ASSIGNMENT NO. 5

Problem Statement:

Assignment on Association Rule Learning Download Market Basket Optimization dataset from below link. Data Set: https://www.kaggle.com/hemanthkumar05/market-basket-optimization. This dataset comprises the list of transactions of a retail company over the period of one week. It contains a total of 7501 transaction records where each record consists of the list of items sold in one transaction. Using this record of transactions and items in each transaction, find the association rules between items. There is no header in the dataset and the first row contains the first transaction, so mentioned header = None here while loading dataset. Follow following steps:

A. Data Preprocessing
B. Generate the list of transactions from the dataset
C. Train Apriori algorithm on the dataset
D. Visualize the list of rules
E. Generated rules depend on the values of hyper parameters. By increasing the minimum confidence value and find the rules accordingly

Related Theory:

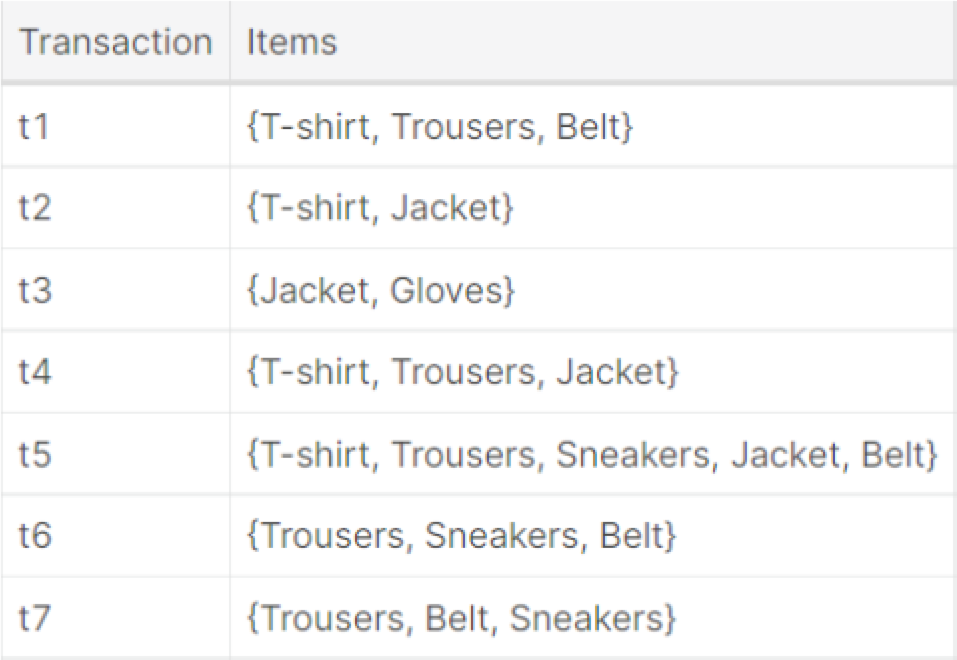
In this assignment we are going to use the Apriori algorithm to perform a Market Basket Analysis. A Market what? Is a technique used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions, providing information to understand the purchase behavior. The outcome of this type of technique is, in simple terms, a set of rules that can be understood as “if this, then that”. For more information about these topics, please check in the following links:

- Market Basket Analysis
- Apriori algorithm
- Association rule learning

First it’s important to define the Apriori algorithm, including some statistical concepts (support, confidence, lift and conviction) to select interesting rules. Then we are going to use a data set containing more than 6.000 transactions from a bakery to apply the algorithm and find combinations of products that are bought together.

