) To analyze data for expected relationships.

(iv) To create a new data warehouse.

(d) Bayesian classifiers is

(i) a class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory

(ii) any mechanism employed by a learning system to constrain the search space of a hypothesis

(Continued)

(iii) an approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation

(iv) None of the above

(e) Case-based learning is

(i) a class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory

(ii) any mechanism employed by a learning system to constrain the search space of a hypothesis

(iii) an approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation

(iv) None of the above

(Turn Over)

(f) Cluster is

(i) group of similar objects that differ significantly from other objects

(ii) operation on a database to transform or simplify data in order to prepare it for a machine-learning algorithm

(iii) symbolic representation of facts or idea from which information can potentially be extracted

(iv) None of the above

(g) A snowflake schema is which of the following types of tables?

(i) Fact

(ii) Dimension

(iii) Helper

(iv) All of the above

(h) Statistical significance is

(i) the science of collecting, organizing and applying numerical facts

(ii) measure of the probability that a certain hypothesis is incorrect given certain observations

(iii) one of the defining aspects of a data warehouse, which is specially built around all the existing applications of the operational data

(iv) None of the above

(i) Prediction is

(i) the result of the application of a theory or a rule in a specific case.

(ii) one of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table

(iii) discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces

(iv) None of the above

(j) Noise is

(i) a component of a network

(ii) in the context of KDD and data mining, this refers to random errors in a database table

(iii) one of the defining aspects of a data warehouse

(iv) None of the above

2. (a) What is data mining? Explain the process of Knowledge Discovery in Databases (KDD) with diagram. 7

(b) With suitable example, explain different types of attributes used in data mining. 7

3. (a) Suppose that you are employed as a data mining consultant for an E-commerce company. Describe how data mining and social media analysis can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining and anomaly detection can be applied. 7

(b) What is over-fitting? What is under-fitting? What can be done to address over-fitting/under-fitting in decision tree induction? 7

4. (a) Compare Naive Bayes, SVM and Neural network algorithms working with the help of example. 7

(b) What is clustering? What are the different types of clusters? Explain basic *K*-means clustering. 7

5. *(a)* In many applications, new data sets are incrementally added to the existing large data sets. Thus an important consideration for computing descriptive data summary is whether a measure can be computed efficiently in incremental manner. Use count, standard deviation and median as examples to show that distributive or algebraic measure facilitates efficient incremental computation, whereas a holistic measure does not. 7

(b) Write a short note on spatial data mining. What are the different types of dimensions and measures in a spatial data cube? 7

6. Suppose that we have the following data :

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
(2, 0)	(1, 2)	(2, 2)	(3, 2)	(2, 3)	(3, 3)	(2, 4)	(3, 4)	(4, 4)	(3, 5)

(a) Identify the cluster by applying the *K*-means algorithm, with $k = 2$. 7

(b) Apply the agglomerative clustering algorithm on (*a*, *b*, *c*, *d*, *e*, *f*) data only and draw the **Dendrogram and Proximity matrix**. 7

7. 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.

(a) Draw a snowflake schema diagram for the data warehouse. 5

(b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student. 5

(c) If each dimension has five levels (including all), such as "student<major <status<university<all", how many cuboids will this cube contain (including the base and apex cuboids)? 4

8. The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

(a) Draw a contingency table for each of the following rules using the transactions shown in Table below : Rules : 7

$\{b\} \rightarrow \{c\}$, $\{a\} \rightarrow \{d\}$, $\{b\} \rightarrow \{d\}$, $\{e\} \rightarrow \{c\}$, $\{c\} \rightarrow \{a\}$

Transaction ID	*Items Bought*
1	{*a*, *b*, *d*, *e*}
2	{*b*, *c*, *d*}
3	{*a*, *b*, *d*, *e*}
4	{*a*, *c*, *d*, *e*}
5	{*b*, *c*, *d*, *e*}
6	{*b*, *d*, *e*}
7	{*c*, *d*}
8	{*a*, *b*, *c*}
9	{*a*, *d*, *e*}
10	{*b*, *d*}

(b) Use the contingency tables in part (*a*) to compute and rank the rules in decreasing order according to the following measures : **Support, Confidence. Interest, Odds ratio**. 7

9. Write short notes on (any *four*) : 3½×4=14

(a) OLAP, MOLAP, HOLAP

(b) Data mining applications

(c) Association rule mining

(d) Support vector machine

(e) Decision tree method

(f) Back propagation

(g) KNN algorithms

★★★