MARKET BASKET ANALYSIS USING ASSOCIATION RULES

Objective:

- To decipher how actionable patterns can be leveraged contextually.
- Association rules as unsupervised classification.
- Usage of the support and confidence measures to decide upon the important rules

MARKET BASKET ANALYSIS AND ASSOCIATIONS:

- Is the technique used for evaluating the purchasing patterns of individuals in order to target customers of interest.
- The mechanism of identifying and making decisions based on the associations obtained from large databases can add value to many other domain areas aside from supermarkets
- Association rules are set up in a transaction by the stipulation of an itemset on the LHS implying another item or item set on the right hand side.
- For illustration sake { bread, peanut , butter, jelly } will typically be found with bread in a transaction.
- {peanut , butter, jelly} ⟶ {bread} Association Rule
- The Association rules that make use of transactional databases are not used for prediction but only classification.
- This is an unsupervised algorithm and implements an exploratory paradigm for the purpose of discovery.
- This learning algorithm although similar to Decision rules are not supervised therefore the algorithm does not need to be trained ahead of time.
- The data is not labeled ahead of time. The algorithm work on the data and provides insights on the substantive patterns that it deciphers
- The common applications of interest are identifying DNA protein sequences in cancer data
- Identifying internet customer trends, medical insurance patterns.
- Association rule is able to work on a data and able to churn out patterns that an expert domain specialist may only be able to identify with experience.

Apriori Algorithm

- Database transactions are very exhaustive therefore the possible itemset becomes a very quantifiably challenging value.
- As the number of items or features grows the potential number of itemset grows exponentially.
- For illustration purposes if there are k items or features then in the rules we can have $2^{\wedge k}$ possible combinations if all the features are used. If the trader sells 100 items the rule complexity would be compounded to $2^{\wedge 100} = 1.27e+30$
- The task of the creating finalized association rules for the aforementioned scenario is daunting therefore a smart rule learning algorithm uses heuristics to evaluate the patterns that are more prevalent as opposed to patterns that are rare.
- For illustration purposes {orange fruit} and {brake oil} will rarely be associated in any way and therefore the rare patterns can be ignored.
- The most popular heuristic used for ascertaining the insightful patterns is Apriori which makes a prior determination of the properties of frequent itemsets.

- This algorithm works better with a large dataset as opposed to small dataset.
- Results are easy to interpret but sometimes it is difficult to differentiate between common sense and substantive meaningful patterns.
- Apriori utilizes statistical measures of an item's "interestingness" to formulate association rules from within transactional databases.

Measuring Rule Interest

- To make a determination of whether a rule is important two statistical measures namely support and confidence are used.
- A minimum threshold is numerically set up as a constraint in order to eliminate the inconsequential rules. It is important to understand what kind of rules are being eliminated.
- Support measures the frequency with which rule set occurs. For example calculating how often does {peanut, jelly, butter}-> bread occur. The support of an individual item can also be calculated as follows

  Support (X)=Count(X)/N

  N=number of transactions in the database
  X is the number of transactions containing the item or itemset.

  If {peanut, jelly, butter} occurs 100 out of 250 transactions then the support is 100/250

- A rule's Confidence measures its predictive power

  $$Confidence\ (X \rightarrow Y) = \frac{support(X,Y)}{support(X)}$$

  Numerator is the support of both X and Y divided by the support of X

- This intuitively informs us that presence of item X results in the presence of item Y.
- The reverse directionality is not implied by this confidence which is usually not true.
- Support (X, Y) is analogous to joint probability P(X and Y) whereas Support X is analogous to unconditional or marginal probability P(X) And Confidence (X→Y) is analogous to Conditional probability P(X|Y).
- Strong rules are those that have strong confidence and strong support.

Transaction number	Purchased items
1	*{flowers, get well card, soda}*
2	*{plush toy bear, flowers, balloons, candy bar}*
3	*{get well card, candy bar, flowers}*
4	*{plush toy bear, balloons, soda}*
5	*{flowers, get well card, soda}*

Support {get well -> flowers} = 3/5=60% Strong support

Confidence {get well -> flowers} =.6/.6 =100%

Association Rule {get well -> flowers} is a strong rule

Association Rule Formulation Steps

- If we are evaluating a itemset {A,B} by how frequently it occurs then it is explicit that if {A} occurs very infrequently then we do not need to use {A} or {A,B} or any other itemset to be included in the rule set
- Using the two statistical metrics the rule determination can be ascertained as follows:
- Identify the itemsets that meet the minimum support threshold.
- Constructing rules from itemsets that meet a minimum confidence threshold.
- The first phase is the evaluate the itemsets of each size starting with i=1.
- The first iteration consisting of computing the support of the itemset composed of 1 item, the second iteration is composed of 2 items etc.
- Each iteration results in i itemsets that meet the threshold criterion of the threshold support.
- The algorithm will eliminate some of the combination even before the next round. If {D} is infrequent in the first iteration from amongst A, B, C, D then {A,B} {A,C} and{B,C} will be evaluated .
- If for the next iteration it is discovered that {A,B} and {B,C} are frequent but {A,C} is not then any further combination consisting of {A,C} like {A,B,C} need not be considered.
- For the third iteration we cannot generate any itemset therefore algorithm will stop.
- Finally for the next phase the association rules are generated using confidence threshold evaluation.