Dr. Pradeep K Singh

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-3 EXAMINATION- JUNE -2016

B.Tech (IT), 6th Semester

COURSE CODE: 10B22CI622 MAX. MARKS: 35

COURSE NAME: Data Mining

COURSE CREDITS: 4 MAX. TIME: 2 HRS

Note: *All questions are compulsory. Carrying of mobile phone during examinations will be treated as case of unfair means. Assumption may taken if necessarily required.*

Q.1 Why concept hierarchies are useful in data mining? **[2 Marks]**

Q.2 How does the discordancy testing work in context to Statistical Distribution-Based Outlier Detection? Name at least two types of procedures used for detecting outlier. **[2 Marks]**

Q.3 Let us suppose that data set I have selected for analysis is HUGE, which is sure to slow down the mining process. Is there any way I can reduce the size of my data set, without the jeopardizing the data mining results? **[3 Marks]**

Q.4 Use the two methods below to *normalize* the following group of data:
200; 400; 600; 800; 1000
(a) min-max normalization by setting ***min*** **= 0 and** ***max*** **= 1**
(b) z-score normalization **[3 Marks]**

Q.5 How OLTP and OLAP perform differently in context to data contents, database design, View, Access Patterns and Number of Users. **[3 Marks]**

Q.6 Imagine that you need to analyze AllElectronics sales and customer data. You note that many tuples have no recorded value for several attributes, such as customer income. How can you go about filling in the missing values for this attribute? **[3 Marks]**

Q.7 Why is *tree pruning* useful in decision tree induction? What is the drawback of using a separate set of tuples to evaluate pruning? **[3 Marks]**

Q.8 A database has four transactions. Let *min sup* = 60% and *min conf* = 80%.

Cust_ID	TID	Items_Bought (In the form of brand-item-category)
01	T100	{Kings-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread }
02	T200	{ Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie,Wonder-Bread}
01	T300	{ Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie }
03	T400	{ Wonder-Bread, Sunset-Milk, Dairyland-Cheese }

(a) At the granularity of *item category* (e.g., *itemi* could be *Milk"*), for the following rule template, *8X 2 transaction; buys(X; item*1) ^ *buys(X; item*2) *) buys(X; item*3) [*s; c*] list the frequent *k*-itemset for the largest *k*, and *all* of the *strong* association rules (with their support *s* and con¯dence *c*) containing the frequent *k*-itemset for the largest *k*.

(b) At the granularity of *brand-item category* (e.g., *itemi* could be *Sunset-Milk"*), for the following rule template, *8X 2 customer; buys(X; item*1) ^ *buys(X; item*2) *) buys(X; item*3) list the frequent *k*-itemset for the largest *k* (but do not print any rules). **[6 Marks]**

Q.9 How we choose a data mining system among the many data mining systems for a specific problem? Name the criteria's adopted in selection of data mining systems. Discuss some of the challenges brought by emerging scientific applications of data mining. **[4 Marks]**

Q.10. Consider the class labeled tuples from the AllElectronics customer database as follows;

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

Predict a class level using Naïve Bayesian Classification. **[6 Marks]**

----------END----------