GANPAT UNIVERSITY

M. Tech SEMESTER-II Computer Engineering
REGULAR EXAMINATION JULY 2013
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: **[12]**

(a) Prove that the naïve Bayesian classifier predicts *buys computer = "yes"* for tuple *X* data given in Figure 1.

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) Select the attribute with the maximum gain ratio as a splitting attribute and construct the classification tree for the data as shown in figure 1.

OR

Q-1 Answer the following: **[12]**

(a) A software engineer at *University* is to design a data mining system to examine the university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and the cumulative grade point average (GPA). Describe the *architecture* you would choose. What is the purpose of each component of this architecture?

(b) Let *A* be the splitting attribute. *A* has *v* distinct values,{*a*1, *a*2,: : : , *av*}, based on the training data. Show the three possibilities for partitioning tuples based on the splitting criterion with example.

Q-2 Answer the following: **[12]**

(a) The data for analysis includes 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) What is the *mean* of the data?What is the *median*?

(ii) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal,trimodal, etc.).

(iii) What is the *midrange* of the data?

(iv) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data?

(v) Give the *five-number summary* of the data.

(b) A data warehouse consists of the three dimensions *time, doctor*, and *patient*, and the two measures *count* and *charge*, where *charge* is the fee that a doctor charges a patient for a visit.

(a) Enumerate three classes of schemas that are popularly used for modeling data warehouses.

(b) Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).

(c) Starting with the base cuboid [*day*, *doctor*, *patient*], what specific *OLAP operations* should be performed in order to list the total fee collected by each doctor in 2004?

(d) To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema *fee* (*day, month, year, doctor, hospital, patient,count, charge*).

Q-3 Answer the following: **[11]**

(a) The data for analysis includes 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. **[05]**

(i) Use *smoothing by bin means* to smooth the data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

(ii) How might you determine *outliers* in the data?

(iii) What other methods are there for *data smoothing*?

(b) A data warehouse for *University* consists of the following four dimensions: **[06]**

student, course, semester, and *instructor*, and two measures *count* and *avg grade*.

When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the average grade for the given combination.

(a) Draw a *snowflake schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each *Big University* student.

OR

Q-3 Answer the following: **[11]**

(a) The Apriori algorithm uses *prior knowledge* of subset support properties. **[05]**

(a) Prove that all nonempty subsets of a frequent itemset must also be frequent.

(b) Prove that the support of any nonempty subset *s*0 of itemset *s* must be at least as great as the support of *s*.

(b) A database has five transactions. Let *min sup* = 60% and *min conf* = 80%. **[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) What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)? [12]

(b) Define each of the following *data mining functionalities*: Preprocessing, Association. Give examples of each data mining functionality, using a real-life database.

OR

Q-4 Answer the following: [12]

(a) Write the steps of working of Naïve Bayesian Classification.

(b) Give the applications of Data Mining in detail.

Q-5 Answer the following: [12]

(a) Explain the STAR schema with an example and query.

(b) Discuss three tire architecture of Data Warehousing

Q-6 Answer the following:

(a) Explain the cube operations in Data Warehousing. [11]

(b) Show the diagram and discuss the data mining as a confluence of multiple disciplines. [05]

[06]

OR

Q-6 Answer the following:

(a) Explain the preprocessing steps of Data Mining [11]

(b) Explain major issues in data mining. [05]

[06]

End of Paper