UNIT-5
UNSUPERVISED MACHINE LEARNING MODELS

Prepared by : Ritesh Prajapati

INTRODUCTION TO MACHINE LEARNING(4350702)

SEMESTER-5 DIPLOMA COMPUTER ENGINEERING

Topics

5.1.1 Introduction of Unsupervised Learning

5.1.2 Example of Unsupervised Learning

5.2.1 Types of Unsupervised Learning

- Clustering (concept and example)
- Association (concept and example)

5.3.1 Differentiate Supervised vs. Unsupervised Learning

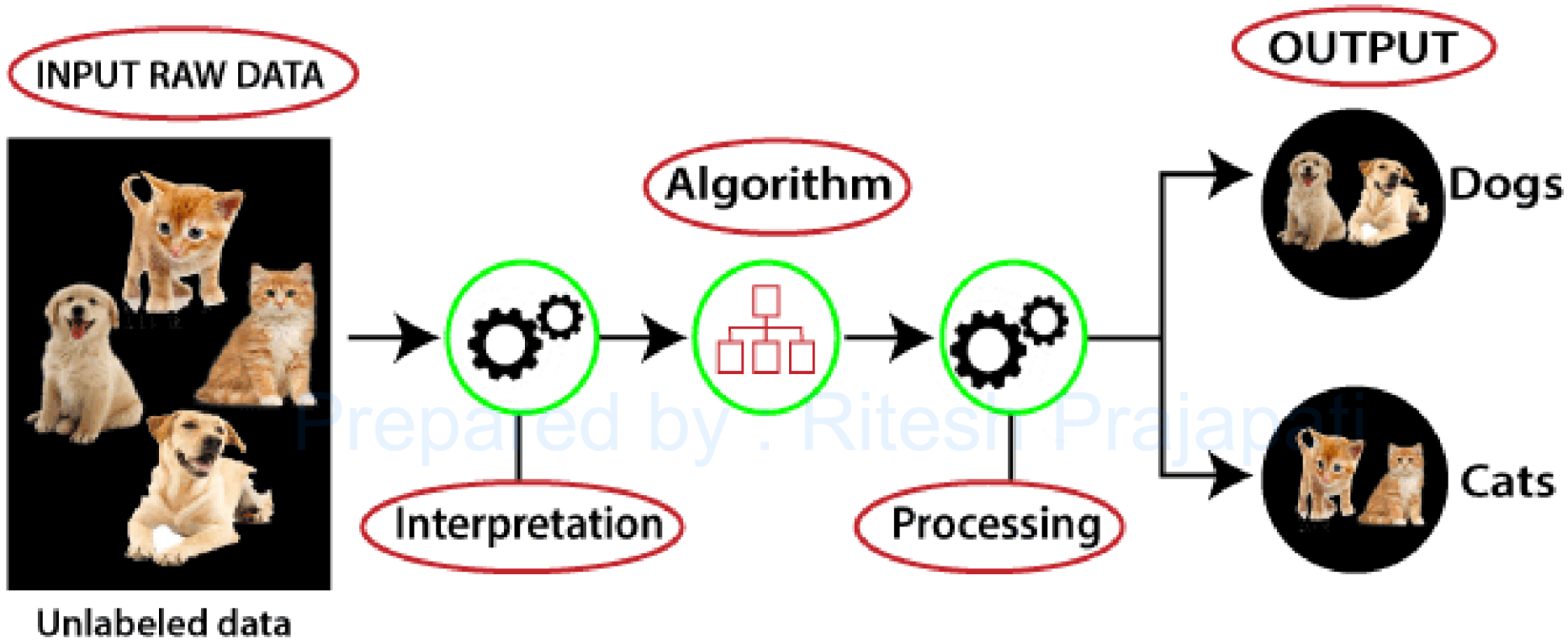
Prepared by : Ritesh Prajapati

Unsupervised Machine Learning

- No training dataset.
- Algorithm learns from unlabeled data, without explicit guidance.
- Models itself find the hidden patterns and insights from the given data.
- ***Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.***
- We have the input data but no corresponding output data.
- Similar as a human learns to think by their own experiences.
- *More close to real AI.*

