Csci 8980
Homework 7

1. **Numeric Association Rules.**
Read the paper "Mining Quantitative Association Rules in Large Relational Tables" by Srikant & Agrawal (available http://www.almaden.ibm.com/cs/people/ragrawal/papers/sigmod96.ps) and answer the following questions.

a. Numeric attributes are often handled by discretizing the range of attribute values into disjoint intervals. Describe, in your own words, what are the main problems with mining numeric association rules.

b. Briefly explain how these problems are handled in the paper.

c. What modifications are needed in the candidate generation step of the Apriori algorithm in order to support numeric attributes.

Consider the transaction matrix given in Table 1 below:

TID	Temperature	Pressure	Alarm 1	Alarm 2	Alarm 3
1	95	1105	0	0	1
2	85	1040	1	1	0
3	103	1090	1	1	1
4	97	1084	1	0	0
5	80	1038	0	1	1
6	100	1080	1	1	0
7	83	1025	1	0	1
8	86	1030	1	0	0
9	101	1100	1	1	1

Table 1

d. Suppose we use the following discretization strategies:
D1: For each attribute, partition the range of attribute values into 3 equal-sized bins. Label the new attributes as X1, X2 and X3 (where X is temperature or pressure).
D2: For each attribute, partition the range of the attribute values into 3 bins; where each bin contains equal number of data points. Label the new attributes as X1, X2 and X3 (where X is temperature or pressure).
For each strategy, answer the following questions:

i. Construct the transaction matrix having the new discretized attributes.

ii. Derive all the frequent itemsets having support ≥ 30%.

iii. From part d(ii), remove all non-maximal frequent itemsets. A frequent itemset X is maximal if no superset of X is frequent. For example, if {A, B, C} is frequent, you can remove all subsets of this itemset. For each of the remaining itemsets, derive all the association rules. Compute the confidence and lift factor[1] for these rules.

e. In the conclusion of this paper, it was suggested that clustering can be used for partitioning the numeric attributes :

i. Draw a graph of Temperature vs Pressure for the data points given in Table 1.

ii. From the graph, what do you think is a reasonable number of clusters for the data points? Label the clusters in the graph as CID1, CID2, CID3, etc.

iii. Which clustering techniques can be used to find the above clusters?

iv. Replace the temperature and pressure columns in Table 1 with attributes CID1, CID2, etc. Construct a transaction matrix using the new attributes (along with attributes Alarm1, Alarm2 and Alarm3).

v. Derive all the frequent itemsets having support ≥ 30%.

[1] Lift factor for A → C is confidence(A → C)/P(C).

vi. Repeat problem d(ii) using the itemsets derived in e(v).

f. Comparing the three discretization strategies, which method do you think works the best for the data set given in Table 1. Explain your reasons clearly.

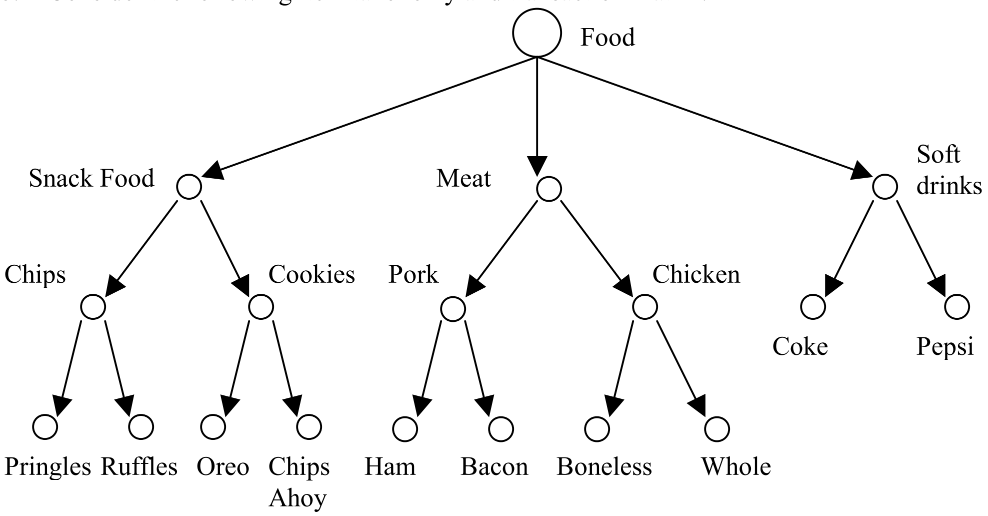
2. Association Rules with Item Taxonomy.