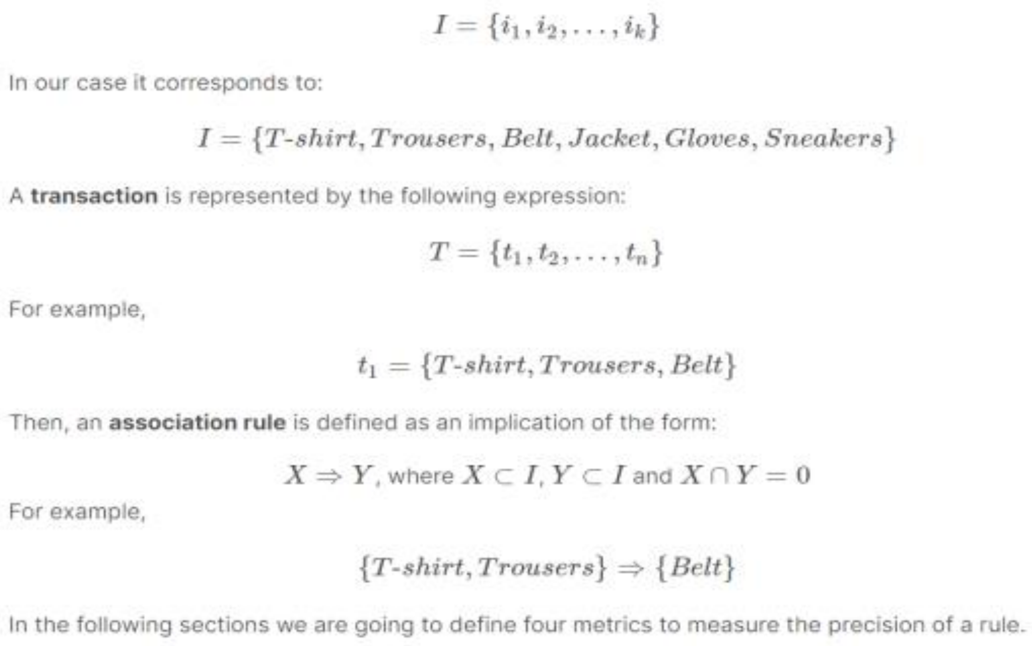
Association rules:

The Apriori algorithm generates association rules for a given data set. An association rule implies that if an item A occurs, then item B also occurs with a certain probability. Let's see an example,



Transaction	Items
t1	{T-shirt, Trousers, Belt}
t2	{T-shirt, Jacket}
t3	{Jacket, Gloves}
t4	{T-shirt, Trousers, Jacket}
t5	{T-shirt, Trousers, Sneakers, Jacket, Belt}
t6	{Trousers, Sneakers, Belt}
t7	{Trousers, Belt, Sneakers}

In the table above we can see seven transactions from a clothing store. Each transaction shows items bought in that transaction. We can represent our items as an item set as follows:



$$I = \{i_1, i_2, \ldots, i_k\}$$

In our case it corresponds to:

$$I = \{T\text{-}shirt, Trousers, Belt, Jacket, Gloves, Sneakers\}$$

A **transaction** is represented by the following expression:

$$T = \{t_1, t_2, \ldots, t_n\}$$

For example,

$$t_1 = \{T\text{-}shirt, Trousers, Belt\}$$

Then, an **association rule** is defined as an implication of the form:

$$X \Rightarrow Y, \text{ where } X \subset I, Y \subset I \text{ and } X \cap Y = 0$$

For example,

$$\{T\text{-}shirt, Trousers\} \Rightarrow \{Belt\}$$

In the following sections we are going to define four metrics to measure the precision of a rule.

Support:

Support is an indication of how frequently the item set appears in the data set.

$$supp(X \Rightarrow Y) = \frac{|X \cup Y|}{n}$$

In other words, it’s the number of transactions with both XX and YY divided by the total number of transactions. The rules are not useful for low support values. Let’s see different examples using the clothing store transactions from the previous table.

- $supp(T\text{-}shirt \Rightarrow Trousers) = \frac{3}{7} = 43\%$
- $supp(Trousers \Rightarrow Belt) = \frac{4}{7} = 57\%$
- $supp(T\text{-}shirt \Rightarrow Belt) = \frac{2}{7} = 28\%$
- $supp(\{T\text{-}shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2}{7} = 28\%$

Confidence:

For a rule X⇒YX⇒Y, confidence shows the percentage in which YY is bought with XX. It’s an indication of how often the rule has been found to be true.

