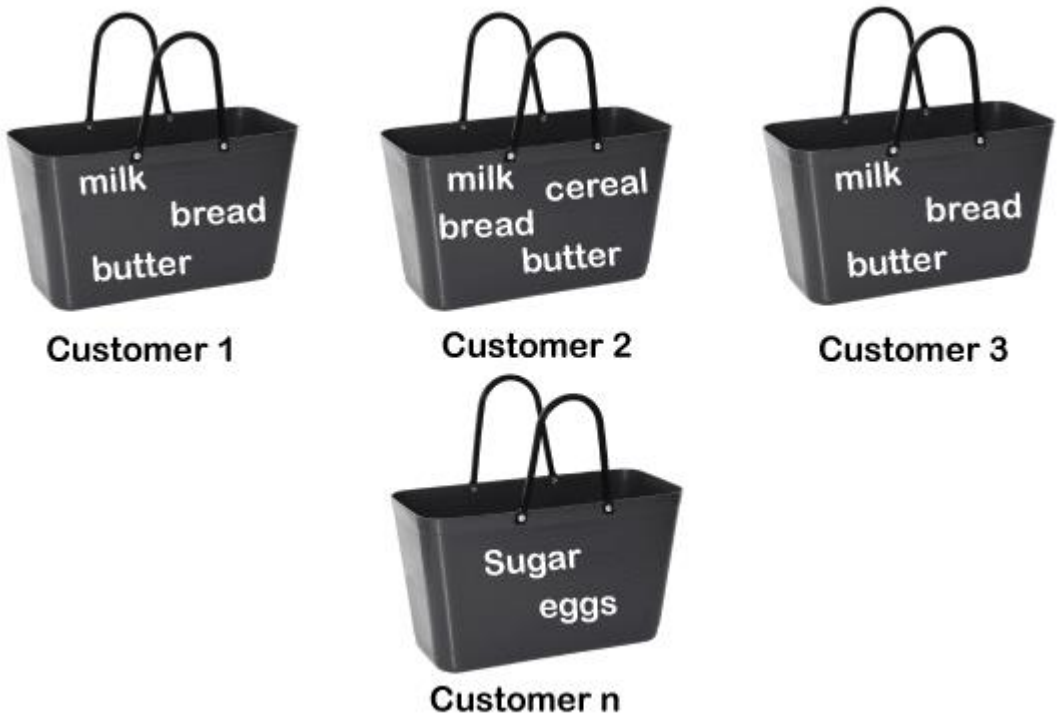
Association rules learning

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of the dataset. **It is based on different rules to discover the interesting relations between variables in the database.**

The association rule learning is one of the very important concepts of machine learning, and it is employed in **Market Basket analysis**, **Web usage mining**, **continuous production**, etc. Here market basket analysis is a technique used by various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

Association rules learning

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:



Association rules learning

Association rule learning can be divided into **three** types of algorithms:

1. **Apriori**
2. **Eclat**
3. **F-P Growth Algorithm**

Association rules learning

How does Association Rule Learning work?

Association rule learning works on the concept of If and Else Statement, such as if A then B.

Here the If element is called antecedent, and then statement is called as Consequent. These types of relationships where we can find out some association or relation between two items is known as single cardinality. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

- **Support**
- **Confidence**
- **Lift**

Association rules learning

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{Freq(X)}{T}$$

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

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

Association rules learning

Lift

It is the strength of any rule, which can be defined as below formula:

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

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

- If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Apriori Algorithm in Machine Learning

The Apriori algorithm uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rules, it determines how strongly or how weakly two objects are connected. This algorithm uses a **breadth-first search** and **Hash Tree** to calculate the itemset associations efficiently. It is the iterative process for **finding the frequent itemsets** from the large dataset.

This algorithm was given by the R. Agrawal and Srikant in the year 1994. It is mainly used for market basket analysis and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Apriori Algorithm in Machine Learning

What is Frequent Itemset?

Frequent itemsets are those items whose support is greater than the threshold value or user-specified minimum support. It means if A & B are the frequent itemsets together, then individually A and B should also be the frequent itemset.

Suppose there are the two transactions: A= {1,2,3,4,5}, and B= {2,3,7}, in these two transactions, 2 and 3 are the frequent itemsets.

Note: To better understand the apriori algorithm, and related terms such as support and confidence, it is recommended to understand the association rule learning.

Apriori Algorithm in Machine Learning

Generate strong association rules from the frequent item sets:

1. **Must satisfy the minimum support.**
2. **Must satisfy minimum confidence.**

Apriori Algorithm in Machine Learning

Example: Suppose we have the following dataset that has various transactions, and from this dataset, we need to find the frequent itemsets and generate the association rules using the Apriori algorithm:

Transaction ID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

$$\boldsymbol{Given\ minmum\ support} = \mathbf{30\%} = \frac{30}{100} * 4 = \mathbf{1.2}$$

$$\mathbf{and}\ \boldsymbol{minmum\ confidence} = \mathbf{80\%}$$

Apriori Algorithm in Machine Learning

Step-1: Calculating C1 and L1:

In the first step, we will create a table that contains support count (The frequency of each itemset individually in the dataset) of each itemset in the given dataset. This table is called the **Candidate set or C1.**

Note: We will take unique items.

