ASSIGNMENT 6

Title : Association Rule Learning

Problem Statement :

Downloading the Market Basket Optimization dataset. It contains a total of 7501 transaction records where each record consists of the list of items sold in one transaction. Find association rules betn. items.

a. Data Preprocessing
b. Generate list of transactions from dataset
c. Train apriori algo. on dataset
d. Visualize list of rules

Objective :

It will help us understand and implement Apriori algorithm.

S/W Packages & H/W apparatus :

Linux OS : Ubuntu / Windows, Jupyter notebook
Pentium IV 1.7 GHz, 128 MB Ram, 40 GB HDD, etc.

Theory :

The Apriori algo. is used for the purpose of associative rule mining. A.R.M is a technique to identify frequent patterns and association among items eg. By finding correlations and association betn. diff. items and that customers place in their basket, patterns can be derived.

Apriori Algorithm :

It assumes that any subset of a frequent itemset must be frequent.

Say a transaction containing (wine, chips, bread) also contains (wine, bread). So acc. to principle of Apriori, if (wine, chips, bread) is frequent, then (wine, bread) must also be frequent.

The key concept in this algo. is that it assumes in all its subsets of a frequent itemset to be frequent. Similarly, for any infrequent itemset, all its superset must also be infrequent.

Here is a dataset containing of six transactions in an hour. Each transaction is a combination of 0's and 1's where 0 represents the absence of an item and 1 represents the presence of it.

Transaction ID	Wine	Chips	Bread	Milk
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

We can find multiple rules from this scenario. In a transaction of wine, chips & bread, if wine & chips are brought, then so is bread.
(wine, chips) ⇒ (bread).

In order to select the interesting rules out of multiple possible rules from this small business scenario, we will be using the following measures:

- Support
- Confidence
- Lift
- Conviction

① Support
It's nothing but the ratio of the no. of transactions in which item x appears to the total no. of transactions.

i.e. $\text{Support(wine)} = \frac{\text{No. of transactions containing wine}}{\text{Total no. of transactions}}$

$\text{Support(wine)} = 4/6 = 0.6667$

② Confidence
Confidence $(x \Rightarrow y)$ signifies the likelihood of y being purchased when x is purchased.

i.e. $\text{Conf.}((\text{wine, chips}) \Rightarrow (\text{bread})) = \frac{\text{Support(wine, C, B)}}{\text{Support(W, C)}}$

$\text{Conf.}((w, c) \Rightarrow (B)) = \frac{2/6}{3/6} = 0.6667$

③ Lift :

Lift $(x \Rightarrow y)$ is nothing but the 'interestingness' or the likelihood of y being purchased when x is sold. Unlike confidence $(x \Rightarrow y)$, this method takes into account, popularity of item y.

$$\text{Lift}(\{W, C\} \Rightarrow \{B\}) = \frac{\text{Support}(W, C, B)}{\text{Support}(W, C) * \text{Support}(B)}$$

$$\text{Lift}(\{W, C\} \Rightarrow \{B\}) = \frac{2/6}{3/6 * 4/6} = 1$$

- Lift $(x \Rightarrow y) = 1$ means there's no correlation within itemset.
- Lift $(x \Rightarrow y) > 1$:— there's a +ve correlation
- Lift $(x \Rightarrow y) < 1$:- there's a -ve correlation

Conclusion :

Understood and successfully implemented the association rule learning algo. on given dataset.