05-2023

Paper / Subject Code: 89324 / Data Warehousing and Mining

T.E. Sem VI (C Scheme, R.2019) ECS May 2023.

Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory
2. Answer any three out of the remaining five questions.
3. Assume any suitable data wherever required and justify the same.

Q1 a) Every data structure in the data warehouse contains the time element. Why? [5]

b) Define initial load, incremental load, and full refresh. [5]

c) Calculate Accuracy, Recall and Precision with the help of following data: True Positive (TP)= 50, True Negative (TN) = 20, False Positive (FP)= 20, False Negative (FN)= 10 [5]

d) Elucidate Market Basket analysis with an example. [5]

Q2 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. [10]

(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 one perform in order to list the total charge paid by student spectators at GM Place in 2004?

b) For the given set of points identify clusters using a single linkage algorithm. Draw dendrogram. [10]

Object	Attribute(X)	Attribute(Y)
A	2	2
B	3	2
C	1	1
D	3	1
E	1.5	0.5

Q3 a) Name the set of basic transformation tasks. Give an example for each. [10]

b) A database has five transactions. Let min sup count = 2 and min conf = 60%. [10]

TID	Items
10	1, 3, 4
20	2, 3, 5
30	1, 2, 3, 5
40	2, 5
50	1, 3, 5

Find all frequent itemsets and strong association rules using Apriori Algorithm.

Q4 a) Describe slowly changing dimensions. What are the three types? Explain each type very briefly. [10]

b) The following table contains a training set D, of class-labeled tuples randomly selected from the AllElectronics customer database. Let buys_computer be the class label attribute. Using Naïve Bayesian classification predict the class label of a tuple X = (age = youth, income = medium, student = yes, credit rating = fair). [10]

RID	age	income	student	credit_rating	buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Q5 a) Explain Data mining as a step in KDD. Give the architecture of typical data mining System. [10]

b) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only [10]

(i) The three cluster centers after the first-round execution

(ii) The final three clusters

Q6 a) What are the three major areas in the data warehouse? Relate and explain the architectural components to the three major areas. [10]

b) The following table shows the time spent writing an essay and essay grades obtained for students in an English course. [10]

Hours spent on writing an essay	Grades
6	82
10	88
2	56
4	64
6	77
7	92
0	23
1	41
8	80
5	59
3	47

(i) Use the method of least squares to find an equation for the prediction of a student's essay grade based on the hours spent on writing an essay in the English course.

(ii) Predict the essay grade of a student who spent 2.35 hours on writing an essay in the English course.
