The Hong Kong Polytechnic University
Department of Computing
COMP 578 Data Mining and Data Warehousing
Semester 1, 03-04
Assignment 2 (Part I)
Due: TBA

Section A: Regular Questions

1. Your friend has been following through the recommendations of some of the most popular financial analysts on ABC Bank's stock. He collected the following data set and would like you to help discover hidden patterns in it so that he knows when he should believe in whom.
   a. Using the ID3, discover the hidden rules for your friend.
   b. If Luk recommends Sell, Tang recommends Hold, Pong recommends Buy and Cheng recommends Sell, what action should you take?
   c. If, instead of the ID3, you use the Naive Bayesian Approach, what action should you take?

Luk	Tang	Pong	Cheng	Most appropriate action
Sell	Sell	Buy	Sell	Sell
Buy	Buy	Buy	Buy	Buy
Sell	Buy	Buy	Buy	Buy
Sell	Hold	Sell	Buy	Buy
Sell	Sell	Buy	Sell	Sell
Buy	Buy	Buy	Sell	Sell
Buy	Sell	Buy	Buy	Buy
Hold	Hold	Sell	Buy	Buy
Hold	Buy	Sell	Sell	Sell
Sell	Hold	Buy	Sell	Buy
Buy	Buy	Sell	Buy	Buy
Hold	Buy	Sell	Sell	Sell
Sell	Sell	Buy	Buy	Sell
Buy	Sell	Sell	Sell	Buy

2. Suppose that you are asked by your manager to determine if the customers of your company can be grouped according to their demographics and their buying patterns.

A Sample of Records from The Customer Database

Customer ID	Income	Age	Marital Status	Frequency Purchase	Average Amt Per Purchase	Children's Clothings
9100123	Low	Old	Single	High	Low	May be
9303034	High	Young	Married	Low	High	No
9210126	Low	Old	Married	High	Medium	Yes
9142020	High	Young	Single	Low	Low	May be
9221233	High	Young	Married	Low	Medium	No
9717273	Low	Old	Single	High	Low	Yes
9912113	High	Young	Single	High	Medium	May be
9576776	High	Old	Married	Low	High	Yes
9125022	Low	Old	Married	Low	Low	May be
9687786	Low	Young	Single	High	Medium	No

a) Find an inherent grouping of records using the Condoset approach (you need only perform the first iteration).
b) How many different groups are there?
c) Repeat the same using the Hierarhical Agglomerative Single-Linkage clustering algorithm and the Hamming distance as a dissimilarity measure. How are your results different from that of the Condorset approach?

Section B: Independent Learning

1. Search the library or the Web for an article that describe a clustering problem involving symbolic (or qualitative, categorical, or discrete-valued) data. You are to prepare for a powerpoint presentation as if you are to give it to the class. Your presentation should contain these sections:
   - Motivation – explain why is the problem worth.
   - Problem statement – describe what exactly the problem is, what data are given and what, if any, assumptions are made?
   - Critical review of existing techniques -- if the problem has been addressed before, describe the techniques used. Explain if they are effective.
   - The proposed solution – Try to understand the solution if you can and give a high level description of it.
   - Evaluation – describe how the technique is evaluated and discuss if the proposed solution is effective.
   - Conclusion -- give a summary and discuss possible future work. Explain if you would consider the problem or the proposed solution for your project or dissertation.

**** END ****