CI603 Data Mining

Tutorial 3

1. Consider the data set shown below:

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, c}

   a. Compute the **support** for itemsets **{e}**, **{b,d}**, and **{b,d,e}** by treating each transaction as a market basket.

   b. Use the results from *part a* to compute the **confidence** for the association rules **{b,d}** → **{e}** and **{e}** → **{b,d}**. Is confidence a symmetric measure?

   c. Repeat *part a* by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, 0 otherwise)

   d. Use the results in part (c) to compute the **confidence** for the association rules **{b,d}** → **{e}** and **{e}** → **{b,d}**.

2. Consider the data set shown below

    a. Compute the **support count** and **support** for each item.

TID	Items Bought
1	Milk, Bread, Butter
2	Bread, Cheese
3	Bread, Jam
4	Milk, Bread, Cheese
5	Milk, Jam
6	Bread, Jam
7	Milk, Jam
8	Milk, Bread, Jam, Butter
9	Milk, Bread, Jam

    b. Generate the **frequent itemsets** with *minsup* = 2 using the $F_{K-1} \times F_{k-1}$ candidate itemset generation method.

    c. Determine all the rules from the largest k-itemset(s) from *part b* with the consequents that contain only 1-itemsets.

    d. Using the formula $X_1 \cap X_2 \rightarrow Y_1 \cup Y_2$ merge the rules determined in *part c* to determine all the rules for the k-itemset(s).

    e. Determine the confidence $(X \rightarrow Y)$ for each rule generated in *parts c and d*.

    f. Determine the lift $(X \rightarrow Y)$ for each rule generated in *parts c and d*.