# Model Question Paper

## Seventh Semester B. Tech. Degree Examination

(2013 scheme)

Computer Science & Engineering

**13.706.2 DATA MINING AND INFORMATION RETRIEVAL (FR)**

Time : 3 hours

Max Marks : 100

**PART A**

(Answer all questions. Each question carries 4 marks )

**1.** How is data warehouse different from a database? How they are similar?

**2.** What do you mean by dimensionality reduction? How significant this operation in data mining?

**3.** What are demerits of Apriori algorithm? Write any one technique address the problem.

**4.** Define the hyper plane. What is maximum margin hyper plane?

**5.** Let ($x_1$, $x_2$, .....$x_n$) are set elements in a cluster. Write equation to find Centroid and Radius of the cluster

**PART B**

(Answer one full question from each module)

**Module I**

**6 a.** Describe three-tier data warehousing architecture (12)

**b**. Discuss issues to be consider during data integration (8)

**OR**

**7 a.** In Data ware house technology, a multiple dimensional view can be implemented by a relational database technique (ROLAP), or by a multidimensional database technique (MOLAP) or by a hybrid database technique (HOLAP). Briefly describe each implementation technique (12)

**b.** Consider Sales data with dimensions LocationID, TimeID, ProductID. Explain Roll-up and Drilldown with respect to sales data (8)

**Module II**

**8 a.** Assume we wish to find whether play is possible or not on a particular day by building a decision tree. The properties to be considered are Temperature, Humidity and Wind. Use ID3 algorithm and find the best attribute to split at the first level

| Temperature | Humidity | Windy | Play |
|---|---|---|---|
| Hot | high | FALSE | no |
| Hot | normal | TRUE | no |
| Hot | normal | FALSE | yes |
| Mild | high | FALSE | yes |
| Cool | low | FALSE | yes |
| Cool | normal | TRUE | no |
| Cool | normal | FALSE | yes |
| Mild | high | TRUE | no |
| Cool | normal | FALSE | yes |
| Mild | low | FALSE | yes |
| Mild | normal | TRUE | yes |
| Hot | normal | TRUE | no |

Hot low TRUE no

(14)

**b.** Suppose that 1000 people attended a disease prediction test. Among 300 patients having heart related disorders, 280 of them tested positive, 20 tested negative. Among the 700 people, without having any heart diseases, 685 tested negative and 15 tested positive. Find accuracy, precision, recall and specificity (6)

**OR**

**9 a.** Explain back propagation algorithm in Artificial Neural Networks (ANN) (10)

b. How the data that are linearly inseparable be classified using Support Vector Machines (10)

**Module III**

**10 a.** What do you meant by hierarchical clustering? How is it represented? What are differences between single link and complete link algorithms (10)

**b.** Write K means algorithm and separate {5, 11, 19, 27, 23, 25, 6, 18, 2, 8, 10, 12, 31, 29, 4} into 3 clusters (10)

**OR**

**11 a.** What are density based clustering methods? How DBSCAN algorithm works? (10)

**b.** Discuss the parameters to evaluate results of a clustering method? (10)

**MODULE IV**

**12 a.** Consider the following transactional database, with set of items I={I1, I2, I3, I4, I5}. Let minimum support is 40% and confidence is 60%. Find all frequent itemsets using Apriori and FP-growth. Compare efficiency of two mining process. (15)

| TID | List of items |
|---|---|
| T1 | I1, I2, I5 |
| T2 | I2, I4 |
| T3 | I2, I3 |
| T4 | I1, I2, I4 |
| T5 | I1, I3 |
| T6 | I2, I3 |
| T7 | I1, I3 |
| T8 | I1, I2, I3, I5 |
| T9 | I1, I2, I3 |
| T10 | I2, I4, I5 |

b. What do you meant by constraint based association mining (5)

**OR**

**13 a.** Explain how the spatial data structures Quad Tree, R-Tree and KD Tree differs ? (12)

**b.** What do you mean by temporal mining (8)