THE SPRING END SEMESTER EXAMINATION-2013

4th Semester MCA (Regular & Back)

DATA WAREHOUSING AND MINING

[MCA 414]

[Regular-2011 Admitted batch & Back]

Full Marks: 60 **Time: 3 Hours**

Answer any six questions including question No. 1 which is compulsory.
The Figures in the margin indicate full marks.
Candidates are required to give their own words as far as practicable and all parts of a question should be answered at one place only. No marks awarded for extra questions.

1. a) Describe the steps involved in data mining when viewed as a process of knowledge discovery. [2x10]

   b) List and describe the primitives for specifying a data mining task.

   c) Briefly outline how to compute the dissimilarity between objects described by the numeric attributes?

   d) List different factors of data quality when preprocess the data for data warehousing.

   e) Define different data warehouse model from the architecture point of view.

   f) What is data loading? Explain refresh mode and update mode data loading.

   g) Describe, how to index OLAP data by bitmap indexing?

   h) Differentiate between density- reachable and density-connected in density based clustering.

   i) How does classification works? How it is different from prediction?

   j) Differentiate between support and confidence in association rules.

2. a) There are several disciplines which strongly influence the development of data mining methods. Justify your answer by defining all those disciplines/ technologies. [4]

b) Describe the major challenges to data mining research regarding data mining methodology and user interaction issues. [4]

3. Suppose that a data warehouse for *University- X consists* of the following four dimensions: *student, course, semester*, and *instructor*, and two measures *count* and *avg grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg_grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

a) Draw a *snowflake schema* diagram for the data warehouse. [3]

b) Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each *University* student. [2]

c) If each dimension has five levels (including all), such as "*student* < *major* < *status* < *university* < all", how many cuboids will this cube contain (including the base and apex cuboids)? [1]

d) Illustrate with suitable examples DMQL syntax for specifying the kind of knowledge (Association) to be mined. [2]

4. a) Describe DBSCAN clustering algorithm in terms of the following criteria [4]
   (i) Shapes of clusters that can be determined.
   (ii) Input parameters that must be specified
   (iii) limitations

b) Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strength and weakness of k- medoids in comparison with the k-means algorithm. [4]

5. a) The following table shows the midterm and final exam grades obtained for students in a database course. [6]

Midterm exam (X)	Final exam(Y)
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
33	77
34	52

i) Plot the data to specify X and Y seem to have a linear relationship?

ii) Predict the final exam grade of a student who received an 86 on the midterm exam.

b) How information gain is used in attribute selection measure while creating decision trees. [2]

6. a) Why data reduction technique? Describe any two dimensionality reduction strategies. [5]

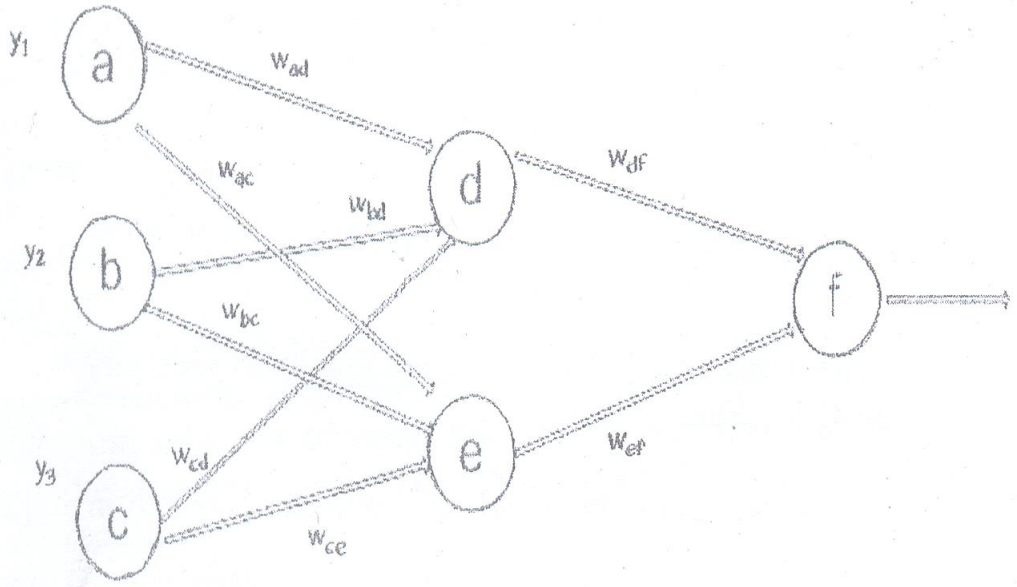
b) Discuss how can parametric and non-parametric numerosity reduction methods replace the original data volume by alternative smaller forms of data representation [3]

7. a) What are different views of different users regarding a data warehouse design? Discuss various approaches to the data warehouse design process and the steps involve in it. [4]

b) What is data mart? How do you differentiate between dependent and independent data mart? [2]

c) What is metadata repository? What it should contain? [2]

8. The following figure shows a multi layer feed-forward neural network. Let the learning rate be 0.8. The initial weight and bias values of the network are given in the table, along with the first training tuple Y=(1,1,0) , whose class label is 1. Using the back propagation algorithm update the weight and bias to get the class label. [8]



y_1	y_2	y_3	W_{ad}	W_{ae}	W_{bd}	W_{be}	W_{cd}	W_{ce}	W_{df}	W_{ef}	Θ_d	Θ_e	Θ_f
1	1	0	0.2	0.3	0.4	-0.1	-0.5	0.3	-0.2	-0.4	-0.2	0.2	0.1

XXXXX