TE Sem V (R-2019 C. scheme) "AI & DS" May '2023

31/5/2023

Time: 3 hours

Max. Marks: 80

N.B. (1) Question one is Compulsory.
(2) Attempt any 3 questions out of the remaining.
(3) Assume suitable data if required.

Q. 1 (a) Every data structure in the data warehouse contains the time element. Why? 05

(b) 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 05

(c) What is Market basket analysis? 05

(d) Draw and explain KDD process. 05

Q. 2 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.

a) Draw a star schema diagram for the data warehouse.
b) Draw the base cuboid [date, spectator, location] and apply any four OLAP operations. 10

b) What is clustering? Explain K-mean clustering algorithm. Suppose that the data mining task is to cluster the following items into two clusters. {2, 4, 10, 12, 3, 20, 30, 11, 25, 56, 23}. Apply k-means algorithm. 10

Q.3 a) A database has five transactions. Let min sup = 50% and min conf = 70%.

T_id	Items
T100	a,b
T200	a,c,d
T300	e,c,a
T400	c,d,b
T500	a,c,d,b,e

Find all frequent itemsets and strong association rules using Apriori Algorithm. 10

b) What is data preprocessing? Explain different data cleaning techniques. 10

Q. 4 a) 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).

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

b) What is web mining? Explain HITS algorithm. 10

Q. 5 a) Explain with example multilevel association mining and multidimensional rule mining. 10

b) Clearly explain the working of DBSCAN algorithm using appropriate diagram. 10

Q.6 (a) Explain with example different data sampling techniques. 10

(b) Write short note on any 2 10

i. Differentiate between OLTP and OLAP

ii. Web Content mining

iii. Data Loading in ETL

****