























shows how frequently an itemset occurs in a transaction. A typical example is a Market Based Analysis.

Market-Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently. Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

There are three common ways to measure association. So that before we start defining the rule, let us first see the basic definitions.

Important Definitions:

Support

Support is an indication of how frequently the itemset appears in the dataset. It is one of the measures of interestingness. This tells about the usefulness and certainty of rules. **50% Support** means a total 50% of transactions in the database follow the rule.

$$\text{Support}\{X\} = \frac{\text{Number of transaction in which } X \text{ appears}}{\text{Total number of transactions}}$$

$$\text{Support}\{\text{🍎}\} = \frac{4}{8}$$

Confidence

Confidence is an indication of how often the rule has been found to be true. It signifies the likelihood of item Y being purchased when item X is purchased. A confidence of 75% means that 75% of the customers who purchased an apple also bought beer.

$$\text{Confidence}\{X \rightarrow Y\} = \frac{\text{Support}\{X \cup Y\}}{\text{Support}\{X\}}$$

$$\text{Confidence}\{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support}\{\text{🍎} \cup \text{🍺}\}}{\text{Support}\{\text{🍎}\}}$$

If a rule satisfies both **minimum support** and **minimum confidence**, it is a **strong rule**.

Lift

how likely item Y is purchased when item X is purchased while controlling for how popular item Y is. If the lift value is 1, which implies no association between items. A lift value greater than 1 means that item Y is *likely* to be bought if item X is bought, while a value less than 1 means that item Y is *unlikely* to be bought if item X is bought. A lift of 12.5% means that 12.5% of the customers who likely to be bought beer if an apple is bought.

$$\text{Lift}\{\text{🍏} \rightarrow \text{🍺}\} = \frac{\text{Support}\{\text{🍏} \cup \text{🍺}\}}{\text{Support}\{\text{🍏}\} \times \text{Support}\{\text{🍺}\}}$$

So, Let's learn about the Association Rules:

For this dataset, we can write the following association rules: (Rules are just for illustrations and understanding of the concept. They might not represent the actuals).

Rule 1: If apple is purchased, Then the beer is also purchased in 75% of the transactions.

Rule 2: If beer is purchased, Then the meat is also purchased in 33.33% of the transactions.

Generally, association rules are written in the “IF-THEN” format. We can also use the term “Antecedent” for IF and “Consequent” for THEN.

In order to understand the concept better, let's take a simple dataset and find **frequent itemsets** and generate **association rules** on this.

TID	Items
T1	I1,I2,I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2,
that means, minimum support threshold is
 $2/9 \times 100\% = 22.2\%$
minimum confidence is 60%

A minimum support threshold is applied to find all frequent itemsets in a database.
A minimum confidence constraint is applied to these frequent itemsets in order to form rules.

Step 1: Create a frequency table of all the items that occur in all the transactions. For our case:

Item set	Sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Step 2: We know that only those elements are significant for which the support is greater than or equal to the threshold support. Here, the support threshold is 22.2%, hence only those items are significant which occur in more than 2 transactions and such items are l1, l2, l3, l4, l5. Therefore, we are left with:

Item set	Sup_count
l1	6
l2	7
l3	6
l4	2
l5	2

The table above represents the single items that are purchased by the customers frequently.

Step 3: The next step is to make all the possible pairs of the significant items keeping in mind that the order doesn't matter, i.e., AB is same as BA. To do this, take the first item and pair it with all the others such as {l1, l2}, {l1, l3}, {l1, l4}, {l1, l5}. Similarly, consider the second item and pair it with preceding items, i.e., {l2, l3}, {l2, l4}, {l2, l5}. We are only considering the preceding items because {l2, l1} (same as {l1, l2}) already exists. So, all the pairs in our example are {l1, l2}, {l1, l3}, {l1, l4}, {l1, l5}, {l2, l3}, {l2, l4}, {l2, l5}, {l3, l4}, {l3, l5}, {l4, l5}.

Step 4: We will now count the occurrences of each pair in all the transactions.

Item set	Sup_count
l1, l2	4
l1, l3	4
l1, l4	1
l1, l5	2
l2, l3	4
l2, l4	2
l2, l5	2
l3, l4	0
l3, l5	1
l4, l5	0

Step 5: Again only those itemsets are significant which cross the support threshold, and those are {l1, l2}, {l1, l3}, {l1, l5}, {l2, l3}, {l2, l4} and {l2, l5}.

Step 6: Now let's say we would like to look for a set of three items that are purchased together. We will use the itemsets found in step 5 and create a set of 3 items.

