UNSUPERVISED LEARNING

Prepared by

P.Subhashree

UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

Prepared by
P.Subhashree

ASSOCIATION RULE

An association rule is an implication of the form X → Y where X is the association rule antecedent and Y is the consequent of the rule.

- One example of association rules is in basket analysis where we want to find the dependency basket analysis between two items X and Y.
- Support of the association rule X → Y:

$$\text{Support}(X,Y) \equiv P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

Prepared by
P.Subhashree

Confidence in the association rule X → Y:

$$\text{Confidence}(X \rightarrow Y) \equiv P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

Lift, also known as an interest of the association rule X → Y:

$$\text{Lift}(X \rightarrow Y) = \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)}$$

•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.

Prepared by
P.Subhashree

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

1.Apriori

2.Eclat

3.F-P Growth Algorithm

Prepared by

P.Subhashree

Apriori Algorithm

- This algorithm uses frequent datasets to generate association rules. It is designed to work on databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.
- It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in healthcare to find drug reactions for patients.

Prepared by

P.Subhashree

Steps for Apriori Algorithm

Below are the steps for the apriori algorithm:

Step-1: Determine the support of item sets in the transactional database, and select the minimum support and confidence.

Step 2: Take all supports in the transaction with a higher support value than the minimum or selected support value.

Step 3: Find all the rules of these subsets that have a higher confidence value than the threshold or minimum confidence.

Step 4: Sort the rules in the decreasing order of lift.

Clustering in Machine Learning

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as

"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.“

Prepared by
P.Subhashree

K-Means

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

EM algorithm

The expectation-Maximization algorithm can be used for the latent variables (variables that are not directly observable and are inferred from the values of the other observed variables) too to predict their values with the condition that the general form of probability distribution governing those latent variables is known to us.

Prepared by
P.Subhashree

- **Algorithm:**

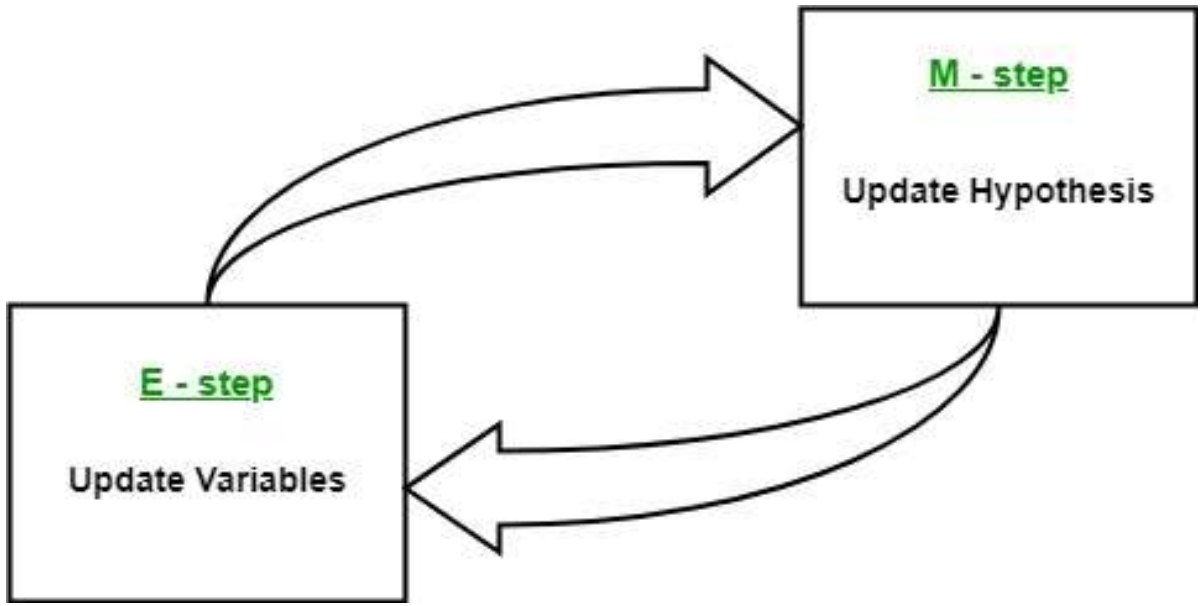
1.Given a set of incomplete data, consider a set of starting parameters.

2.Expectation step (E – step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.

