53 (IT 701) DMDW

2017

DATA MINING AND DATA WAREHOUSING

Paper : IT 701

Full Marks : 100

Time : Three hours

The figures in the margin indicate full marks for the questions.

Answer ***any five*** *questions out of* ***seven.***

1. *(a)* How does data mining differ from Knowledge Discovery in Databases (KDD) ?

   (b) Explain the architecture of a data mining system with a suitable schematic diagram. 10+10

2. (a) Write a short note on association rule mining. Explain constraint based association mining.

(b) Use the two methods below to normalize the following group of data :

200, 300, 400, 600, 1000

- min-max normalization by setting min = 0 and max = 1
- 2-score normalization.

(c) What are the value-ranges of the following normalization methods ?

- min-max normalization
- 2-score normalization
- normalization by decimal scaling.

6+8+6

3. (a) Write and explain pseudo code for apriori algorithm.

(b) Consider a database, D, consisting 9 transactions. Suppose minimum support count is 2 (i.e. min-sup = 2/9 = 22%) and minimum confidence required is 70%.

TID	List of items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2, I3, I5
T100	I1, I2, I3

Find out frequent item set using Apriori algorithm. Explain each step with diagram. 10+10

4. Draw decision tree for the following data sets. Use entropy as a node selection mechanism : 20

Outlook	Temp (F)	Humidity	Windy	Class
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

5. *(a)* Differentiate between :
   - Hierarchical *vs* Partitioning clustering.

(b) Write short notes on following clustering methods :

- Density based method
- Grid based method
- Model based method.

10+10

6. (a) With a neat diagram explain the architecture of data warehouse. Explain the terms :

ROLAP, MOLAP, and HOLAP.

(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). 10+10

7. (a) Explain the *k*-means clustering algorithm.

(b) Assume the following data set is given (2, 2), (4, 4), (5, 5), (6, 6), (9, 9), (0, 4), (4, 0). It will have $k = 3$ cluster and the Manhattan distance is used as the distance function. The k-means initial clusters C1, C2, and C3 are as follows :

C1 : {(2, 2), (4, 4), (6, 6)}

C2 : {(0, 4), (4, 0), (1, 2)}

C3 : {(5, 5), (9, 9), (0, 3)}

Find the new cluster and their centroid after the first interaction of k-means. 10+10
