IV B.Tech I Semester Supplementary Examinations, March 2013
DATA WAREHOUSING AND DATA MINING
(Computer Science & Engineering)

Time: 3 hours **Max Marks: 80**

Answer any FIVE Questions
All Questions carry equal marks
⋆ ⋆ ⋆ ⋆ ⋆

1. (a) Draw and explain the architecture of typical data mining system.
   (b) Differentiate OLTP and OLAP. [8+8]

2. Explain various data reduction techniques. [16]

3. (a) Briefly discuss about specifying the kind of knowledge to be mined.
   (b) Explain the syntax for specifying the kind of knowledge to be mined. [8+8]

4. Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order):
   13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70.

   (a) What is the mean of the data?
   (b) What is the median?
   (c) What is the mode of the data? Comment on the data's modality.
   (d) What is the mid range of the data?
   (e) Can you find (roughly) the first quartile(Q1),and third quartile(Q3) of the data?
   (f) Give the five number summaries of the data.
   (g) Show a box plot of the data.
   (h) How is the quantile-quantile plot different from a quantile plot? [16]

5. (a) Explain about constraint-based Association mining.
   (b) Give an example for Association rule mining? Classify Association rules.[8+8]

6. (a) Give the algorithm to generate a decision tree from the given training data.
   (b) Explain the concept of integrating data warehousing techniques and decision tree induction.
   (c) Describe multilayer feed-forward neural network. [8+4+4]

7. (a) Give an example of how specific clustering methods may be integrated, for example, where one clustering algorithm is used as a preprocessing step for another.
   (b) Write CURE algorithm and explain. [10+6]

8. (a) What kinds of association can be mined in multimedia data? What are the differences between mining association rules in multimedia databases versus transactional databases?

   (b) How does latent semantic indexing reduce the size of the term frequency matrix? Explain.

   (c) Describe the construction of a multilayered web information base.[3+3+6+4]

⋆⋆⋆⋆⋆

Code No: M0522/R07 **Set No. 2**

IV B.Tech I Semester Supplementary Examinations, March 2013
DATA WAREHOUSING AND DATA MINING
(Computer Science & Engineering)

Time: 3 hours **Max Marks: 80**

Answer any FIVE Questions
All Questions carry equal marks

⋆⋆⋆⋆⋆

1. Discuss the methods for the efficient implementation of data warehouse systems. [16]

2. Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order):
13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46, 52,70.

   (a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of the technique for the given data.

   (b) How might you determine outliers in the data?

   (c) What other methods are there for data smoothing?

   [16]

3. (a) Briefly describe the concept hierarchy specification.

   (b) Explain the syntax for concept hierarchy specification. [8+8]

4. Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order):
13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70.

   (a) What is the mean of the data?

   (b) What is the median?

   (c) What is the mode of the data? Comment on the data's modality.

   (d) What is the mid range of the data?

   (e) Can you find (roughly) the first quartile(Q1),and third quartile(Q3) of the data?

   (f) Give the five number summaries of the data.

   (g) Show a box plot of the data.

   (h) How is the quantile-quantile plot different from a quantile plot? [16]

5. When mining cross-level association rule, suppose it is found that the item set "IBM desktop computer, printer" does not satisfy minimum support. Can this information be used to prune the mining of a "descendent " item set such as "IBM desktop computer, b/w printer"? Give a general rule explaining how this information may be used for pruning the search space. [16]

6. The following table shows a set of paired data where X is the number of years of work experience of a college graduate and Y is the corresponding salary of the graduate.

X	Y
Years experience	Salary (in $ 1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

   (a) Plot the data. Do X and Y seem to have a linear relationship?

   (b) Use the method of least squares to find an equation for the prediction of the salary of a college graduate with some years of experience. [8+8]

7. (a) Given the following measurement for the variable age:
   16, 25, 28, 46, 29, 44, 38, 37, 54, 27
   Standardize the variable by the following:

      i. Compute the mean absolute deviation of age.

      ii. Compute the Z-score for the first four measurements.

   (b) Explain clustering using representatives algorithm with example.

   (c) Write an algorithm for DBSCAN and give an example of DBSCAN.[4+4+4+4]

8. (a) What is information retrieval? What methods are there for information retrieval?

   (b) What is sequential pattern mining? Explain.