Now, we will take out all the itemsets that have the greater support count that the Minimum Support (1.2). It will give us the table for the **frequent itemset L1**. Since all the itemsets have greater or equal support count than the minimum support, except the yellow marking row, so yellow marking row itemset will **be removed**.

Transaction ID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

$$\textbf{Given minmum support} = 30\% = \frac{30}{100} * 4 = \mathbf{1.2}$$

$$\textbf{and minmum confidence} = 80\%$$

C1

Items	Support
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L1

Items	Support
{1}	2
{2}	3
{3}	3
{5}	3

So, item sets: {1, 2, 3, 5}

Apriori Algorithm in Machine Learning

Step-2: Candidate Generation C2, and L2:

In this step, we will generate C2 with the help of L1. In C2, we will create the pair of the itemsets of L1 in the form of subsets.

After creating the subsets, we will again find the support count **from the main transaction table of datasets**, i.e., how many times these pairs have occurred together in the given dataset. So, we will get the below table for C2.

Now, we will take out all the itemsets that have the greater support count that the Minimum Support (1.2). It will give us the table for the **frequent itemset L2**. Since all the itemsets have greater or equal support count than the minimum support, except the yellow marking row, so yellow marking row itemset will **be removed**.

Transaction ID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

L1

Items	Support
{1}	2
{2}	3
{3}	3
{5}	3

C2

Items	Support
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

$$\textbf{Given minmum support} = 30\% = \frac{30}{100} * 4 = \mathbf{1.2}$$
$$\textbf{and minmum confidence} = 80\%$$

Apriori Algorithm in Machine Learning

Step-3: Candidate Generation C3, and L3:

In this step, we will generate C3 with the help of L2. In C3, we will create the pair of the itemsets of L2 in the form of subsets.

After creating the subsets, we will again find the support count **from the main transaction table of datasets**, i.e., how many times these pairs have occurred together in the given dataset. So, we will get the below table for C3.

Now, we will take out all the itemsets that have the greater support count that the Minimum Support (1.2). It will give us the table for the **frequent itemset L3**. Since all the itemsets have greater or equal support count than the minimum support, except the yellow marking row, so yellow marking row itemset will **be removed**.

Transaction ID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

L2

Items	Support
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2

C3

Items	Support
{1,2,3}	1
{1,3,5}	1
{2,3,5}	2

$$\boldsymbol{Given\ minmum\ support} = \mathbf{30\%} = \frac{30}{100} * 4 = \mathbf{1.2}$$

$$\mathbf{and}\ \boldsymbol{minmum\ confidence} = \mathbf{80\%}$$

Final item set: {2,3,5}

Apriori Algorithm in Machine Learning

Step-4: Finding the association rules for the subsets or the final item set:

To generate the association rules, first, we will create a new table with the possible rules from the occurred combination {A, B.C}. For all the rules, we will calculate the Confidence using below formula. After calculating the confidence value for all rules, we will exclude the rules that have less confidence than the minimum threshold(80%).

$$Confidence(A \rightarrow B) = \frac{support(A\ U\ B)}{support(A)}$$

Items	Support
{2,3,5}	2

Final item set: {2,3,5}

$$minmum\ confidence = 80\%$$

Rules	Support
(2,3) → 5	2
(2,5) → 3	2
(3,5) → 2	2
5 → (2,3)	2
3 → (2,5)	2
2 → (3,5)	2

Confidence (A → B) = Support(A U B) / Support(A)

So,

(2 , 3) → 5 = S((2 , 3) U 5) / S(2 , 3)
= 2 / 2
= 100%

(3 , 5) → 2 = S(3 , 5) U 2) / S(3 , 5)
= 2 / 2
= 100%

(2 , 5) → 3 = S((2 , 5) U 3) / S(2 , 5)
= 2 / 3
= 67%

3 → (2 , 5) = S(3 U (2 , 5)) / S(3)
= 2/ 3
= 67%

2 → (3 , 5) = S(2 U (3 , 5)) / S(2)
= 2 / 3
= 67%

5 → (2 ,3) = S(5 → (2 ,3)) / S(5)
= 2/3
= 67%

Apriori Algorithm in Machine Learning

Step-4: Finding the association rules for the subsets or the final item set:

As the given threshold or minimum confidence is 80%, so the last four rules 2 ^ 3 → 5, 3 ^ 5 → 2 can be considered as the strong association rules for the given problem.

$$Confidence(A \rightarrow B) = \frac{support(A\ U\ B)}{support(A)}$$

$$minmum\ confidence = 80\%$$

Rules	Support
(2,3) → 5	2
(2,5) → 3	2
(3,5) → 2	2
5 → (2,3)	2
3 → (2,5)	2
2 → (3,5)	2

Step 5:

Rules	Support	Confidence
(2 , 3) → 5	2	2/2 = 100%
(3 , 5) → 2	2	2/2 = 100%
(2 , 5) → 3	2	2/3 = 67%
3 → (2 , 5)	2	2/3 = 67%
2 → (3 , 5)	2	2/3 = 67%
5 → (2 , 3)	2	2/3 = 67%

After compare with threshold confidence (80%):

Final rules are, (2 ,3) → 5 & (3 , 5) → 2