GANPAT UNIVERSITY

M. Tech SEMESTER-II Computer Engineering
REGULAR EXAMINATION MAY 2014
3CE203: Data Mining & Data Warehousing

Time: 3 Hours] **[Total Marks: 70**

Instructions:

1. Figures to the right indicate full marks
2. Each section should be written in a separate answer book
3. Be precise and to the point in your answer

SECTION-I

Q-1 Answer the following:

(a) What is naïve Bayesian classifier? With the same classifier prove that the dataset given in Figure 1 never predicts *buys computer* = *"no"* for tuple *X*. [06]

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Figure 1

(b) What is gain ratio? Construct classification tree for the data as shown in Figure 1 by using the maximum gain ratio for selection of an attribute. [06]

OR

Q-1 Answer the following:

(a) The following table consists of training data from an employee database. The data have been generalized. For example, "31 : : : 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. (Refer Figure 2). [06]

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Figure 2

Let status be the class label attribute.

(i) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

(ii) Use your algorithm to construct a decision tree from the given data.

(iii) Given a data tuple having the values "systems," "26. . . 30," and "46–50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?

(b) Write an algorithm for Naïve Bayesian Classification. [06]

Q-2 Answer the following:

(a) The data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 14, 16, 17, 17, 20, 21, 21, 22, 23, 23, 26, 26, 26, 26, 31, 34,34, 36, 36, 36, 36, 37, 41, 46, 47, 53, 71. [06]

(i) Find the *mean & median* of the data

(ii) Find the *mode* of the data?

(iii)Find the *midrange* of the data?

(iv)By using binning method, smooth above data.

(b) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): [06]

(i) Compute the Euclidean distance between the two objects.

(ii) Compute the Manhattan distance between the two objects.

(iii) Compute the Minkowski distance between the two objects, using q = 3.

Q-3 Answer the following:

(a) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each [05] category having its own charge rate.

(a) Draw a star schema diagram for the data warehouse.

(b) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2014?

(b) A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer. [06]

OR

Q-3 **Answer the following:**

(a) With an example you show that the Apriori algorithm uses prior knowledge of subset support properties. Prove that all nonempty subsets of a frequent itemset must also be frequent. [05]

(b) A database has five transactions. Let *min sup* = 20% [06]

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Find all frequent itemsets using Apriori and FP-growth, respectively.

SECTION-II

Q-4 Answer the following:

(a) Explain the role of data Mining for Biological Data Analysis. [06]

(b) Explain the data preprocessing techniques with an example. [06]

OR

Q-4 Answer the following:

(a) Explain clustering of a set of objects based on the k-means method. [06]

(b) What is the use of data mining in various fields? Discuss in detail. [06]

Q-5 Answer the following:

(a) Explain major issues in data mining. [06]

(b) Discuss three tire architecture of Data Warehousing [06]

Q-6 Answer the following:

(a) Explain the cube operations in Data Warehousing. [05]

(b) Show the diagram and discuss the data mining as a confluence of multiple disciplines. [06]

OR

Q-6 Answer the following:

(a) Show & Explain the diagram of KDD process. [05]

(b) Explain the typical framework of a data warehouse with an example. [06]

End of Paper