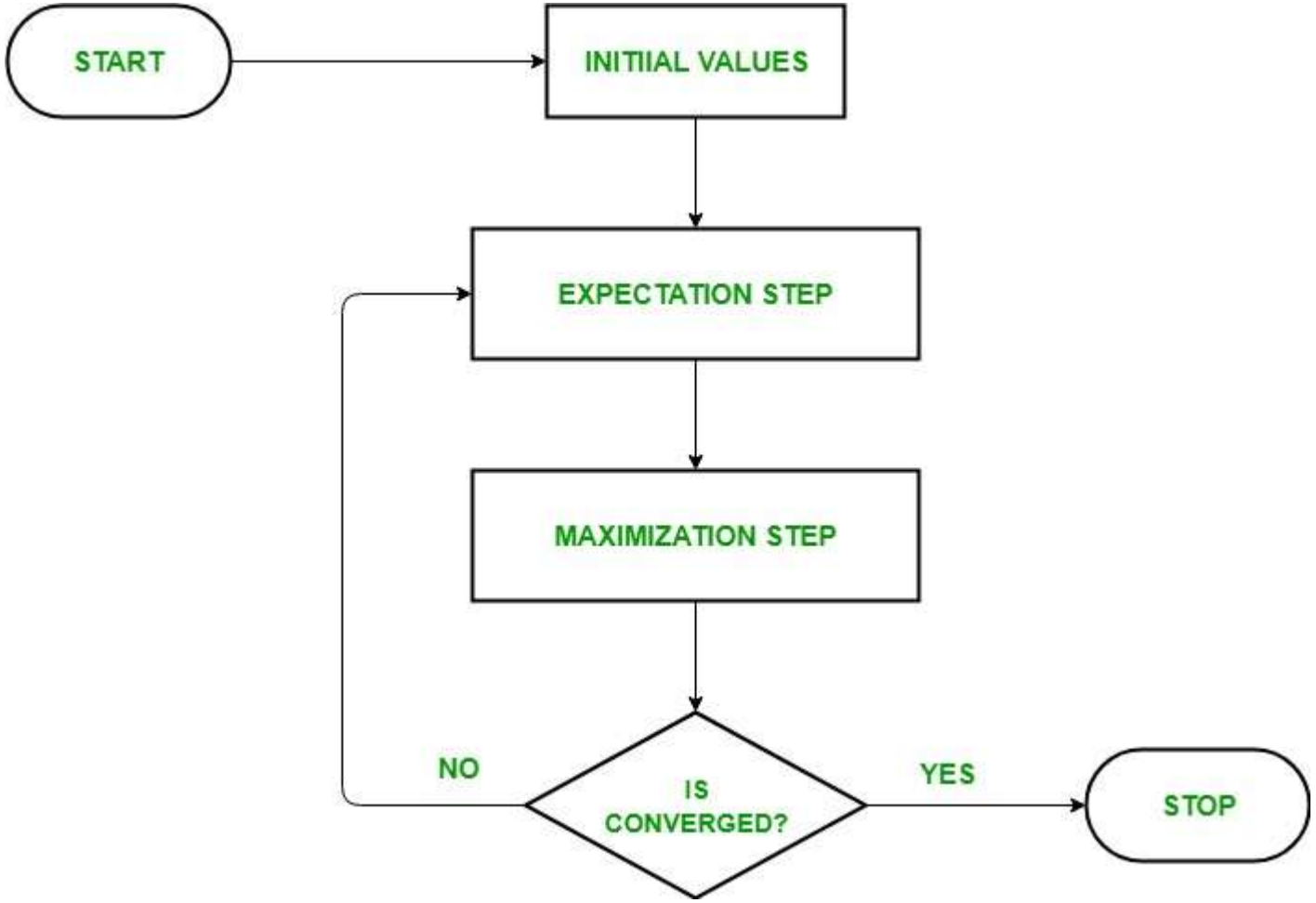
3.Maximization step (M – step): Complete data generated after the expectation (E) step is used in order to update the parameters.

4.Repeat step 2 and step 3 until convergence.

Prepared by
P.Subhashree


M - step
Update Hypothesis
E - step
Update Variables

Flow chart for EM algorithm



Gaussian Mixture Model

- Gaussian Mixture Model or Mixture of Gaussian as it is sometimes called is not so much a model as it is a probability distribution. It is a universally used model for generative unsupervised learning or clustering.
- A Gaussian is a type of distribution is a popular and mathematically convenient type of distribution. A distribution is a listing of outcomes of an experiment and the probability associated with each outcome

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where μ= Mean

σ=Standard Deviation

- if we have three Gaussian Distribution as GD1, GD2, and GD3 having mean as μ1, μ2,μ3 and variance 1,2,3 then for a given set of data points GMM will identify the probability of each data point belonging to each of these distributions.
- It is a probability distribution that consists of multiple probability distributions and has Multiple Gaussians.

Prepared by
P.Subhashree