Prepared by : Ritesh Prajapati

Unsupervised Machine Learning -Example



Prepared by : Ritesh Prajapati

Unsupervised Machine Learning -Example

- Input dataset containing images of different types of cats and dogs.
- The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset.
- **The task of the unsupervised learning algorithm is to identify the image features on their own.**
- Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.
- Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

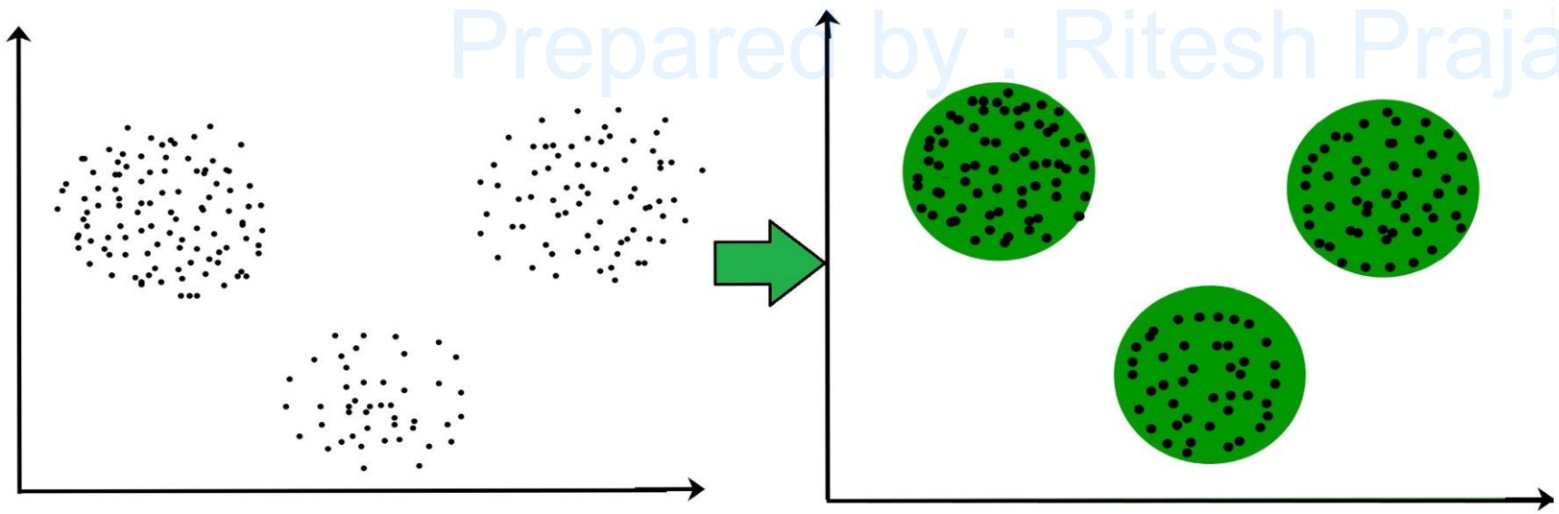
Prepared by : Ritesh Prajapati

Unsupervised Machine Learning -Types

Clustering:

- Grouping **the objects into clusters** such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.
- The goal of clustering is to discover inherent **groupings** in unlabelled data and form clusters.