To create a set of 3 items another rule, called self-join is required. It says that from the item pairs {l1, l2}, {l1, l3}, {l1, l5}, {l2, l3}, {l2, l4} and {l2, l5} we look for two pairs with the identical letter and so we get different set of items

- $\{l_1, l_2\}$ and $\{l_1, l_3\}$, this gives $\{l_1, l_2, l_3\}$
- $\{l_1, l_2\}$ and $\{l_1, l_5\}$, this gives $\{l_1, l_2, l_5\}$
- $\{l_1, l_2\}$ and $\{l_2, l_4\}$, this gives $\{l_1, l_2, l_4\}$
- $\{l_1, l_3\}$ and $\{l_1, l_5\}$, this gives $\{l_1, l_3, l_5\}$
- $\{l_2, l_3\}$ and $\{l_2, l_4\}$, this gives $\{l_2, l_3, l_4\}$
- $\{l_2, l_3\}$ and $\{l_2, l_5\}$, this gives $\{l_2, l_3, l_5\}$
- $\{l_2, l_4\}$ and $\{l_2, l_5\}$, this gives $\{l_2, l_4, l_5\}$

Next, we find the frequency for these itemsets. Among them only two item sets can get as frequent itemset, those are $\{l_1, l_2, l_3\}$ and $\{l_1, l_2, l_5\}$

Item set	Sup_count
$\{l_1, l_2, l_3\}$	2
$\{l_1, l_2, l_5\}$	2

Again we can apply the self-join rule, then we get $\{l_1, l_2, l_3, l_5\}$ item set, but it is not a frequent itemset or check all subsets of these itemsets are frequent or not (Here itemset formed by joining above table is $\{l_1, l_2, l_3, l_5\}$ so its subset contains $\{l_1, l_3, l_5\}$ which is not frequent). We stop here because no frequent itemset is found frequent further.

General Process of the Apriori algorithm

The entire algorithm can be divided into two steps: **Step 1:** Apply minimum support to find all the frequent sets with k items in a database. **Step 2:** Use the self-join rule to find the frequent sets with $k+1$ items with the help of frequent k -itemsets. Repeat this process from $k=1$ to the point when we are unable to apply the self-join rule.

Thus we discovered all frequent item-sets now the generation of strong association rule comes into the picture. For that, we need to calculate the confidence of each rule.

As an example a confidence of 60% means that 60% of the customers who purchased a milk and bread also bought the butter. So here By taking example of any frequent itemset we will show rule generation. Let's take Itemset $\{I_1, I_2, I_3\}$, So rules can be

- $[I_1 \wedge I_2] \Rightarrow [I_3]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_3)}{\text{sup}(I_1 \wedge I_2)} = \frac{2}{4} * 100 = 50\%$ // **Rejected**
- $[I_1 \wedge I_3] \Rightarrow [I_2]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_3)}{\text{sup}(I_1 \wedge I_3)} = \frac{2}{4} * 100 = 50\%$ // **Rejected**
- $[I_2 \wedge I_3] \Rightarrow [I_1]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_3)}{\text{sup}(I_2 \wedge I_3)} = \frac{2}{4} * 100 = 50\%$ // **Rejected**
- $[I_1] \Rightarrow [I_2 \wedge I_3]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_3)}{\text{sup}(I_1)} = \frac{2}{6} * 100 = 33\%$ // **Rejected**

- $[I_2] \Rightarrow [I_1 \wedge I_3]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_3)}{\text{sup}(I_2)} = \frac{2}{7} * 100 = 28\%$ // **Rejected**
- $[I_3] \Rightarrow [I_1 \wedge I_2]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_3)}{\text{sup}(I_3)} = \frac{2}{6} * 100 = 33\%$ // **Rejected**

Let's take Itemset $\{I_1, I_2, I_5\}$,So rules can be

- $[I_1 \wedge I_2] \Rightarrow [I_5]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_5)}{\text{sup}(I_1 \wedge I_2)} = \frac{2}{4} * 100 = 50\%$ // **Rejected**
- $[I_1 \wedge I_5] \Rightarrow [I_2]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_5)}{\text{sup}(I_1 \wedge I_5)} = \frac{2}{2} * 100 = 100\%$ // **Selected**
- $[I_2 \wedge I_5] \Rightarrow [I_1]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_5)}{\text{sup}(I_2 \wedge I_5)} = \frac{2}{2} * 100 = 100\%$ // **Selected**
- $[I_1] \Rightarrow [I_2 \wedge I_5]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_5)}{\text{sup}(I_1)} = \frac{2}{6} * 100 = 33\%$ // **Rejected**
- $[I_2] \Rightarrow [I_1 \wedge I_5]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_5)}{\text{sup}(I_2)} = \frac{2}{7} * 100 = 28\%$ // **Rejected**
- $[I_5] \Rightarrow [I_1 \wedge I_2]$ //confidence = $\frac{\text{sup}(I_1 \wedge I_2 \wedge I_5)}{\text{sup}(I_5)} = \frac{2}{2} * 100 = 100\%$ // **Selected**

The minimum confidence threshold is 60%. So, We have found three strong association rules.