**Association rule mining** is a technique which is used to find associations between many variables. It is mainly used in grocery stores, online shopping website and anything with large transactional database. The most common example that we see in everyday life like amazon for shopping purpose. When we order an item on Amazon website, our search information is tracked and will be used to give us suggestions to buy related items.

Market basket analysis is like ARM (Association Rule Mining). It is a modelling technique based on the concept, if you buy a certain group of items, you are more (or less) likely to buy another group of items. For example, if you are in a spa for facial treatment, you are more likely to take threading for eyebrows too, than somebody who came for haircut.

The measures of effectiveness of the association rules, are as follows:

**Support:** This indicates how frequently the itemset appears in the dataset.

**Confidence:** This indicates how often the rule has been found to be true.

These two measures are curial lets me discuss them with example in detail.

|  |  |
| --- | --- |
| TID | Items |
| 1 | Oranges, Apples |
| 2 | Oranges, Napkins, Alcohol, Eggs |
| 3 | Apples, Napkins, Alcohol, Soda |
| 4 | Oranges, Apples, Napkins, Alcohol |
| 5 | Oranges, Apples, Napkins, Soda |

This above dataset can also be represented like this:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Alcohol | Oranges | Apples | Napkins | Eggs | Soda |
| T1 | 0 | 1 | 1 | 0 | 0 | 0 |
| T2 | 1 | 1 | 0 | 1 | 1 | 0 |
| T3 | 1 | 0 | 1 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 1 | 0 | 0 |
| T5 | 0 | 1 | 1 | 1 | 0 | 1 |

**Support** can be calculated as the fraction of rows containing both X and Y or joint probability of X and Y. support is the percentage of transactions that contains XUY, it is taken to be probability P(XUY).

Support(X=>Y) = P(XUY)

**Confidence** is the fraction of rows containing Y or conditional probability of Y given X. confidence is the percentage of transactions in XUY containing X that also contains Y, it is taken to be the conditional probability, P(Y/X).

Confidence(X=>Y) = P(Y/X)

**Example:**

{Napkins, Alcohol}-k> Apples

Support = 2/5, confidence = 2/3

{Apples, Napkins}-> Oranges

Support = 2/5, confidence = 2/3

**Lift**: it is a ratio of confidence to support. If the lift is Less than 1 then X and Y are negatively correlated else positively correlated and if it is equal to 1 it is not correlated.

**Apriori algorithm** is used for mining frequent itemset for Boolean association rules. This is proposed by R. Agrawal and R. Srikant in 1994. Outline of Apriori algorithm. It is the first candidate generation and test approach for frequent pattern data mining and it is level-wise candidate generation and test approach. Initially, we scan the database once to get the frequent one itemset. By taking this itemset we generate length for (k+1) candidate item sets for length of k frequent item sets. At the case iteration we take the length k frequent item sets to generate length (k+1), then we go against database to test these candidates generated and to find the real frequent (k+1) item sets. For every iteration we set k=k+1. So, you can go until no frequent item sets are generated or no candidate item sets can be generated. After you exit from the loop, we must return all the frequent item sets derived. That’s the algorithm. There are two steps to generate candidates one is self-joining, where candidate item sets are generated by joining with itself and other is pruning, candidate itemset may or may not be frequent.

**Reference:**

1. <https://towardsdatascience.com/association-rule-mining-in-r-ddf2d044ae50>
2. <https://towardsdatascience.com/association-rule-mining-be4122fc1793>
3. Class material.