Prepared by : Ritesh Prajapati



Unsupervised Machine Learning -Types

- **Clustering**: Examples
  - Customer Segmentation:
  - Identification of Cancer Cells:
  - Search Engines:
  - In Biology:
- **Clustering**: Common algorithms
  - K-Means:
  - Partitioning Clustering
  - Density-Based Clustering
  - Distribution Model-Based Clustering
  - Hierarchical Clustering
  - Gaussian Mixture Models (GMM):

Prepared by : Ritesh Prajapati

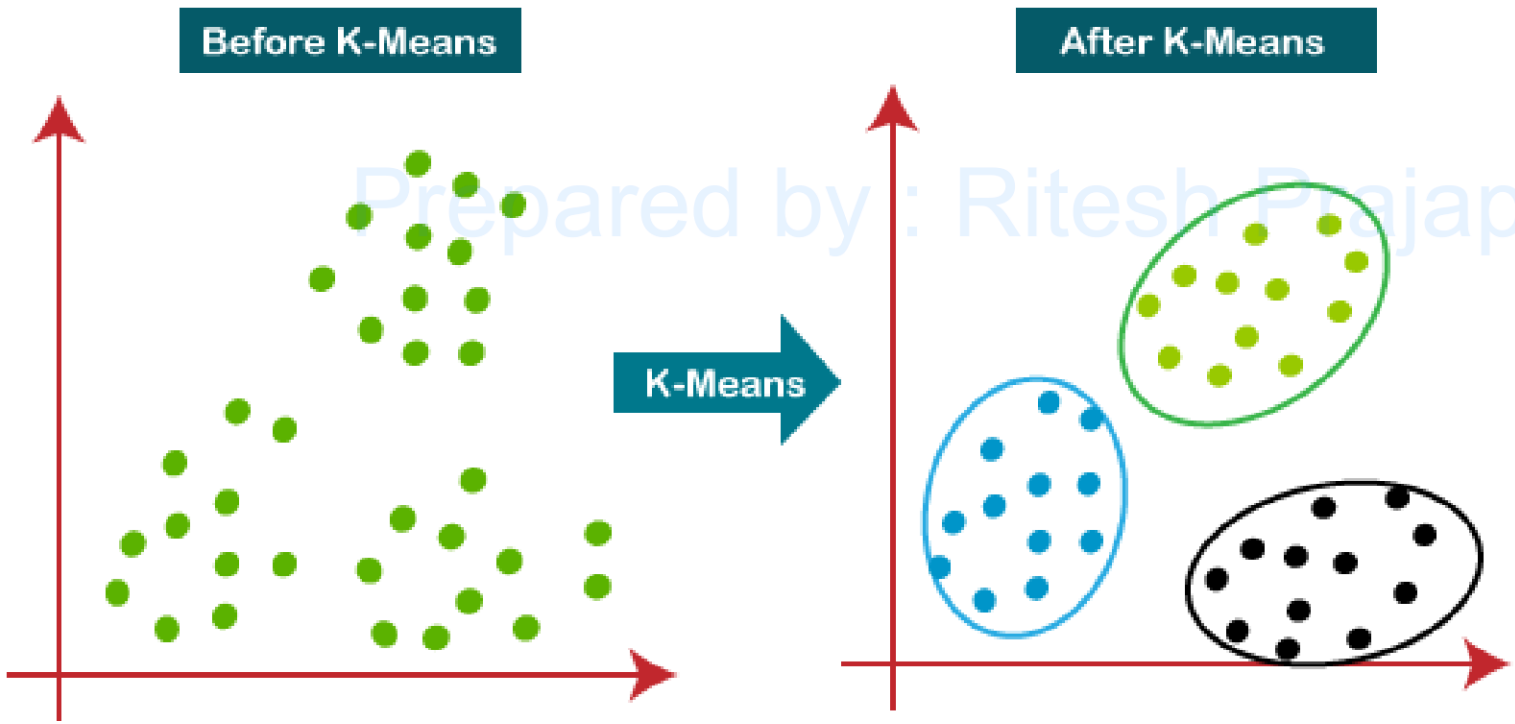
K-Means Clustering

- It groups the unlabeled dataset into different clusters.
- K defines the number of pre-defined clusters that need to be created.
- if K=2, there will be two clusters, K=3, there will be three clusters.
- *It is an **iterative algorithm** that divides the **unlabeled dataset into k different clusters** in such a way that **each dataset belongs only one group** that has similar properties.*
- It is a **centroid-based algorithm**, where each cluster is associated with a centroid.
  - *Centroid : the centre of mass of a geometric object*.
- The main aim is to minimize the sum of distances between the data point and their corresponding clusters.
- The algorithm takes the **unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process** until it does not find the best clusters.

