Total No. of Questions : 8]

SEAT No. :

P3658

[Total No. of Pages : 4

[4859] - 1076
B.E. (Semester - I)
DATA MINING TECHNIQUES & APPLICATIONS
(2012 Pattern) (End Sem.)

Time : $2\frac{1}{2}$ *Hours]* *[Max. Marks : 70*

Instructions to the candidates:

1) *Answer Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.*
2) *Neat diagrams must be drawn wherever necessary.*
3) *Figures to the right side indicate full marks.*
4) *Assume suitable data if necessary.*

Q1) a) For each of the following queries, identify and write the type of data mining task. **[6]**

i) Find all credit applicants who are poor credit risks.

ii) Identify customers with similar buying habits.

iii) Find all items which are frequently purchased with milk.

b) Consider the following: **[6]**

Transaction	Items
t_1	Bread, Jelly, Peanut Butter
t_2	Bread, Peanut Butter
t_3	Bread, Milk, Peanut Butter
t_4	Beer, Bread
t_5	Beer, Milk

Calculate the support and confidence for the following association rules

i) Bread $\rightarrow$ PeanutButter.

ii) Jelly $\rightarrow$ Milk.

iii) Beer $\rightarrow$ Bread.

P.T.O.

c) Consider the ten records given below **[8]**

ID	Income	Credit	Class	X_i
1	4	Excellent	h_1	X_4
2	3	Good	h_1	X_7
3	2	Excellent	h_1	X_2
4	3	Good	h_1	X_7
5	4	Good	h_1	X_8
6	2	Excellent	h_1	X_2
7	3	Bad	h_2	X_{11}
8	2	Bad	h_2	X_{10}
9	3	Bad	h_3	X_{11}
10	1	Bad	h_4	X_9

Calculate the prior probabilities of each of the class h_1, h_2, h_3, h_4 and probabilities for data points X_2, X_4, X_7 and X_8 belonging to the class h_1.

OR

Q2) a) Define Data Mining. Draw a pyramid showing relationship between Data Mining and Business Intelligence. Write types of users at different levels in the Pyramid. **[6]**

b) Write a pseudo code for Apriori algorithm and explain. **[6]**

c) Write a pseudo code for the construction of Decision Tree. State and justify its time complexity also. **[8]**

Q3) a) Using K-means Clustering, cluster the following data into 2 clusters. **[8]**

{2, 4, 10, 12, 3, 20, 30, 11, 25}

Show each of the steps.

b) Draw a diagram showing different approaches used for Clustering. **[3]**

c) Many clustering algorithms need to determine the distance between two clusters. Write the formula to determine the distance between two given clusters K_1 and K_2 using Single-link, Complete-link and Average methods. **[6]**

OR

Q4) a) Define Clustering. State the space complexity, time complexity and type (hierarchical, Partitional etc.) for following clustering algorithms **[9]**

i) Composite-link.

ii) MST.

iii) K-means.

b) Differentiate between K-means and K-mediod Clustering algorithms. **[2]**

c) Consider the following matrix. **[6]**

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

Draw the resultant Dendrograms when Single-link and composite-link clustering algorithms are used.

Q5) a) Consider the following 4 documents D_1, D_2, D_3 and D_4. For each document D_i, various terms that occur in document D_i is provided. **[6]**

D_1 = {To do is to be to be is to do}

D_2 = {To be or not to be I am what I am}

D_3 = {I think therefore I am}

D_4 = {Do do do da da da Let it be let it be}

Write the equations and calculate the Term frequency "tf", Inverse Document Frequency "idf" and "tf-idf" for the terms "to" and "do" for the documents D_1 and D_4.

b) A webmaster at XYZ organization learns that high percentage of users have following pattern of reference to pages

<P, Q, P, R>

What modification would be suggested if Web Usage Mining is used? State various purposes of Web Usage mining. **[5]**

c) Enumerate various Text Operations (also called Pre-processing) that are used by an Information Retrieval System. **[6]**

OR

Q6) a) Draw a neat diagram showing retrieval process of an IR system and briefly describe its components. **[6]**

b) Explain Precision and Recall. When a query "q" was fired for an IR system having 100 relevant documents w.r.t. the query "q", the system in all retrieved 68 documents out of the total collection of 600 documents. Out of 68 retrieved documents, 40 documents found to be relevant. What is the Precision and Recall of the system w.r.t. the given query "q". **[4]**

c) What is web crawler? Explain the working of a basic crawler. **[7]**

Q7) a) Define Big Data. State a few challenges of Big Data. **[6]**

b) Compare Business Intelligence and Big Data. **[4]**

c) Write a note on Reinforcement Learning. **[6]**

OR

Q8) a) Draw a diagram showing generalized systemic Machine Learning Framework. Briefly explain. **[6]**

b) What are the sources for Structured, semi-structured and Unstructured data? **[6]**

c) Write a note on Multi-perspective learning. **[4]**