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

IV B.Tech I Semester Examinations,December 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 mining frequent item sets without candidate generation.
   (b) Explain about multidimensional Association rules in detail. [8+8]
2. Write the syntax for the following data mining primitives:
   (a) The kind of knowledge to be mined.
   (b) Measures of pattern interestingness. [16]
3. (a) Explain data mining as a step in the process of knowledge discovery.
   (b) Differentiate operational database systems and data warehousing. [8+8]
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) 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]
6. (a) What is spatial data warehouse? What are the different types of dimensions in a spatial data cube? What are the different types of measures in a spatial data cube?
   (b) What is keyboard-based association analysis? How can automated document classification be performed?
   (c) Briefly discuss about mining the World Wide Web. [2+2+2+2+2+6]
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 then 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) 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]

⋆ ⋆ ⋆ ⋆ ⋆

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

IV B.Tech I Semester Examinations,December 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. Briefly discuss the Discretization and concept hierarchy techniques. [16]

2. (a) What is Cluster Analysis? What are some typical applications of clustering? What are some typical requirements of clustering in data mining?
   (b) Discuss about model-based clustering methods. [2+2+5+7]

3. (a) Explain the design and construction process of data warehouses.
   (b) Explain the architecture of a typical data mining system. [8+8]

4. Write the FP-growth algorithm for discovering frequent item sets without candidate generation. Explain an example. [16]

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

6. (a) How can object identifiers be generalized, if their role is to uniquely identify objects? Can inherited properties of objects be generalized.
   (b) What kinds of association can be mined in multimedia data? Explain.
   (c) Describe similarity search in time-series analysis. [4+6+6]

7. (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]

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

⋆ ⋆ ⋆ ⋆ ⋆

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

IV B.Tech I Semester Examinations,December 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:
   (a) Mining spatial databases
   (b) Mining the World Wide Web. [8+8]
2. (a) Explain about multilevel Association rules from transaction databases.
   (b) What are the steps involved in Association rule clustering system? Explain. [8+8]
3. (a) How can you go about filling in the missing values in data cleaning process?
   (b) Discuss the data smoothing techniques. [8+8]
4. (a) Discuss the importance of establishing a standardized data mining query language. What are some of the potential benefits and challenges involved in such a task?
   (b) How can we standardize data mining primitives? [8+8]
5. Briefly discuss about the following data warehouse implementation methods:
   (a) Indexing OLAP data
   (b) Metadata Repository. [16]
6. (a) Why naive Bayesian classification called "naive"? Briefly outline the major ideas of naive Bayesian classification.
   (b) Define regression. Briefly explain about linear, non-linear and multiple regressions. [8+8]
7. Suppose that you are to allocate a number of automatic teller machines (ATMs) in a given region so as to satisfy a number of constraints. Households or places of work may be clustered so that typically one ATM is assigned per cluster. The clustering, however, may be constrained by factors involving the location of bridges, rivers, and highways that can affect ATM accessibility. Additional constraints may involve limitations on the number of ATMs per district forming the region. Given such constraints, how can clustering algorithms be modified to allow for constraint-based clustering? [16]
8. (a) How can we perform discrimination between different classes? Explain.
   (b) Explain the analytical characterization with an example. [8+8]

⋆ ⋆ ⋆ ⋆ ⋆

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

IV B.Tech I Semester Examinations,December 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:
   (a) Centroid-based technique
   (b) Representative object-based technique
   (c) OPTICS
   (d) Deviation-based outlier detection. [4+4+4+4]
2. (a) Explain similarity search in multimedia data.
   (b) Explain similarity search in time-series analysis.
   (c) What is meant by authoritative web pages? Explain about mining the webs link structures to identify authoritative web pages. [5+6+5]
3. Write short note on the following architectures of data mining systems:
   (a) No coupling
   (b) Loose coupling
   (c) Semitight coupling
   (d) Tight coupling. [16]
4. (a) How can we perform discrimination between different classes? Explain
   (b) Briefly explain about data summarization based characterization. [8+8]
5. Which algorithm is used for discovering frequent item sets without candidate generation. Explain with an example. [16]
6. (a) Draw and explain the architecture for on-line analytical mining.
   (b) Briefly discuss the data warehouse applications. [8+8]
7. The following table shows the midterm and final exam grades obtained for students in a database course:

X Midterm exam	Y Final exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

(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 a student's final exam grade based on the student's midterm grade in the course. [8+8]

8. 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]

⋆ ⋆ ⋆ ⋆ ⋆