Homework # 6
Csci 8980

1. Consider the database of transactions given in Table 1 below.

Customer Id	Transaction Id	Items Bought
1	0001	{A, B, E}
2	0002	{B, C, D}
2	0003	{C, E}
3	0004	{A, C, D}
3	0005	{A, B, E}
3	0006	{A, E}
4	0007	{A, B, C, E}
5	0008	{A, C, E}
6	0009	{D, E}
7	0010	{B, C}

Table 1: Database of transactions

a) **Compute the support for itemsets {A, B}, {C} and {A,B,C} using each transaction Id as a market basket.**

b) **Repeat part (a) by using each customer Id as a market basket. Explain the method you use to handle items bought in multiple transactions by the same customer.**

c) **Use your results in part (a) to compute the confidence for the association rules** AB → C **and** C → AB. **Is confidence a symmetric measure? What about support?**

d) **Compute the confidence for the association rule** C → AB **using your result in part (b).**

e) **Transitivity: Suppose the confidence for the rules** A → B **and** B →C **are larger than the minimum confidence threshold. Is it possible that** A → C **has a confidence less than the minimum confidence? Justify your answer.**

2. A function f(x) is said to be monotone in x if $x_1 < x_2$ implies that $f(x_1) \leq f(x_2)$. On the other hand, the function is anti-monotone in x if $x_1 < x_2$ implies that $f(x_1) \geq f(x_2)$. For each of the metrics given below, determine if it is a monotone, anti-monotone or non-monotone function with respect to the subset operator.

Example: support, $s(X) = \dfrac{\sigma(X)}{|T|}$ is anti-monotone in X with respect to the subset operator because support(X) ≥ support(Y) whenever $X \subset Y$.

a) $Laplace(X \rightarrow Y) = \dfrac{\sigma(X \cup Y)+1}{\sigma(X)+2}$ **with respect to subset operator for Y (i.e. when** $Y_1 \subset Y_2$, **compare** $Laplace(X \rightarrow Y_1)$ **with** $Laplace(X \rightarrow Y_2)$**).**

b) **Piatetsky-Shapiro interest measure,** $RI(X \rightarrow Y) = s(X \cup Y) - s(X)s(Y)$ **with respect to subset operator for** $X \cup Y$**; where** $s(X)$ **denotes the support for itemset X.**

c) **A characteristic rule is a rule of the form** $A \rightarrow B_1 B_2 \ldots B_k$, **where the premise (left-hand side) of the rule consists of a single item. An itemset of size k can be used to generate k characteristic rules. Show that the minimum confidence of characteristic rules generated from a given itemset is anti-monotone with respect to the subset operator for the itemset; i.e. show that** $Min(conf(B_1 \rightarrow B_2 B_3 \ldots, B_k), conf(B_2 \rightarrow B_1 B_3 \ldots B_k), \ldots, conf(B_k \rightarrow B_1 B_2 \ldots B_{k-1})) \geq Min(conf(B_1 \rightarrow B_2 B_3 \ldots B_{k+1}), conf(B_2 \rightarrow B_1 B_3 \ldots B_{k+1}), \ldots, conf(B_{k+1} \rightarrow B_1 B_2 \ldots B_k))$.

d) **A discriminant rule is a rule of the form B_1 B_2 … $B_k \rightarrow A$, where the consequent (right-hand side) of the rule consists of a single item. An itemset of size k can be used to generate k discriminant rules. Does the minimum confidence of discriminant rules generated from the same itemset satisfy any monotonicity property? Justify your answer.**

3. Support and confidence are not the only rule quality measures that have been proposed in association rule literature. The purpose of the following exercise is to illustrate the applicability of other rule interest measures in assessing the quality of a rule. For each of the measure below, compute and rank the following rules according to their interest measures. Use the transaction database given in Table 1 (where each market basket is defined at transaction-level). Show your work clearly.

Rules : A → B, B → C, A → C, B → D, C → E.

a) Support.

b) Confidence.

c) Interest(X → Y) = $\dfrac{s(X \cup Y)}{s(X)s(Y)}$

d) Piatetsky-Shapiro interest measure, RI(X → Y) = $s(X \cup Y) - s(X)s(Y)$

e) Lift factor = $\dfrac{conf(X \rightarrow Y)}{s(Y)}$

f) Gini Index =

$$s(X)\left(conf(X \rightarrow Y)^2 + conf(X \rightarrow \overline{Y})^2\right) + s(\overline{X})\left(conf(X \rightarrow Y)^2 + conf(X \rightarrow \overline{Y})^2\right) - s(Y)^2 - s(\overline{Y})^2$$

g) **(Optional)** J-Measure =

$$s(X)\left(conf(X \rightarrow Y)\log\frac{conf(X \rightarrow Y)}{s(Y)} + conf(X \rightarrow \overline{Y})\log\frac{conf(X \rightarrow \overline{Y})}{1 - s(Y)}\right)$$

where $s(X)$ is the support of itemset X and $conf(X \rightarrow Y)$ denotes the confidence of the rule.