Prepared by : Ritesh Prajapati

K-Means Clustering

- It mainly performs two tasks:
- Determines the best value for K center points (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.



Prepared by : Ritesh Prajapati

K-Means Clustering -Steps

- **Step 1**: Select K points in the data space and mark them as initial centroids

loop

- **Step 2**: Assign each point in the data space to the nearest centroid to form K clusters
- **Step 3**: Measure the distance of each point in the cluster from the centroid
- **Step 4**: Calculate the Sum of Squared Error (SSE) to measure the quality of the clusters
- **Step 5**: Identify the new centroid of each cluster on the basis of distance between points
- **Step 6**: Repeat Steps 2 to 5 to refine until centroids do not change

end loop

Prepared by : Ritesh Prajapati

K-Means Clustering -Working

- Suppose we have two variables M1 and M2. Scatter plot is given in Figure1.
- Let's K=2, to identify the dataset and to put them into different clusters.
- Choose some random k points or centroid to form the cluster. we are selecting the below two points as k points(not the part of dataset). Figure2.
- Assign each data point of the scatter plot to its closest K-point. Compute it by calculating the distance between two points.
- So, we will draw a median between both the centroids. Figure3

Prepared by : Ritesh Prajapati

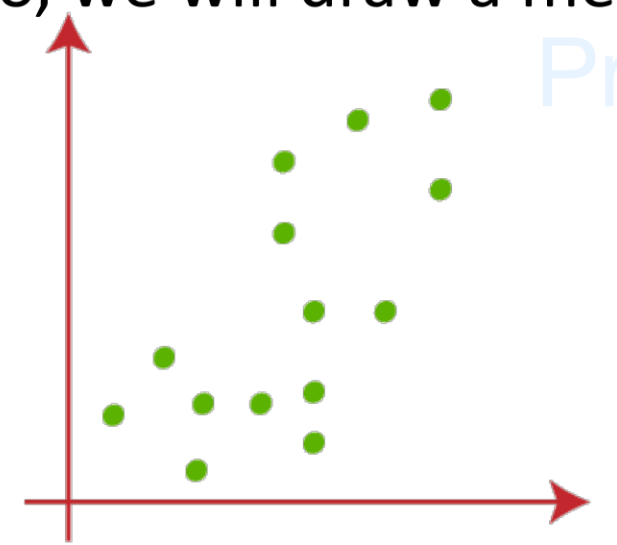


Figure1

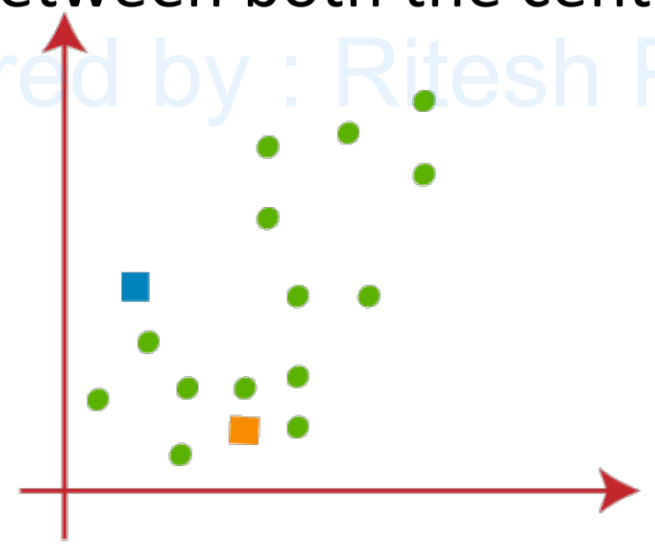


Figure2

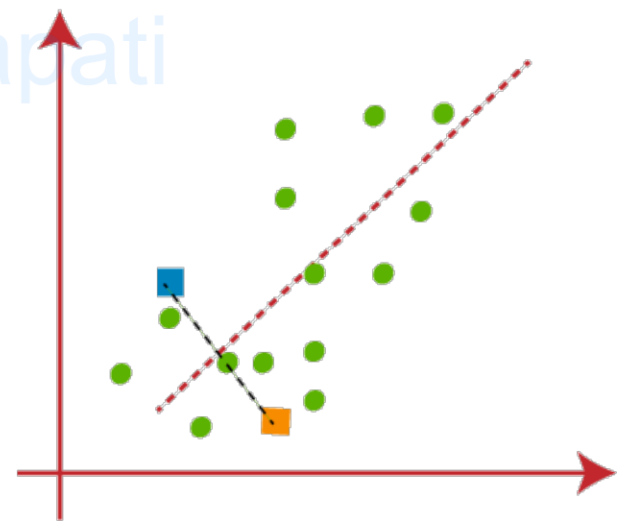


Figure3

K-Means Clustering -Working

- From the image, points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. color them as blue and yellow for clear visualization.Figure4
- we need to find the closest cluster, so repeat the process by choosing **a new centroid**. Figure5
- Reassign each datapoint to the new centroid, Again finding a median line. Figure6

