Unveiling purchase patterns: Market Basket Analysis using the Apriori Algorithm

Vanshaj Vidyan (NISER)

CS460: Machine Learning

Market Basket Analysis

- Market Basket Analysis (MBA) is a data mining technique that analyzes patterns of co-occurence and determines the **strength of the link** between products purchased together. It is also called *association analysis.*
- It leverages these patterns recognized in any retail setting to understand the behaviour of the customer by identifying the relationships between the items bought by them. It is often used in conjunction with *customer segmentation*, another application of **unsupervised learning.**
- MBA helps the retailers know about the products frequently bought together, so as to keep those items *closer together* in the shopping aisles, as well as keep them stocked.

Here, we discuss *descriptive* MBAs, which offer actionable insights based on historical data.

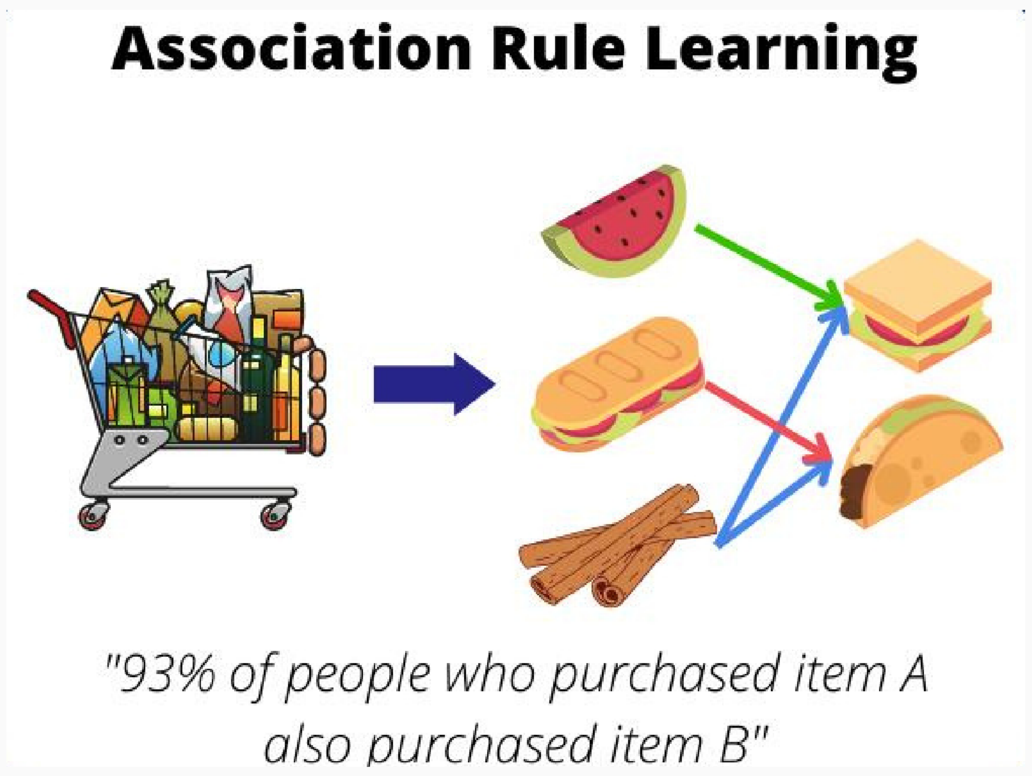
Association rules

- Association rules are an important machine learning concept being used in MBA. They represent relationships or patterns between items in a dataset, indicating that the presence of one item (the antecedent) is associated with the presence of another item (the consequent) in a transaction or dataset.
- Formally, an association rule may be written as:

  X ⇒ Y
- For example, a simple association rule can be written as:

  IF [{milk} AND {bread}] THEN {honey}

This rule suggests that if a customer purchases both milk and bread, they are likely to also purchase eggs.



Using this information, retailers may:
- offer discounts on only one of the associated products.
- place the associated items near to each other.

This will help to **increase the sales and revenue** of the company.

Terminology

- Itemset: refers to the set of items that are purchased together by a customer at the same time. Here, I={bread,butter,milk}
- Support count: is the frequency of a particular item set appearing in the transactional database. It is also stated as *probability*.
  - For instance, if milk has a support count of 50 out of 500 total transactions, then the probability is 50/500=0.1
- The antecedent is the IF itemset on the LHS, and the consequent is the THEN itemset on the RHS.

But, the possible association rules increase exponentially with the number of products: **computationally expensive** (just 10 products ---> 57,000 rules). How to prune them?