Read the following two papers :

- **Discovery of Multiple-Level Association Rules in Large Databases by Han and Fu (available at ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/vldb95.ps).**
- **Mining Generalized Association Rules by Srikant and Agrawal (available at http://www.almaden.ibm.com/cs/people/ragrawal/papers/vldb95_tax.ps)**

a. Describe what are the main problems of mining association rules with item taxonomy.

b. Consider the following item taxonomy and transaction matrix.



TID	Items
1	Pringles, Oreo, Coke, Ham
2	Ruffles, Pringles, Ham, Boneless Chicken, Pepsi
3	Ham, Bacon, Coke, Whole Chicken
4	Ruffles, Chips Ahoy, Ham, Boneless Chicken, Pepsi, Coke
5	Chips Ahoy, Bacon, Boneless Chicken
6	Ruffles, Ham, Bacon, Whole Chicken, Pepsi
7	Ruffles, Oreo, Pepsi, Boneless Chicken

Table 2.

i. Derive all the frequent itemsets for the above data set using Han's approach (algorithm 3.1) with support ≥ 50%.

ii. Derive the frequent itemsets for the above data set using Srikant's approach (algorithm Cumulate) with minimum support ≥ 50%.

iii. Are there any differences between the frequent itemsets discovered in (i) and (ii).

3. Interestingness Measures.

Read the subjective interestingness measure paper, "What makes Patterns Interesting in Knowledge Discovery Systems" by Silberschatz and Tuzhilin (available at http://www.bell-labs.com/user/avi/publication-dir/tkde.ps) and answer the following questions.

i. In this paper, what are the two major criteria for determining the subjective interestingness measure of a pattern?

ii. Suppose you are working as a data analyst for an online grocery store. Your company collects the transactions information for each of their customers as shown in Table 2 above.

a. Suppose the itemsets given below are found to be frequent. Comment on the subjective interestingness of each of the itemsets. Specifically, in your opinion, are these patterns interesting?

1. {Pepsi, Coke}.
2. {Oreo, Boneless Chicken}.
3. {Chips, Food}.
4. {Pringles, Ruffles, Oreo, Ham, Bacon, Whole Chicken}.
5. {Snack Food, Soft Drinks}.

b. Besides the items bought for each transaction, the database also contains other information such as the price of each item and whether discounts are offered for any combinations of these items. Will the knowledge about the price of items and promotions impact the assessment of interestingness measures? In other words, can this information be used to prune uninteresting patterns and will there be any interesting patterns (based on price or promotion criteria) missed by the existing algorithms? Explain your answer clearly (with examples).

4. Examine the applicability of association rule mining in the following domains.

a. Text documents (e.g. news articles from an online newspaper archive). Each document consists of a collection of words that appear in the document.

b. Stock market data (from NASDAQ or Dow Jones Index – e.g.: http://quote.yahoo.com). The raw data contains the closing price of each stock for the last 5 years. We are only interested in the fluctuations of stock prices (i.e. whether the stock price goes up or down compared to the previous closing price).

c. Census data (such as the 1998 Internet and Computer Use Census data which is available at http://www.bls.census.gov/cps/computer/1998/sdata.htm). The data contains, for each person, information such as salary group, number of computers, level of education, occupation category, number of hours spent on the Internet, etc.

Transaction	Item$_1$	Item$_2$	Item$_3$	Item$_4$	…
Basket$_1$	1	0	1	0	…
Basket$_2$	0	0	1	0	…
…	…	…	…	…	…

Table 3

In each of the application domain, answer the following questions:

i. Give an example of a hypothetical association rule that can be generated from the domain.

ii. Describe how you would construct the transaction matrix (Table 3) in order to derive the example pattern given above. Specifically, what are the baskets and items? For numeric and categorical attributes (e.g. for census data), you can use the technique described in Srikant & Agrawal's paper (see the reference given for problem 2).

iii. What are the limitations of your proposed transaction matrix? Specifically, will there be any interesting rules missed out by your choice of transaction matrix?

5. Indirect Association.

Read the indirect association paper by P.N. Tan, V. Kumar and J. Srivastava (available at http://159.84.66.212/PKDDForum/messages/7/24.html?WednesdayOctober2520001124am) and answer the following questions.

a. Why do we need the dependence condition for mediators? Give an example to show that using the mediator support condition alone is insufficient.

b. **BONUS (5points):** Give an application domain, other than the ones specified in the paper, for which indirect association can be useful.