Code No: 07A70503 **R07** **Set No. 2**

IV B.Tech I Semester Examinations,MAY 2011
DATA WAREHOUSING AND DATA MINING
Computer Science And Engineering

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

Answer any FIVE Questions
All Questions carry equal marks

⋆⋆⋆⋆⋆

1. Explain the following terms in detail.
   (a) Concept description
   (b) Variance and Standard deviation.
   (c) Mean, median, and mode.
   (d) Quartiles, outliers, and boxplots. [16]

2. Define the terms: classification, prediction, decision tree, backpropagation, case-based reasoning, rough set approach, linear regression and multiple regression.[16]

3. Explain the following:
   (a) Spatial association analysis
   (b) Sequential pattern mining
   (c) Latent semantic indexing
   (d) Term frequency matrix. [4+4+4+4]

4. (a) Briefly discuss the various forms of Presenting and visualizing the discovered patterns.
   (b) Discuss about the objective measures of pattern interestingness. [8+8]

5. (a) Explain how a data warehouse is different from database
   (b) Discuss various steps involved in the design process of a data warehouse. [8+8]

6. (a) Explain about iceberg queries with example.
   (b) Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, explain with example. [8+8]

7. 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]

8. (a) Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8):
    i. Compute the Euclidean distance between the two objects.
    ii. Compute the Manhanttan distance between the two objects.
    iii. Compute the Minkowski distance between the two objects, using q=3.

   (b) Explain about Statistical-based outlier detection and Deviation-based outlier detection. [16]

⋆⋆⋆⋆⋆

Code No: 07A70503 **R07** **Set No. 4**

IV B.Tech I Semester Examinations,MAY 2011
DATA WAREHOUSING AND DATA MINING
Computer Science And Engineering

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

Answer any FIVE Questions
All Questions carry equal marks

⋆ ⋆ ⋆ ⋆ ⋆

1. (a) Briefly discuss about Task-relevant data specification.
   (b) Explain the syntax for Task-relevant data specification. [8+8]

2. (a) Briefly discuss about data integration.
   (b) Briefly discuss about data transformation. [8+8]

3. (a) Explain the construction of spatial data cube with suitable example.
   (b) What methods are there for information retrieval? Explain.
   (c) Describe web usage mining. [8+4+4]

4. (a) Attribute-oriented induction generates one or a set of generalized descriptions. How can these descriptions be visualized?
   (b) Discuss about the methods of attribute relevance analysis? [8+8]

5. (a) Define object-by-variable structure and object-by-object structure.
   (b) Explain representative object-based technique.
   (c) Write CURE algorithm and explain. [4+6+6]

6. (a) Describe three challenges to data mining regarding data mining methodology and user interaction issues.
   (b) Draw and explain the Three-tier architecture of a data warehouse. [8+8]

7. (a) Given a decision tree, you have the option of (i) converting the decision tree to rules and then pruning the resulting rules, or (ii) pruning the decision tree and then converting the pruned tree to rules. What advantages does former option have over later one. Explain.
   (b) Can any ideas from association rule mining be applied to classification? Explain. [8+8]

8. 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]

⋆ ⋆ ⋆ ⋆ ⋆

Code No: 07A70503 **R07** **Set No. 1**

IV B.Tech I Semester Examinations,MAY 2011
DATA WAREHOUSING AND DATA MINING
Computer Science And Engineering

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

Answer any FIVE Questions
All Questions carry equal marks

⋆ ⋆ ⋆ ⋆ ⋆

1. Write the algorithm for Apriori. Explain with a suitable example. [16]

2. Write the syntax for the following data mining primitives:

   (a) Task-relevant data.

   (b) Concept hierarchies. [16]

3. (a) Describe latent semantic indexing technique with an example.

   (b) Discuss about mining time-series and sequence data. [4+12]

4. (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]

5. (a) Differentiate between predictive and descriptive data mining.

   (b) State and explain algorithm for attribute-oriented induction. [8+8]

6. Write short notes on the following:

   (a) Efficient computation of data cube.

   (b) Indexing OLAP data.

   (c) Efficient processing of OLAP queries

   (d) Metadata repository. [16]

7. (a) Briefly discuss the data smoothing techniques.

   (b) 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.

      i. 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.

      ii. How might you determine outliers in the data?

      iii. What other methods are there for data smoothing? [16]

8. 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]

⋆⋆⋆⋆⋆

Code No: 07A70503 **R07** **Set No. 3**

IV B.Tech I Semester Examinations,MAY 2011
DATA WAREHOUSING AND DATA MINING
Computer Science And Engineering

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

Answer any FIVE Questions
All Questions carry equal marks

⋆⋆⋆⋆⋆

1. (a) Discuss about binary, nominal, ordinal, and ratio-scaled variables.

   (b) Explain about grid-based methods. [8+8]

2. (a) Discus about Association rule mining.

   (b) Define multidimensional Association rule. Discuss mining distance-based Association rules. [8+8]

3. (a) What is Concept description? Explain.

   (b) What are the differences between concept description in large data bases and OLAP? [8+8]

4. (a) Can any ideas from association rule mining be applied to classification? Explain.

   (b) Explain training Bayesian belief networks.

   (c) How does tree pruning work? What are some enhancements to basic decision tree induction? [6+5+5]

5. (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]

6. Write short notes on the following data reduction techniques:

   (a) Dimensionality reduction

   (b) Concept hierarchy generation for categorical data. [16]

7. Briefly discuss the following data mining primitives:

   (a) The kind of knowledge to be mined

   (b) Background knowledge

   (c) Interestingness measures

   (d) Presentation and visualization of discovered patterns. [16]

8. (a) Draw and explain the architecture of typical data mining system.

(b) Differentiate OLTP and OLAP. [8+8]

⋆⋆⋆⋆⋆