The Hong Kong Polytechnic University
Department of Computing
COMP5121 Data Mining and Data Warehousing

Assignment 2
Due Date: October 28, 2011

1. Negative Association Rules are association rules that have either a negative antecedent or consequent (not A → B or A → not B). In mining a transaction database, for example, one may discover that "30% of the people who buy diapers do not buy beer" or "30% of the people who do not buy diapers buy beer", etc. These negative association rules may sometimes provide a useful source of information for various applications. Read the two articles attached:

   a) "Mining for Strong Negative Associations in a Large Database of Customer Transactions"
   b) "Mining Positive and Negative Association Rules: An Approach for Confined Rules"

   Write a brief report of about 1,000 words of what you have learnt and/or found in this topic. In your report, explain, among other things that you have read, what negative association rules are, how they should be mined and how they can be used.

2. PolyMobile is a mobile phone service provider. It provides different package plans to customer. In viewing the rate of loss of customers is going high. PolyMobile would like to find out what kind of customers may have a higher chance of leaving PolyMobile to join another carrier.

   a) The table below shows 20 records (*training dataset*) retrieved from its customer database. Find a decision tree, using ID3 without any data mining tool, to determine if a customer would renew his or her contract with PolyMobile.

Table 1: Training Dataset

Item	Sex	Age	Married	Income	Plan	Renew Contract
1	Female	Young	YES	High	A	Yes
2	Male	Middle	NO	High	A	Yes
3	Female	Senior	NO	Middle	A	Yes
4	Male	Senior	NO	Middle	C	Yes
5	Female	Senior	YES	Middle	C	Yes
6	Female	Young	YES	Low	B	No
7	Female	Young	NO	High	C	No
8	Female	Young	YES	Low	B	Yes
9	Male	Young	NO	Low	A	No
10	Female	Middle	YES	Low	C	Yes
11	Female	Middle	YES	Low	C	No
12	Female	Middle	YES	High	B	Yes
13	Male	Young	YES	High	B	Yes
14	Male	Senior	NO	Middle	A	No
15	Female	Senior	YES	Middle	A	Yes

Item	Sex	Age	Married	Income	Plan	Renew Contract
16	Female	Young	NO	Middle	C	Yes
17	Male	Young	NO	High	C	No
18	Male	Young	YES	High	B	Yes
19	Male	Middle	YES	Middle	B	Yes
20	Female	Senior	YES	Low	A	No

b) You are given 5 more records below. Using these 5 records, discuss how much you should trust the decision tree that you have found in a)?

Table 2: Testing Dataset

Item	Sex	Age	Married	Income	Plan	Renew Contract
1	Female	Senior	No	Middle	B	Yes
2	Female	Senior	Yes	High	C	No
3	Male	Young	Yes	Low	C	Yes
4	Female	Middle	No	Middle	A	No
5	Male	Young	Yes	High	B	Yes

c) Use the C5.0 in PASW Modeler to obtain a classification model based on the data in Table 1. Compare your results with those you obtained in a) and b). What are the differences? Can you explain why there are such differences?

d) Repeat a) and b) using the Naïve Bayesian Approach.

e) Given a choice among ID3, PASW and Naïve Bayesian Approach for this task, which one would you choose? Why?

3. Assume that you are a sales manager for a telecommunication company and that your company has a mobile phone operation. Due to the introduction of a special subscription plan by your competitor, you have lost some customers to them within the first several days. In order to prevent further loss, your manager would like you to take a close look at the following data set sampled from those customers who responded to a survey. In the survey, the customers were asked if they would remain with the current subscription plan or switch to the competitor's.

Customer No.	Average Monthly Payment	Average Duration of Calls	Total Calling Time	Decision
1	215.16	6.84	41.63	Stay
2	320.32	14.76	97.25	Switch
3	218.24	3.96	124.88	Stay
4	352.00	8.04	56.00	Undecided
5	259.16	11.88	82.50	Stay
6	462.44	2.64	162.38	Undecided
7	220.66	2.16	69.75	Switch
8	214.72	8.04	96.88	Switch
9	373.78	9.24	122.50	Undecided
10	409.86	17.28	180.50	Switch
11	378.62	18.60	83.50	Undecided
12	195.36	5.88	138.88	Switch

Customer No.	Average Monthly Payment	Average Duration of Calls	Total Calling Time	Decision
13	291.94	9.00	111.13	Stay
14	342.10	12.00	96.38	Stay
15	317.68	10.68	126.38	Switch
16	197.54	4.4	147.3	Stay
17	404.43	17.4	171.7	Switch
18	264.11	5.7	107.1	Switch
19	430.44	3.14	100.74	Undecided
20	273.43	8.7	98.7	Undecided

a) Assume that k = 5, using the k-NN algorithm, what do you expect the decision of a customer, who has an average monthly payment of 293.26, an average duration of calls of 6.96 and a total calling time of 110.25, to be?

b) Assume again that k = 5 and ignoring the "decision" of the customers. Using the k-NN algorithm, what do you expect the average duration of calls of a customer to be, given that his average monthly payment of 271.48 and a total calling time of 114.00?

c) If you are free to choose the value of k, what will your choice be? Why?

4. A database was collected at Chiba University hospital. Some patients who visited the outpatient clinic of the hospital had collagen diseases. They were recommended the hospital by a family physician in a private clinic nearby. For collagen diseases, *thrombosis* is one of the most important and severe complications and one of the major causes of death. Thrombosis is an increased coagulation of blood that clogs blood vessels.

   Thrombosis must be treated as an emergency and it is important to detect and predict the possibilities of its occurrence. An analysis of historical data may help with this task. For such purposes, a well-known physician and medical researcher decided to donate some datasets that he had previously collected from treating his own patients. He hopes that regularities can be discovered behind patients' observations.

   A description of the data sets is given below in TSUM_A and TSUM_B. The patients in these two tables are connected by their ID numbers.

TSUM_A.csv

Item	Meaning	Remark
ID	identification of the patient	
Sex		
Birthday		m/d/yyyy

TSUM_B.csv

Item	Meaning	Remark
ID	identification of the patient	
Examination Date	date of the test	m/d/yyyy
ANA	anti-nucleus antibody concentration	
ANA Pattern	pattern observed in the sheet of	

Item	Meaning	Remark
	ANA examination	
Diagnosis	disease names	multivalued attribute
Symptoms	other symptoms observed	multivalued attribute
Thrombosis	degree of thrombosis	0: negative (no thrombosis) 1: positive (the most severe one) 2: positive (severe) 3: positive (mild)

a) Without pre-processing your data, use the decision-tree based algorithms, ID3 and C5 (i.e. ID3 with tree-pruning) provided by PASW Modeler to mine the data set, what can you discover in it? Please summarize your findings in no more than one A4-size page.

b) Discretize the data values in the data set using your own criteria rather than relying on the tool. Re-run the decision-tree based algorithm you used in a). What are the differences when comparing your new results with the old ones? Which, in your opinion, are more accurate and more meaningful?

c) In discretizing the data, how did you decide how many intervals to discretize the data into? Try to vary the number of intervals to see if there is any difference in classification accuracy of the resulting decision tree or rules that you discovered. Explain why there are such differences.

d) Other than discretization, what other ways of pre-processing of the data do you think you need to perform? Explain your pre-processing method in details. Show whether or not pre-processing of data actually leads to improved classification accuracy.