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

Assignment 1

Due: October 7, 2011

1. Your friend owns a computer store in Shatin, selling Desktop and Notebook PCs and other computer peripherals. Having been rather successful with his business there, he decided to venture into the infamous Mongkok Computer Center and he has already been there for three months. As expected, compared to his Shatin store, his new store has been recording much higher revenue but when it comes to profit, he is not so sure. He needs to pay several times more in rent! In order to stimulate sales, your friend feels that he needs to understand his customers in Mongkok more. To help him do so, you have asked for a sample of the transactional data he collected and they are shown in Table 1.

   A) Before we go through the ***Apriori*** algorithm, explain whether or not we should include the item, "***Maintenance***" for analysis.

   (10%)

   B) Set the ***Minimum Support*** to **14%** and ***Minimum Confidence*** to **85%**, find all interesting rules using the ***Apriori*** algorithm. (*Please do it manually and show your work step by step.*)

   (30%)

   C) Set the minimum ***Lift Ratio*** to **2**, which rules you discovered in **`Part A`** are still interesting?

   (10%)

   D) Given the results with the sampled data, you decided that data mining might well give your friend some useful information about customer behavior. You have decided to obtain all 20,000+ transaction records (***data-q1e.csv***) your friend collected for analysis.

   i. Using the Apriori algorithm available in PASW Modeler 13, show how interesting association relations can be discovered. Discuss the results and, based on the interesting associations discovered, make some suggestions to your friend to improve their sales.

   (10%)

   ii. In mining the data, you may like to note that the data are "noisy" and you may need to "clean" them. If cleaning is necessary, please describe the steps you took to do so.

   (10%)

   iii. Also, please explain how you arrived at the Minimum Support and Minimum Confidence you used in finding the interesting association rules.

   (10%)

E) By taking into considerations the product taxonomy shown in *Figure 1* below, can you discover anything more useful? Again, you may use the Apriori algorithm in PASW or other tools to help you to prepare your work or findings.

Read the attached article, "*Mining Multiple-Level Association Rules in Large Database*" (*ieee1999_v11_5_c417.pdf*), and write a brief report (in less than 800 words) to explain how the product taxonomy could be used during the data mining process to discover patterns at different levels.

(20%)

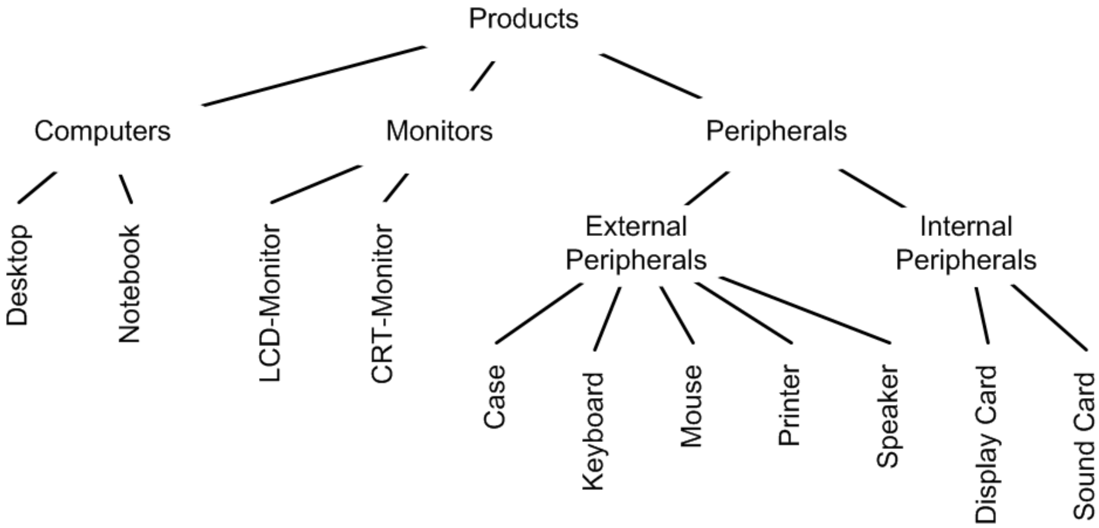

Figure 1: Computer products taxonomy

Table 1. A sample the transactional database kept by the computer store

```
TransID,Items
1,Display Card
1,Speaker
2,Desktop
2,Display Card
3,Case
3,Speaker
4,Speaker
5,Display Card
5,Mouse
6,Maintenance
7,Mouse
7,Desktop
8,Case
9,Desktop
10,Mouse
11,Maintenance
12,Case
13,Case
13,Desktop
13,Display Card
13,Mouse
13,Speaker
14,Display Card
15,Maintenance
16,Mouse
17,Maintenance
18,Display Card
19,Case
19,Desktop
19,Display Card
19,Mouse
19,Speaker
20,Maintenance
21,Case
22,Case
22,Display Card
23,Desktop
23,Display Card
23,Mouse
23,Speaker
24,Desktop
24,Display Card
25,Case
25,Desktop
25,Display Card
25,Mouse
25,Speaker
26,Case
26,Display Card
26,Speaker
27,Case
27,Desktop
27,Display Card
28,Case
28,Desktop
28,Display Card
28,Mouse
28,Speaker
29,Desktop
30,Display Card
31,Display Card
32,Maintenance
33,Case
34,Maintenance
35,Case
36,Display Card
37,Case
37,Desktop
37,Mouse
38,Desktop
38,Case
39,Maintenance
40,Desktop
41,Case
42,Desktop
43,Maintenance
44,Case
45,Display Card
46,Desktop
46,Display Card
46,Mouse
47,Case
48,Maintenance
49,Case
49,Display Card
49,Speaker
50,Case
50,Desktop
50,Display Card
50,Mouse
50,Speaker
```

*** END ****