4. **The Apriori algorithm[1] is often used to find all itemsets having at least the minimum support. The basic principle behind this algorithm is that any subset of a large itemset must also have minimum support. The large itemsets of size-(k+1) generated from itemsets of size-(k) using the apriori-gen function are called candidate itemsets. Candidate itemsets that satisfy the minimum support threshold are known as frequent itemsets.**

a. Suppose you were to apply the Apriori algorithm on the data set given in Table 1. Let the minimum support threshold be equal to 20%. How many candidate itemsets will be generated? How many frequent itemsets are there?

b. **Draw a lattice structure representing all possible itemsets that can be generated from the data set given in Table 1. Figure 1 illustrates an example of such structure for a transaction database with 4 items. The arrows in the diagram points to the larger itemsets that can be generated from a particular node. In your diagram, label each node as C (for candidate itemsets), F (for frequent itemsets) or N (not generated by Apriori at all).**

[1] R.Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules", *Proc. Of the 20th VLDB Conference*, 1994, pp. 487-499. http://www.almaden.ibm.com/cs/people/ragrawal/pubs.html#overview

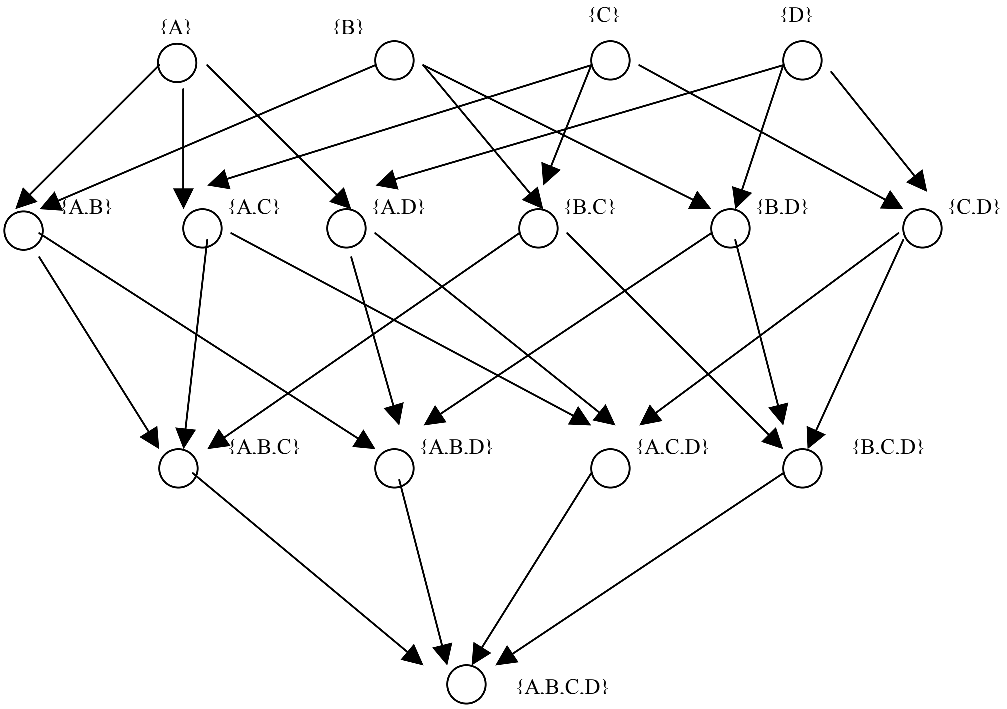


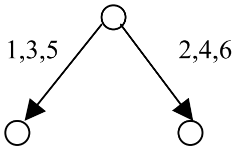
Figure 1: Lattice of itemsets.

5. **The Apriori algorithm uses a hash-tree data structure to efficiently count the support of candidate itemsets.**

   a. **Construct the hash tree for the following candidate 3-itemset:**

   {1,2,3}, {1,2,6}, {1,3,4}, {1,4,5}, {2,3,5}, {2,4,6}, {3,4,6}, {4,5,6}

   Assume the tree uses a hash function shown below (i.e. odd-numbered items will be hashed to the left child of the current node, while even-numbered items will be hashed to the left).



   b. **How many leaf nodes are there in the candidate hash tree? How many internal nodes are there?**

   c. **Suppose a transaction contains the following items : {1,2, 4, 5, 6} and you would like to find all candidate itemsets of size 3 contained in this transaction. Using the hash tree constructed in part a, how many leaf nodes of the hash tree are checked against the transaction? What are the candidate itemsets of this transaction?**