Market Basket Analysis: Association Rules

Market Basket Analysis

Data mining technique used to better understand customer purchasing patterns

Key techniques used to uncover associations between items

Analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”

Market Basket Analysis is one of the fundamental techniques used by large retailers to uncover the association between items.

In other words, it allows retailers to identify the relationship between items which are more frequently bought together.

Why Market basket analysis?

- ❑ Gives insight into which items are frequently bought together
- ❑ Helps improve marketing strategy
- ❑ Helps customer experience

Apriori/Association Rule

Association rule mining:

- Technique to identify underlying relations between various items

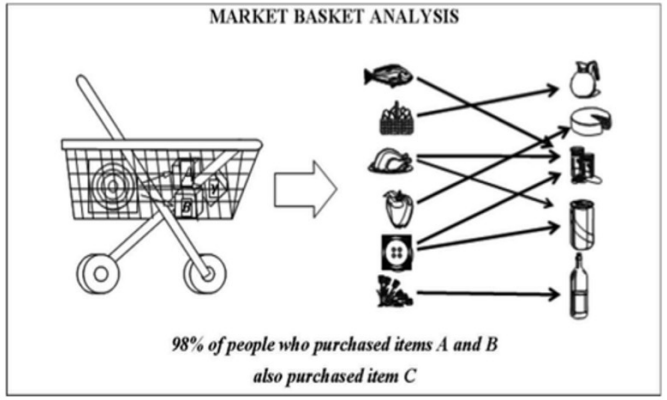
Most common approach: **Apriori algorithm**

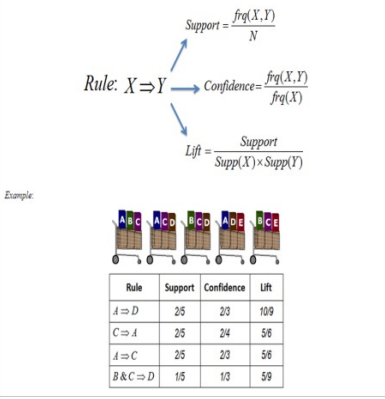
Main idea: “All non-empty subsets of a frequent itemset must also be frequent”

Bottom-up approach

- Start from every individual item
- Generate candidates by self-joining
- Itemsets that contain infrequent subsets are pruned
- Repeat until no more successful itemsets are formed

Apriori/Association Rule





One of the key techniques in Market Basket Analysis

Market basket analysis tells you about items that are “frequently bought together”

Eg., Amazon.com - “Customers also bought”,etc

Idea of Market Basket Analysis: if item x is bought, item/ itemset y is bound to be or not be bought

eg. , if one buys bread, chances of buying jam/butter is high

Key metrics for association rules

Key metrics for association rules

Assume we have a data set of 20 customers who visited the grocery store out of which 11 made the purchase:

Customer 1: Bread, egg, papaya and oat packet

Customer 2: Papaya, bread, oat packet and milk

Customer 3: Egg, bread, and butter

Customer 4: Oat packet, egg, and milk

Customer 5: Milk, bread, and butter

Customer 6: Papaya and milk

Customer 7: Butter, papaya, and bread

Customer 8: Egg and bread

Customer 9: Papaya and oat packet

Customer 10: Milk, papaya, and bread

Customer 11: Egg and milk

Support

Support: Percentage of orders that contain the item set.

In the example above, there are 11 orders in total, and {bread, butter} occurs in 3 of them.

Support = Freq(X,Y)/N

Support = 3/11 = 0.27

Confidence

Given two items, X and Y, confidence measures the percentage of times that item Y is purchased, given that item X was purchased.

This is expressed as:

Confidence = Freq(X,Y)/Freq(X)

Looking back to the example, percentage of times that butter(X) is purchased, given that bread(Y) was bought:

Confidence (butter -> bread) = 3/3 = 1

Confidence

Confidence values range from 0 to 1,

where 0 indicates that Y is never purchased when X is purchased, and 1 indicates that Y is always purchased whenever X is purchased.

Note that the confidence measure is directional.

This means that we can also compute the percentage of times that bread is purchased, given that item butter was purchased:

Confidence

Confidence (bread->butter) = 3/7 = 0.428

Here we see that all of the orders that contain bread also contain butter.

However, does this mean that there is a relationship between these two items, or are they occurring together in the same orders simply by chance?

To answer this question, we look at another measure which takes into account the popularity of both items.

Lift

Unlike the confidence metric whose value may vary depending on direction (eg: confidence{X ->Y} may be different from confidence{Y ->X}), **lift has no direction**.

This means that the lift{X,Y} is always equal to the lift{Y,X}:

lift{X,Y} = lift{Y,X} = support{X,Y} / (support{X} * support{Y})

lift{butter, bread} = lift{bread, butter} = support{butter, bread} / (support{butter} * support{bread})

lift{butter, bread} = lift{bread, butter} =(3/11)/((3/11)*(7/11))

lift{butter, bread} = lift{bread, butter} =1.571

In the example above, if butter occurred in 27.2% (=3/11)of the orders and bread occurred in 63.6% (= 7/11) of the orders, then if there was no relationship between them, we would expect both of them to show up together in the same order 17.35% of the time (ie: 27.2% * 63.6%).

The numerator, on the other hand, represents how often butter and bread actually appear together in the same order (27.2%).

Taking the numerator and dividing it by the denominator, we get to know how many more times butter and bread appear in the same order, compared to if there was no relationship between them (i.e., they are occurring together simply at random).

Summary

Lift can take the following values:

Lift	Lift	Lift
Lift > 1; implies that there is a positive relationship between X and Y (i.e., X and Y occur together more often than random)	Lift = 1; implies no relationship between X and Y (i.e., X and Y occur together only by chance)	Lift < 1; implies that there is a negative relationship between X and Y (i.e., X and Y occur together less often than random)

In our example, butter and bread occur together 1.57 times *more* than random, so we conclude that there exists a positive relationship between them.

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

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

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9