CAIRO UNIVERSITY

Institute of Statistical Studies and Researches

Course: CS613 (Data Mining) Date: 18th June, 2011
Instructors: Dr. Khaled Wassif Final Exam Allowed Time: THREE Hours

Attempt ALL Questions:

Q1. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

Q2. Noise is a random error or variance in a measured variable. What are the main reasons of and how to handle noisy data?

Q3. Explain the main primitives for specifying a data mining task.

Q4. What tasks should be considered in the design of GUIs based on a data mining query language?

Q5. Given the huge amount of data in a database, it is highly preferable to update data mining result incrementally rather than mining from scratch on each database update. Describe how incremental data mining is done in case of data insertion and data deletion.

Q6. Suppose that the following table is derived by attribute-oriented induction:

class	Birth_place	count
System Analyst	Egypt	180
System Analyst	Others	120
Programmer	Egypt	20
Programmer	Others	80

a) Transform the table into a crosstab showing the associated t-weights and d-weights.

b) Map the class Programmer into a quantitative descriptive rule.

Q7. A database has five transactions. Let *min-sup* = 60% and *min-conf* = 80%

TID	*Date*	*Items_bought*
T100	8/6/2010	{D, A, K, B}
T200	18/6/2010	{E, A, C, D, B}
T300	15/10/2010	{C, A, B, E}
T400	22/10/2010	{B, A, D}
T500	10/12/2010	{C, D, K}

a) Find all frequent itemsets using Apriori algorithm.

b) List all of the strong association rules matching the following metarule:

$$buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)\ [support, confidence]$$

where *X* is a variable representing customers, and *item$_i$* denotes variables representing items.

Q8. The following table summarizes supermarket transaction data:

	Hotdogs	*not Hotdogs*	Sum (row)
Hamburgers	2000	500	2500
not Hamburgers	1000	1500	2500
Sum (column)	3000	2000	5000

where *Hotdogs* refers to the transactions containing hot dogs, *not Hotdogs* refers to the transactions that do not contain hot dogs, *Hamburgers* refers to the transactions containing hamburgers, and *not Hamburgers* refers to the transactions that do not contain hamburgers.

a) Suppose that the association rule "*hot dogs* ⇒ *hamburgers*" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?

With My Best Wishes

CAIRO UNIVERSITY

Institute of Statistical Studies and Researches

Course: CS613 (Data Mining)		Date: June, 2015
Instructors: Dr. Khaled Wassif	Final Exam	Allowed Time: THREE Hours

Attempt ALL Questions:

Q1. What is data mining? Describe the steps involved in data mining when viewed as a process of knowledge discovery. (8 Marks)

Q2. Suppose your task as a software engineer at Big-University is to design a data mining system to examine their 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 their cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture? (6 Marks)

Q3. Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, prediction, clustering, and evolution analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with. (12 Marks)

Q4. What is noisy data? What are the main reasons of noisy data? How can handle noisy data? (8 Marks)

Q5. Give a short example to show that items in a strong association rule may actually be negatively correlated. (6 Marks)

Q6. A database has five transactions. Let min_sup = 60% and min_conf = 80%. (15 Marks)

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 }

a) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

b) List all of the strong association rules matching the following metarule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g., "A", "B", etc.):

$$buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)\ [support, confidence]$$

Q7. The following table presents a training set of class-labeled tuples randomly selected from the AllElectronics customer database. The class label attribute, buys computer, has two distinct values (namely, yes, no): **(15 Marks)**

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

a) Create the classification model of the previous training set using *decision tree induction* (use information gain as the attribute selection measure).

b) Extract classification rules from the decision tree.

Q8. The following table shows the midterm and final exam grades obtained for students in data mining course:

Midterm Exam	72	50	81	74	94	86	59	83	65	33	88	81
Final Exam	84	63	77	78	90	75	49	79	77	52	74	90

Predict the final exam grade of a student who received a 90 on the midterm exam. **(10 Marks)**

With My Best Wishes