53 (IT 701) DMDW

2016

DATA MINING & DATA WAREHOUSING

Paper : IT 701

Full Marks : 100

Time : Three hours

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

Answer **all** *questions.*

1. *(a)* Describe the Numerosity reduction in data reduction and its related two methods Parametric and non-parametric in detail.   7

   (b) What are the four major features of data warehousing ? What is the significance of data mining in the data warehousing ?   7

(c) Define Covariance of numeric data and explain the backpropagation based neural network with a diagram. 6

2. *(a)* What is the role of smoothing in data cleaning and mention the techniques for data smoothing ? Given price value (in INR) such as 160, 40, 32, 60, 50, 122, 125, 142, 150, 121, 128 and 145. Apply the binning methods to partition the data (above mentioned). 6

(b) Use the given below methods to normalize the following group of data : 100, 400, 800, 1200, 1800, and 2000.

(i) min-max normalization by setting min$=0$ and max$=1$.

(ii) Z-score normalization using standard deviation.

(iii) Normalization by decimal scaling.

14

(a) Write down the differences between hierarchical and grid approach based clustering and also the types of these clustering's method. Define k-means algorithm with the pros and cons.

12

(b) Consider the given below *Figure* 1 for a given '∈' represented by the radius of the circles, and MinPts=3. Based on the above definitions, mention the label points that are in 'density-reachability' and 'density connectivity'.
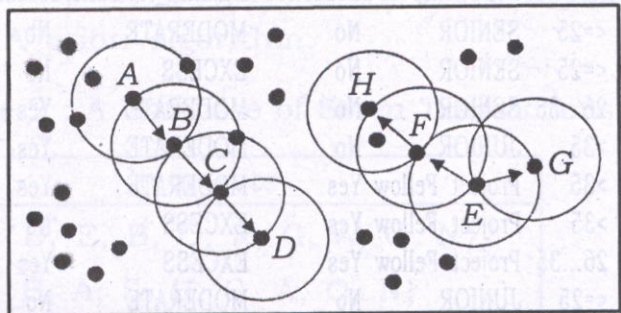

Figure 1 : Label points in DB SCAN Algo.

8

3. *(a)* Explain the classification and prediction methods, and describe the issues regarding classification and prediction.

6

(b) What is the limitation of Naive Bayesian Classification (NBC) and define the two methods to overcome this limitation of NBC ?   4

(c) Classify this training data (given below in Table 1) using Naive Bayesian Classification and predicts this dataset for "buys_Laptop=Yes", in this condition only ('Student : yes', Age : <=25, 'Fellowship : JUNIOR', 'Balance_amount : MODERATE')

Table 1 : TRAINING DATASET

Age	Fellowship	Student	Balance_Amount	Class : Buy_Laptop
<=25	SENIOR	No	MODERATE	No
<=25	SENIOR	No	EXCESS	No
26...35	SENIOR	No	MODERATE	Yes
>35	JUNIOR	No	MODERATE	Yes
>35	Project Fellow	Yes	MODERATE	Yes
>35	Project Fellow	Yes	EXCESS	No
26...35	Project Fellow	Yes	EXCESS	Yes
<=25	JUNIOR	No	MODERATE	No
<=25	Project Fellow	Yes	MODERATE	Yes
>35	JUNIOR	Yes	MODERATE	Yes
<=25	JUNIOR	Yes	EXCESS	Yes
26....35	JUNIOR	No	EXCESS	Yes
26...35	SENIOR	Yes	MODERATE	Yes
>35	JUNIOR	No	EXCESS	No

10

4. (a) What is the role of support and Confidence in Association-Rule mining ? 4

(b) Write down the apriori algorithm and mention the two key step in the implementation of apriori algorithm. 6

(c) A database has given below with seven transactions in the Table 2. Let min_sup=60% and min_conf=75%. Find all frequent item sets using A priori algorithm.

Table 2 : A Database of Seven Transactions

TID	Item sets
T100	{D, E, B, O, R, G, A, O, N}
T200	{B, A, S, U, G, A, O, N}
T300	{K, A, R, I, G, A, O, N}
T400	{B, O, N, G, A, I, G, A, O, N}
T500	{F, A, K, I, R, A, G, R, A, M}
T600	{K, O, K, R, A, J, H, A, R}
T700	{ B, E, S, U, R, G, A, O, N }

10

5. *(a)* True *or* False :

    (i) Data mining uses historical data to improve the decision.

    (ii) OLAP is a major task of traditional relational DBMS.

    (iii) OLTP is a major task of data warehouse system.

    (iv) Data analysis and decision making are processed in OLTP.

    (v) The operation of moving from finer-granularity data to a coarser granularity is called a drill down.

    (vi) The Roll-up operation navigates from less detailed data to more detailed data.

    (vii) The class labels of training data are unknown in supervised learning.

    (viii) In Unsupervised learning, the training data are accompanied by labels indicating the class of the observations.   4

(b) Match the "CLUSTERING APPROACH" with their "CLUSTERING METHODS".

CLUSTERING APPROACH	CLUSTERING METHODS
Hierarchical approach	BIRCH
Hierarchical approach	k-Medoids
Density approach	CLARANS
Density approach	CLIQUE
Grid-based approach	STING
Grid-based approach	AGNES
Partitioning approach	DBSACN
Partitioning approach	OPTICS

4

(c) Differentiate between : *(Any four)*

(i) Partitioning algorithm and sampling algorithm.

(ii) Roll-up and drill-down

(iii) Lazy learning and Eager-learning

(iv) Pre-pruning approach and post-pruning approach

(v) CURE and CHAMELEON. 12