Prepared by : Ritesh Prajapati

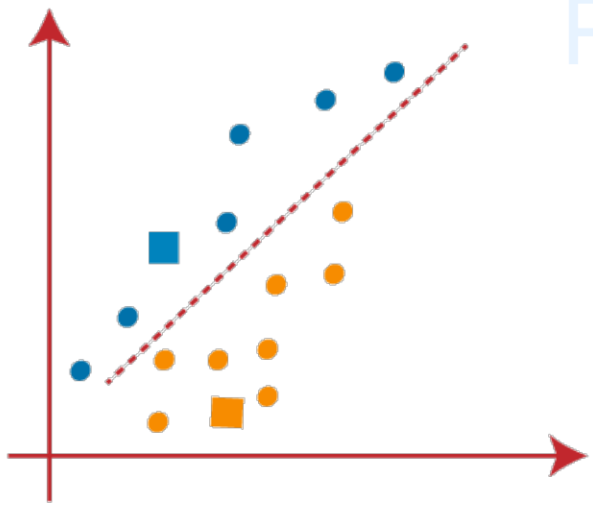


Figure4

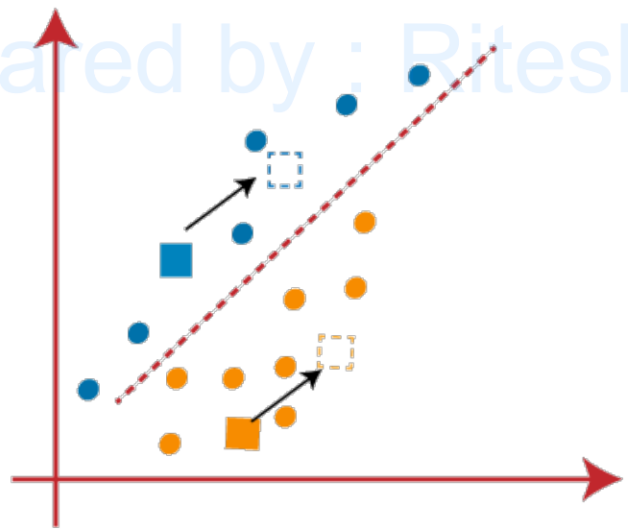


Figure5

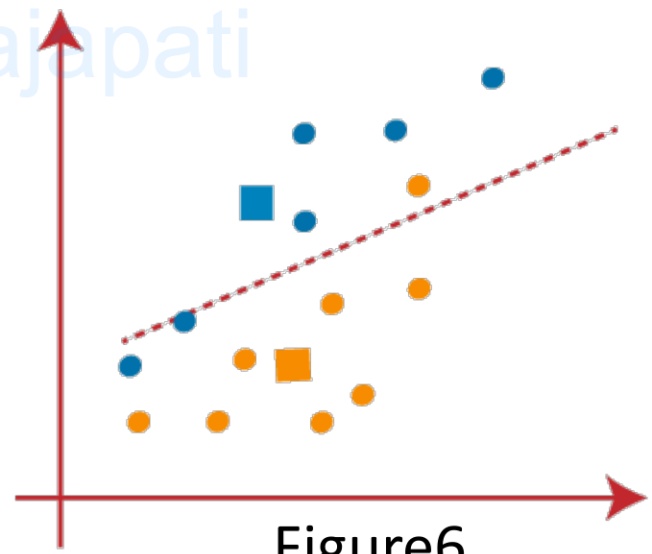
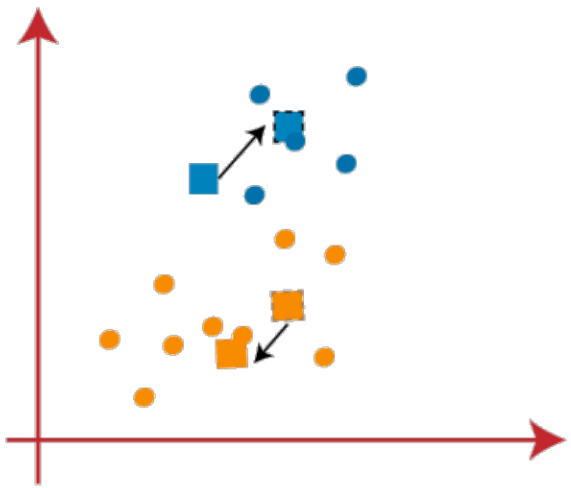


Figure6

K-Means Clustering -Working

- Repeat the process by finding the centroids. Figure7