The *Apriori* algorithm

With the help of the *apriori* algorithm, we can further **reduce the search space** for frequent item sets. There are three components to the algorithm:
- Support: Fraction of total number of transactions in which itemset occurs.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- Confidence: Probability of occurence of {Y} given {X} is present.

But, confidence for an association rule having a very frequent consequent is always high!!

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

- Lift: Ratio of *confidence* to baseline probability of occurence of {Y}

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

- Frequent itemsets are the itemsets for which the *support* value is above a minimum threshold (minsup).

The Apriori algorithm works on the basis of the **apriori principle:**

"All subsets of a frequent itemset must be frequent."

This allows us to prune all supersets of an itemset which is not frequent. Example, if {milk, notebook} is not frequent, neither is any of its supersets. ***Pruning!***

Rule generation is a two step process:
1. Generating itemsets from a list of items: Get all the frequent itemsets using ***minsup,*** in the steps mentioned above. (Pruned using apriori)
2. Generating all possible rules: Rules are formed by partition of each itemset. From a list of all possible rules, we aim to identify rules falling above a min. confidence level ***(minconf)***. (Pruned using apriori).

As an example, if frequent itemset is {Bread, Egg, Milk, Butter}, candidate rules:

(Egg, Milk, Butter → Bread), (Bread, Milk, Butter → Egg), (Bread, Egg → Milk, Butter), (Egg, Milk → Bread, Butter), (Butter→ Bread, Egg, Milk)

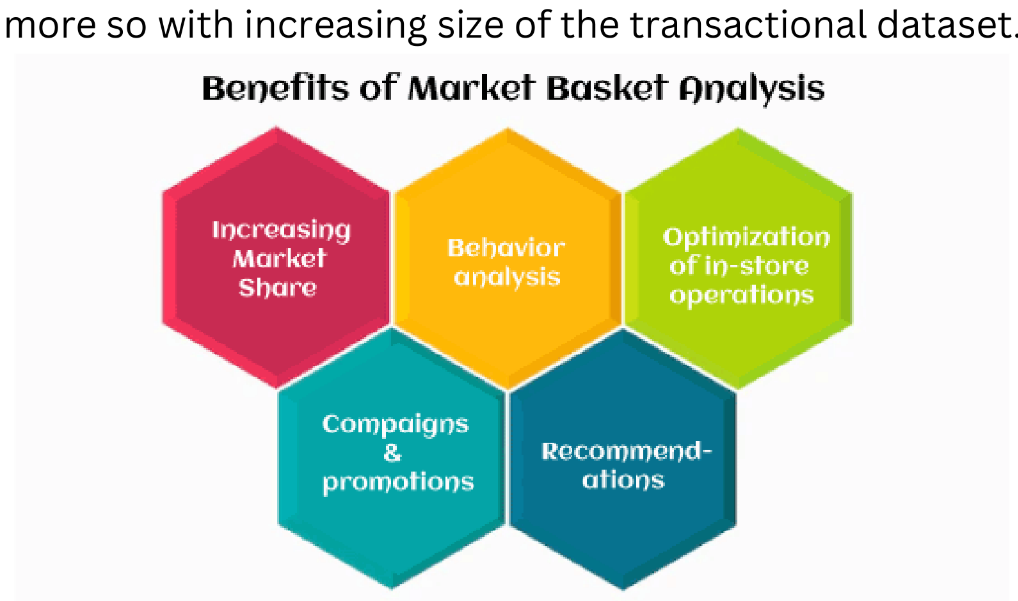
Now, this subset of rules generated can be searched for *highest values of lift* to make business decisions, from any transaction library.

Like k-NN, this algorithm does not require training.

Advantages of MBA

- Increasing revenue: MBA can be used to put together demographic and cultural data to determine the location of new stores or geo-targeted ads.
- Optimization of in-store operations: MBA is invaluable in determining the aisle placements of various products. Also, it can be optimized for each geographical location, as customer preference changes across countries.
- Campaigns and promotions: MBA is also used to determine which products form keystones in a company's product lineup.
- Recommendations: OTT platforms like Netflix and Hulu use MBA to filter recommendations, based on various customer segments.

Using the **simple** Apriori algorithm, MBA can be performed much more efficiently, more so with increasing size of the transactional dataset.



References

1. "Unsupervised Learning and Market Basket Analysis in Market Segmentation". I. Paranavithana et al. Proceedings of the WCE 2021.
2. Toivonen, H. (2011). Apriori Algorithm. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_27

Acknowledgement

- **Fellowship :** Dept. of Science and Technology, Govt. of India
- **Guidance :** Dr. Subhankar Mishra (NISER)