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

For example, the rule T-shirt Trousers T-shirt Trousers has a confidence of 3/4, which means that for 75% of the transactions containing a t-shirt the rule is correct (75% of the times a customer buys a t-shirt, trousers are bought as well.) Three more examples:

- $conf(Trousers \Rightarrow Belt) = \frac{4/7}{5/7} = 80\%$
- $conf(T\text{-}shirt \Rightarrow Belt) = \frac{2/7}{4/7} = 50\%$
- $conf(\{T\text{-}shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2/7}{3/7} = 66\%$

Lift:

The lift of a rule is the ratio of the observed support to that expected if XX and YY were independent, and is defined as,

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)}$$

Greater lift values indicate stronger associations. Let’s see some examples:

- $lift(T\text{-}shirt \Rightarrow Trousers) = \frac{3/7}{(4/7)(5/7)} = 1.05$
- $lift(Trousers \Rightarrow Belt) = \frac{4/7}{(5/7)(4/7)} = 1.4$
- $lift(T\text{-}shirt \Rightarrow Belt) = \frac{2/7}{(4/7)(4/7)} = 0.875$
- $lift(\{T\text{-}shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2/7}{(3/7)(4/7)} = 1.17$

How does Apriori Algorithm Work ?

A key concept in Apriori algorithm is the anti-monotonicity of the support measure. It assumes that

- All subsets of a frequent item set must be frequent
- Similarly, for any infrequent item set, all its supersets must be infrequent too

Step 1: Create a frequency table of all the items that occur in all the transactions.

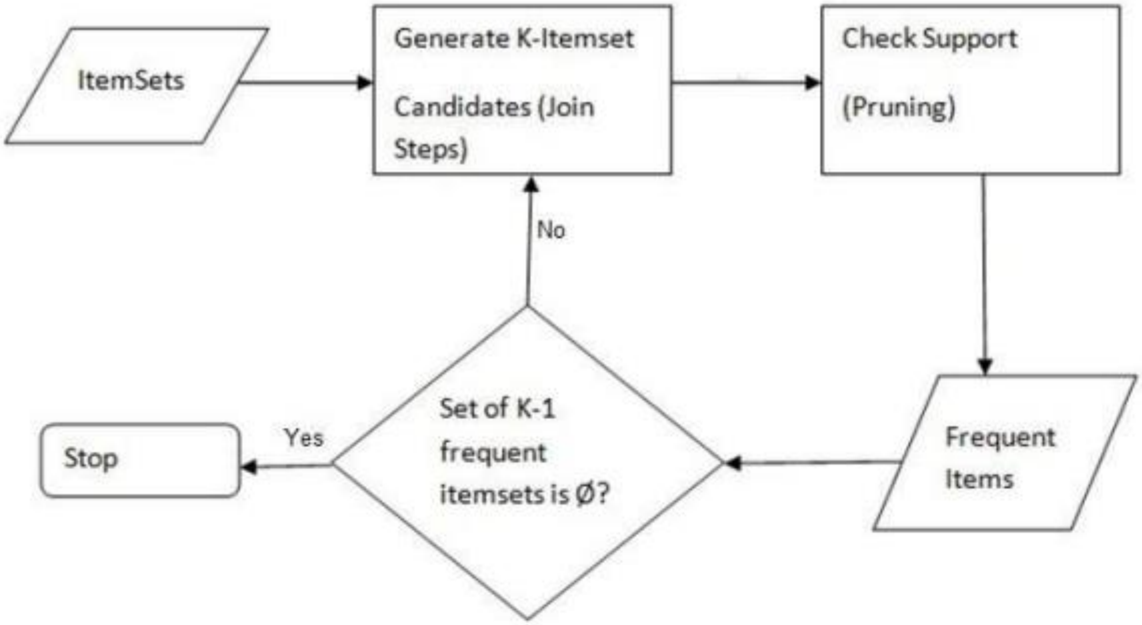
Step 2: We know that only those elements are significant for which the support is greater than or equal to the threshold support.

Step 3: The next step is to make all the possible pairs of the significant items keeping in mind that the order doesn’t matter, i.e., AB is same as BA.

Step 4: We will now count the occurrences of each pair in all the transactions.

Step 5: Again only those item sets are significant which cross the support threshold

Step 6: Now let’s say we would like to look for a set of three items that are purchased together. We will use the item sets found in step 5 and create a set of 3 items.



Conclusion:

In this assignment we have learned about the Apriori algorithm, one of the most frequently used algorithms in data mining. We have reviewed some statistical concepts (support, confidence and lift) to select interesting rules, we have chosen the appropriate values to execute the algorithm and finally we have visualized the resulting association rules.