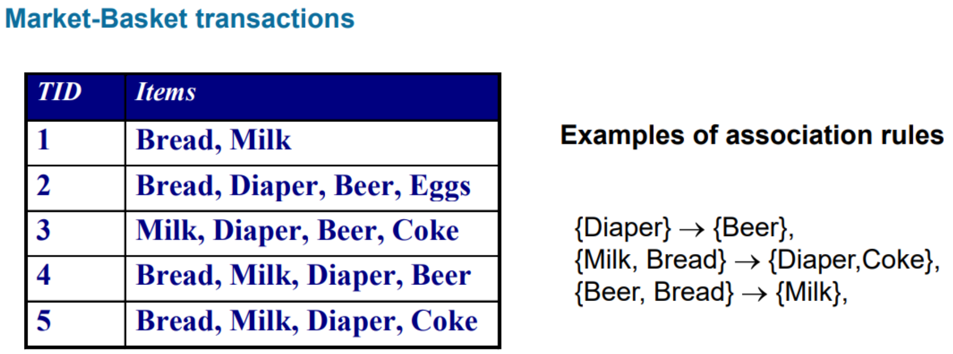
what is association rule?

- It is simple **If/Then** statements that help discover relationships between seemingly **independent relational** databases or other data repositories.

- It is suitable for **non-numeric**, **categorical data** and requires just a little bit more than simple counting.
- It is a procedure which aims to observe frequently **occurring patterns**, **correlations**, or **associations** from datasets found in various kinds of databases such as **relational databases**, **transactional databases**, and other forms of repositories.
- Association rule learning is a type of **unsupervised learning** technique that checks for the **dependency of one data item on another data item**

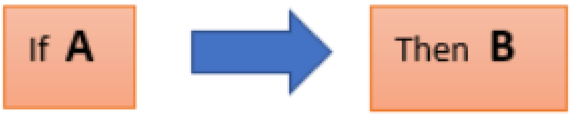


Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Examples of association rules

{Diaper} → {Beer},
{Milk, Bread} → {Diaper,Coke},
{Beer, Bread} → {Milk},



An association rule has 2 parts:

- **an antecedent (if) and**
- **a consequent (then)**

> *"If a customer buys bread, he's 70% likely of buying milk."*

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:

There are several metrics.

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

Support

- Support measures the frequency or occurrence of a particular itemset (a combination of items) in the dataset.
- It is calculated as the number of transactions containing the itemset divided by the total number of transactions.
- High support indicates that the itemset is frequent in the dataset.

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

Confidence

- Confidence measures the strength of the association between two items (antecedent and consequent) in a rule.
- It is calculated as the support of the itemset containing both the antecedent and consequent divided by the support of the antecedent.
- High confidence indicates that when the antecedent is present, there is a strong likelihood of the consequent being present as well.

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

Lift

- Lift measures how much more likely the consequent is to be bought when the antecedent is bought compared to when it is bought independently.
- It is calculated as the confidence of the rule divided by the support of the consequent.
- Lift greater than 1 indicates a positive association, meaning that the presence of the antecedent increases the likelihood of the consequent.

- 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.

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

Example to understand this concept.

We have already discussed above; you need a huge database containing a large no of transactions. Suppose you have 4000 customers transactions in a Big Bazar. You have to calculate the Support, Confidence, and Lift for two products, and you may say Biscuits and Chocolate. This is because customers frequently buy these two items together. Out of 4000 transactions, 400 contain Biscuits, whereas 600 contain Chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. Using this data, we will find out the support, confidence, and lift.

Support

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)

= 400/4000 = 10 percent.

Confidence

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits)

= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

Lift

Lift = (Confidence (Biscuits - chocolates)/ (Support (Biscuits)

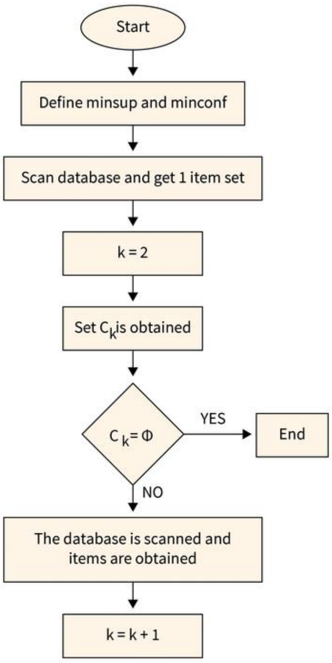
= 50/10 = 5

Association rule learning can be divided into three types of algorithms.

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

Apriori Algorithm?

- Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules.
- Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.
- Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.


Start
Define minsup and minconf
Scan database and get 1 item set
k = 2
Set C_k is obtained
$C_k = \Phi$
YES
End
NO
The database is scanned and items are obtained
k = k + 1

Advantages of Apriori Algorithm

- It is used to calculate large itemsets.
- Simple to understand and apply.

Disadvantages of Apriori Algorithms

- Apriori algorithm is an expensive method to find support since the calculation has to pass through the whole database.
- Sometimes, you need a huge number of candidate rules, so it becomes computationally more expensive

Types Of Association Rules In Data Mining

There are typically four different types of association rules in data mining. They are

- Multi-relational association rules
- Generalized Association rule
- Interval Information Association Rules
- Quantitative Association Rules