   (c) Discuss about mining the webs link structures to identify authoritative web pages. [4+6+6]

⋆⋆⋆⋆⋆

Code No: M0522/R07 **Set No. 3**

IV B.Tech I Semester Supplementary Examinations, March 2013
DATA WAREHOUSING AND DATA MINING
(Computer Science & Engineering)

Time: 3 hours **Max Marks: 80**

Answer any FIVE Questions
All Questions carry equal marks

⋆ ⋆ ⋆ ⋆ ⋆

1. (a) What is data mining? What is data warehousing? Give their applications.
   (b) Briefly discuss data warehouse architecture. [8+8]

2. Write short notes on the following data reduction techniques:
   (a) Dimensionality reduction
   (b) Concept hierarchy generation for categorical data. [16]

3. The four major types of concept hierarchies are: schema hierarchies, set-grouping hierarchies, operation-derived hierarchies, and rule-based hierarchies.
   (a) Briefly define each type of hierarchy.
   (b) For each hierarchy type, provide an example. [16]

4. Write short notes for the following in detail:
   (a) Attribute-oriented induction.
   (b) Efficient implementation of Attribute-oriented induction. [8+8]

5. Propose and outline a level shared mining approach to mining multilevel association rules in which each item is encoded by its level position , and initial scan of the database collects the count for each item at each concept level, identifying frequent and sub frequent items. Comment on the processing cost of mining multilevel associations with this method in comparison to mining single level associations. [16]

6. (a) How scalable is decision tree induction? Explain.
   (b) Explain about prediction. [8+8]

7. The following table contains the attributes name, gender, trait-1, trait-2, trait-3, and trait-4, where name is an object-id, gender is a symmetric attribute, and the remaining trait attributes are asymmetric, describing personal traits of individuals who desire a penpal. Suppose that a service exists that attempt to find pairs of compatible penpals.

Name	gender	trair-1	trait-2	trait-3	trait-4
Kevan	M	N	P	P	N
Caroline	F	N	P	P	N
Erilk	M	P	N	N	P
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

For asymmetric attribute values, let the value P be set to 1 and the value N be set to 0. Suppose that the distance between objects (potential penpals) is computed based only on the asymmetric variables.

(a) Show the contingency matrix for each pair given Kevan, Caroline, and Erik.

(b) Compute the simple matching coefficient for each pair.

(c) Compute the Jaccard coefficient for each pair.

(d) Who do you suggest would make the best pair of penpals? Which pair of individuals would be the least compatible. [4+4+4+4]

8. (a) What is information retrieval? What methods are there for information retrieval?

(b) What is sequential pattern mining? Explain.

(c) Discuss about mining the webs link structures to identify authoritative web pages. [4+6+6]

⋆⋆⋆⋆⋆

Code No: M0522/R07

Set No. 4

IV B.Tech I Semester Supplementary Examinations, March 2013
DATA WAREHOUSING AND DATA MINING
(Computer Science & Engineering)

Time: 3 hours **Max Marks: 80**

Answer any FIVE Questions
All Questions carry equal marks

⋆⋆⋆⋆⋆

1. (a) Discuss about the classification of data mining systems
   (b) Discuss about Multidimensional data model. [8+8]

2. Write short note on the following data reduction techniques:
   (a) Data cube aggregation.
   (b) Concept hierarchy generation for categorical data. [16]

3. (a) List and describe any four primitives for specifying a data mining task.
   (b) Write about Semitight coupling and Loose Coupling. Differentiate them. [8+8]

4. (a) Briefly explain about data generalization.
   (b) Briefly explain about data summarization based characterization. [8+8]

5. A database has four transactions. Let min-sup=60% and min-conf=80%.

TID	Date	items-bought
T100	10/15/99	$\{K, A, D, B\}$
T200	10/15/99	$\{D, A, C, E, B\}$
T300	10/19/99	$\{C, A, B, E\}$
T400	10/22/99	$\{B, A, D\}$

   Find all frequent item sets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes. [16]

6. (a) Write an algorithm for k-nearest neighbor classification given k and n, the number of attributes describing each sample.
   (b) What is linear regression? Give an example of linear regression using the method of least squares. [8+8]

7. (a) Briefly outline how to compute the dissimilarity between objects described by the following types of variables:
      i. Asymmetric binary variables
      ii. Nominal variables
      iii. Ratio-scaled variables
      iv. Numeric (interval-scaled) variables
   (b) Explain about grid-based methods. [2+2+2+2+8]

8. A heterogeneous database system consists of multiple database systems that are defined independently, but that need to exchange transform information among themselves and answer global queries. Discuss how to process a descriptive mining query in such a system using a generalization-based approach. [16]

⋆⋆⋆⋆⋆