The Hong Kong Polytechnic University
Department of Computing
COMP 578 Data Mining and Data Warehousing
Semester 1, 03-04
Assignment 1
Due: October 9, 2003

Section A: Regular Questions

1. An electronic supermarket, *e-supermarket*, has started operation six months ago. Ever since it has started to operate, it has been keeping a record of all transactions. In order to improve its business, they have decided to mine their transaction database for interesting patterns. Assume that you have been hired as their data mining consultant and are given a snapshot of their transactional data as shown in Table 1.
   a) Choose your own definition of interestingness (i.e. set an appropriate minimum support and minimum confidence level), use the Apriori algorithm to find all interesting association rules involving 3 items and 4 items. You can write a simple computer program to perform the data mining tasks if you prefer.
   b) Assume that a user sets the lift ratio to 1.5, which rules you discovered are still interesting?

Table 1. A snapshot of the transactional database kept by *e-supermarket*

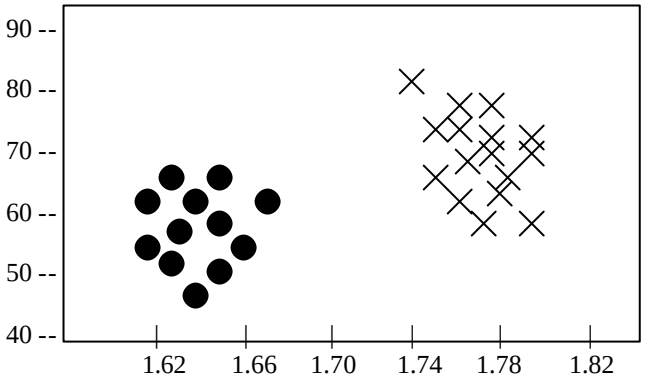
Transaction ID	Egg	Potato Chips	Pasta	Cheese	Beer	Diaper
1	Y	Y			Y	
2		Y		Y		
3		Y	Y		Y	Y
4	Y	Y		Y		
5	Y		Y			Y
6		Y	Y			
7	Y		Y	Y		Y
8	Y	Y	Y		Y	
9	Y	Y	Y			
10	Y	Y				Y
11	Y	Y		Y		
12		Y	Y		Y	
13	Y		Y		Y	
14	Y	Y	Y		Y	
15		Y	Y			Y
16	Y		Y			Y
17		Y		Y		
18	Y	Y	Y			Y

2. 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	97.8	5.7	33.3	Stay
2	145.6	12.3	77.8	Switch
3	99.2	3.3	99.9	Stay
4	160.0	6.7	44.8	Undecided
5	117.8	9.9	66.0	Stay
6	210.2	2.2	129.9	Undecided
7	100.3	1.8	55.8	Switch
8	97.6	6.7	77.5	Switch
9	169.9	7.7	98.0	Undecided
10	186.3	14.4	144.4	Switch
11	172.1	15.5	66.8	Undecided
12	88.8	4.9	111.1	Switch
13	132.7	7.5	88.9	Stay
14	155.5	10.0	77.1	Stay
15	144.4	8.9	101.1	Switch

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 133.3, an average duration of calls of 5.8 and a total calling time of 88.2, 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 123.4 and a total calling time of 91.2?
c) If you are free to choose the value of k, what will your choice be? Why?

3. Find a classifier for the following data sets collected by measuring and plotting the weight avainst height of a group of students in an undergraduate statistics course (the Y-axis represents weights in kg's and the X-axis represents height in cm).



Section B: Independent Learning

Assume that you are to be interviewed for a job tomorrow with an insurance company. To prepare yourself for it, search the library's databases for articles that tell you how data mining can be and has been used in the Insurance Industry. In no more than three A4-sized pages, convince your prospective employer that you are able to help them improve their business decisions if they are willing to hire you to be their Data Mining Consultant.

***** END *****