No. of Printed Pages : **4**

Sl. No.

BE6-R4 : DATA WAREHOUSE AND DATA MINING

NOTE :

1. **Answer question 1 and any FOUR questions from 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time : 3 Hours **Total Marks : 100**

1. (a) Why computing the proximity between two attributes is often simpler than computing the similarity between two objects ?

(b) How is a data warehouse different from a database ? How are they similar ?

(c) Why is tree pruning useful in decision tree induction ? What is a drawback of using a separate set of tuples to evaluate pruning ?

(d) Use an example to show why the k-means algorithm may not find the global optimum, that is, optimizing the within-cluster variation.

(e) In outlier detection by semi-supervised learning, what is the advantage of using objects without labels in the training data set ?

(f) Why is the establishment of theoretical foundations important for data mining ? Name and describe the main theoretical foundations that have been proposed for data mining.

(g) Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules. **(7x4)**

2. (a) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results :

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(i) Calculate the mean, median and standard deviation of age and % fat.

(ii) Draw the box plots for age and % fat.

(iii) Draw a scatter plot and q-q plot based on these two variables.

(b) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8) :

(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 q= 3.

(iv) Compute the supremum distance between the two objects.

(c) What are the major challenges of mining a huge amount of data in comparison with mining a small amount of data ? **(8+8+2)**

3. (a) Consider the following data set for a binary class problem.

A	B	Class Label
T	F	1
T	T	1
T	T	1
T	F	2
T	T	1
F	F	2
F	F	2
F	F	2
T	T	2
T	F	2

(i) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose ?

(ii) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose ?

(iii) Is it possible that information gain and the gain in the Gini index favor different attributes ? Explain.

(b) Briefly outline the major steps of decision tree classification. **(10+8)**

4. (a) The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

Table 1. Market Transactions

TID	Items
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}

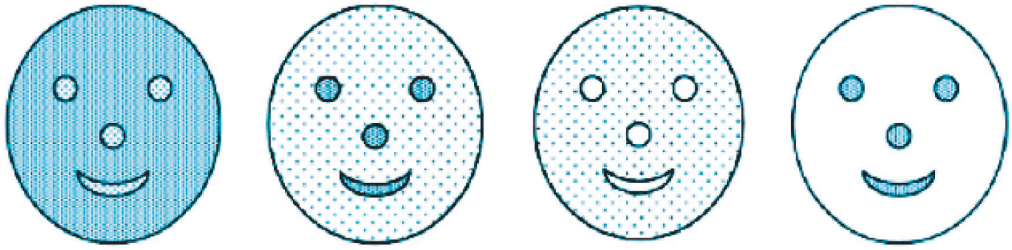
Rules : {b} → {c}, {a} → {d}, {b} → {d}, {e} → {c}, {c} → {a}.

(i) Draw a contingency table for the above-mentioned rules using the transactions shown in **Table 1**.

(ii) Use the contingency tables in part (1) to compute and rank the rules in decreasing order according to support, confidence, interest and Odds Ratio.

(b) Suppose you have the set C of all frequent closed item sets on a data set D, as well as the support count for each frequent closed item set. Describe an algorithm to determine whether a given item set X is frequent or not and the support of X if it is frequent. **(10+8)**

5. (a) Consider the following four faces shown in **Figure 1**. You are given two sets of 100 points that fall within the unit square. Again, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points.



(a) (b) (c) (d)

Figure 1. Smiling Faces

(i) For each figure, could you use single link to find the patterns represented by the nose eyes, and mouth ? Explain.

(ii) For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth ? Explain.

(iii) What limitation does clustering have in detecting all the patterns formed by the points in Figure 1 ?

(b) Both k-means and k-medoids algorithms can perform effective clustering.

(i) Illustrate the strength and weakness of k-means in comparison with k-medoids.

(ii) Illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme. **(10+8)**

6. (a) Suppose that a data warehouse consists of the four dimensions date, spectator, location, and game and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults or seniors, with each category having its own charge rate.

(i) Draw a star schema diagram for the data warehouse.

(ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should you perform in order to list the total charge paid by student spectators at GM Place in 2022 ?

(b) What are the differences between the three main types of data warehouse usage : information processing, analytical processing and data mining ? Discuss the motivation behind OLAP mining (OLAM).

(c) Write an algorithm for attribute oriented induction. **(8+6+4)**

7. (a) Explain data mining task primitives.

(b) Write a general algorithm for web crawling. Explain the meaning of each term used in PageRank algorithm.

(c) Suppose that your local bank has a data mining system. The bank has been studying your debit card usage patterns. Noticing that you make many transactions at home renovation stores, the bank decides to contact you, offering information regarding their special loans for home improvements.

(i) Discuss how this may conflict with your right to privacy.

(ii) Describe another situation in which you feel that data mining can infringe on your privacy.

(iii) Describe a privacy-preserving data mining method that may allow the bank to perform customer pattern analysis without infringing on customers' right to privacy.

(iv) What are some examples where data mining could be used to help society ? Can you think of ways it could be used that may be detrimental to society ? **(4+6+8)**

- o 0 o -