140

Thapar Institute of Engineering & Technology, Patiala
School of Mathematics & Computer Applications
CA-043(Data Mining & Data Warehousing)
End Semester Examination (13/12/2006)

Max Marks: 45 **Time: 3 Hours**

Note: Attempt any FIVE questions.
Attempt all parts of a question together at one place in a given sequence otherwise only first attempted part(s) will be evaluated.
Make assumptions, if missing, suitably with reasoning.

1(a) Describe the steps involved in data mining when viewed as a process of knowledge discovery

(b) Why do you need a separate data staging component.

(c) Distinguish between operational and informational systems.

(d) Explain the formal definition of a data warehouse. **(3, 3, 2, 1)**

2(a) What do you consider to be a core set of team roles for a data warehouse project. Describe the responsibilities of three roles from your set.

(b) Define each class of schema used for designing data warehouse using DMQL syntax. Consider a real life example to explain the design phase.

(c) Explain the difference between the top-down and bottom-up approaches for building data warehouses. Do you have a preference? If so, Why?

(3, 3, 3)

3(a) What is clustering? How it differs from classification. Describe various requirements of clustering in data mining

(b) Describe each of the following clustering algorithms in terms of following criteria:

(i) Shapes of clusters that can be created. (ii) Algorithmic approach

(a) DBSCAN (b) K-Means (c) Agglomerative hierarchical clustering **(3, 6)**

4(a) What do you mean by Entropy and information gain. How information gain helps in classification process?

(b) Construct decision tree from following training data using ID3 algorithm.

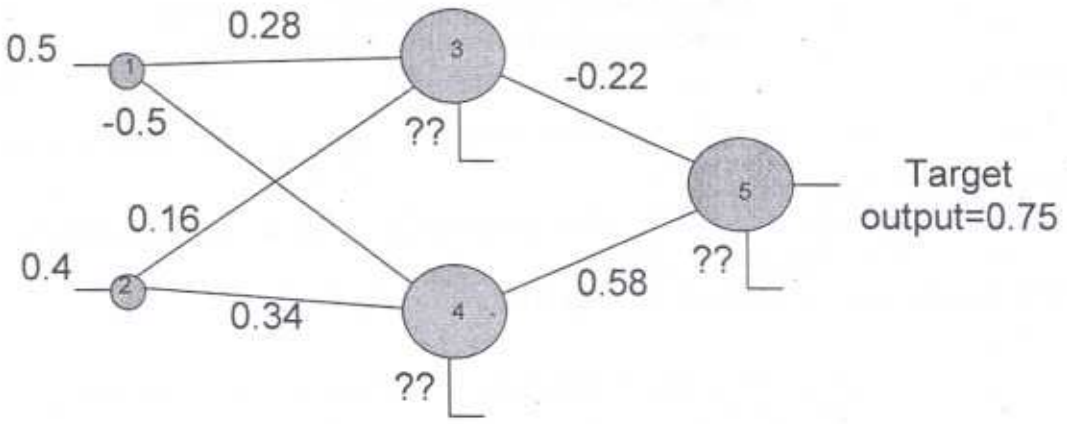
Outlook	Temperature	Windy	Marketing
sunny	hot	false	yes
sunny	hot	true	yes
rainy	cool	true	no
rainy	hot	false	yes
sunny	cool	true	no
rainy	hot	true	yes
sunny	cool	false	yes
rainy	cool	false	no

(c) Explain the following terms:
(a) Gain Ratio (b) Gini Index **(2, 5, 2)**

P. T. O.

Office

5(a) Explain classification of feed forward neural network by back propagation and apply back propagation algorithm to train the following system:



Given learning rate, $\eta = 0.85$

$$\theta_3 = 0.36,\ \theta_4 = 0.25,\ \theta_5 = -0.48$$

Solve the given problem up to two iterations and find the error at output node and hidden nodes. Als modify the weights and bias to train the system.

(b) Illustrate various components of neural network. **(6, 3)**

6(a) Write down Apriori frequent/candidate generation algorithm. Apply Apriori algorithm on following data:

TID	Item sets
T_{100}	Sugar, Butter, Tea, Cheese, Milk, Groundnuts
T_{200}	Milk, Tea, Mustard oil, Bread, Butter
T_{300}	Sugar, Tea, Duster, Maze,Milk
T_{400}	Milk, Tea, Sugar, Bread, Butter
T_{500}	Bread, Butter, Jam, Sugar, Milk
T_{600}	Groundnuts, Bread, Mustard oil, Jam

(i) to generate frequent set **at each level** when **minsupp = 2.**
(ii) to generate association rules **at each level** when **minconf =2.**

(b) Construct FP Tree and generate the frequent sets and association rules for the above problem.
(2, 3.5, 3.5)

7(a) The following table shows the midterm and final exam grades obtained for students in a database course.

x (Midterm exam):	72	50	81	74	94	86	59	83
y (Final exam) :	84	63	77	78	90	75	49	79

Use the method of least square to find linear relationship for the prediction of a student's final grade based on the student's midterm grade in the course.

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

(c) What is 3-4-5 rule? Describe with the help of 3-4-5 rule, the concept hierarchy generation by intuitive partitioning.
(4, 1, 4)

Evaluated answer sheets will be shown on 17 December 2006 (Sunday), at 2:30 P.M. in room B-210.