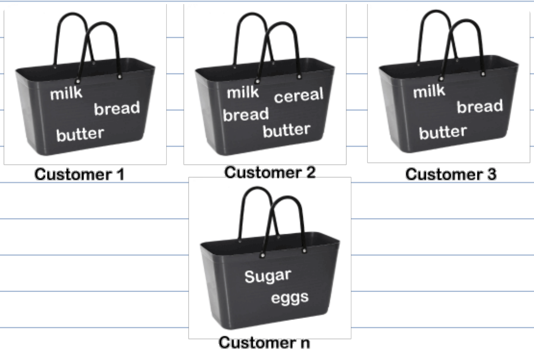
Association Rule Learning

Association Rule Learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.

It tries to find some interesting relations or associations among the variables of dataset. Association Rule Mining is employed in Market Basket Analysis, Web Usages Mining, continuous production etc. Here, Market Basket Analysis is a technique used by the various big retailers to discover the association between items.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.



Association Rule Learning works on the concept of If and Else statements.

Here, the If statement is called antecedent, and then statement is called as consequent. These types of relationships where we can find out some association or relation between two items is known as single cardinality. It is all about creating rules, and if the number of items increases, then the cardinality also increases accordingly.

Most Machine Learning algorithms works with the numeric datasets and hence tend to be mathematical. However, association rule mining is suitable for non-numeric, categorical data.

Association Rule Mining is a procedure which aims to observe frequently occuring patterns, correlations, or associations from datasets found in various Kind of databases such as RDBMS, transactional databases, and other forms of repositories. To measure the association between thousands of data items, there are several metrics.

An antecedent (if) is something that's found in data, and a consequent (else) is an item that is found in combination with the antecedent.

If a customer buys bread, he's 70% likely of buying milk.

↓ Antecedent ↓ Consequent

Support :- Support indicates how frequently the if/then relationship appears in the database. Support is how frequently an item appears in the dataset.

Support is the frequency of A, it can be defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transaction T, it can be written as:

$$\text{Support}(x) = \frac{\text{Frequency}(X)}{\text{Transactions}(T)}$$

or

$$\frac{\text{No. of transactions containing } X}{\text{Total number of transactions}}$$

Frequent Itemset :- An itemset whose support is greater than or equal to minimum support threshold.

Confidence :- Confidence tells about the number of times these relationships have been found to be true. This says how likely item Y is purchased when item X is purchased, expressed as (X → Y). This is measured by the proportion of transactions with item X, in which item Y also appears. (conditional probability of Y given X).

Confidence tells how oftem the items X and Y occurs together in the dataset when the occurance of X is already given (conditional probability). It is the ratio of the transactions that contains X and Y to the number of records that contains X.

$$\text{Confidence} = \frac{\text{Frequency}(X, Y)}{\text{Frequency}(X)}$$

OR

$$\frac{\text{No. of transactions containing } X \text{ and } Y}{\text{No. of transactions containing } X}$$

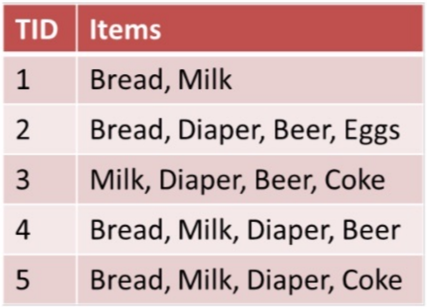
Lift :- Lift is the ratio of confidence to support. If the lift is < 1, then A and B are negatively correlated else positively correlated. And if it is 1, then it is not correlated.

The lift can be calculated as the ratio of the joint probability of two items X and Y, divided by the product of their probabilities.

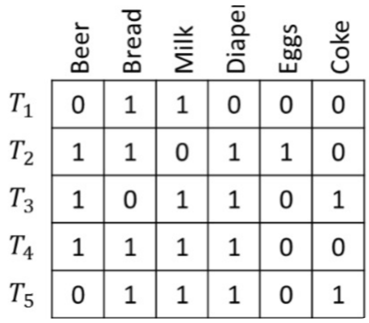
$$\text{Lift} = \frac{\text{Confidence}}{\text{Support}}$$

OR

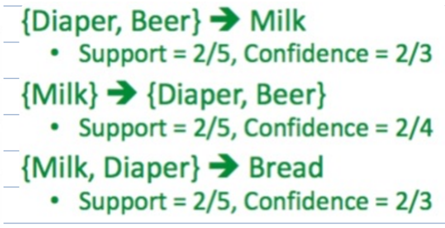
$$\text{Lift}\{A \rightarrow B\} = \frac{\text{Support}\{A, B\}}{\text{Support}\{A\} \times \text{Support}\{B\}}$$



TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke



	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1



{Diaper, Beer} → Milk

- Support = 2/5, Confidence = 2/3

{Milk} → {Diaper, Beer}

- Support = 2/5, Confidence = 2/4

{Milk, Diaper} → Bread

- Support = 2/5, Confidence = 2/3