- Again draw the median line and reassign the data points. Figure8
- There are no dissimilar data points on either side of the line, which means our model is formed. Figure9
- Remove the assumed centroids, and the two final clusters will be as shown in Figure10



Figure

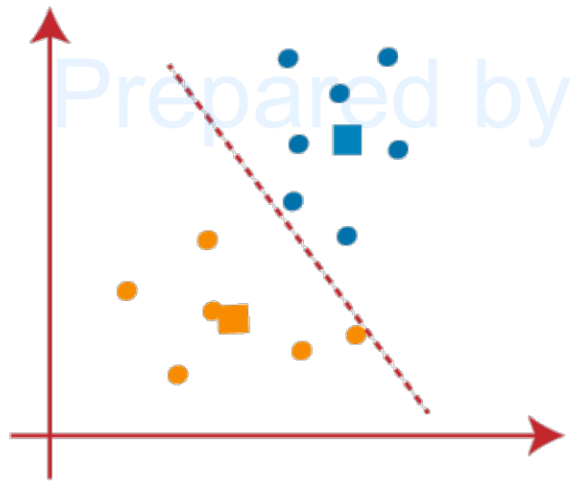


Figure8

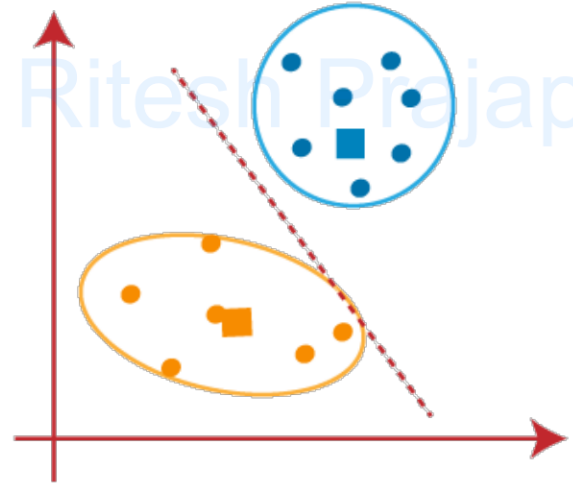


Figure9

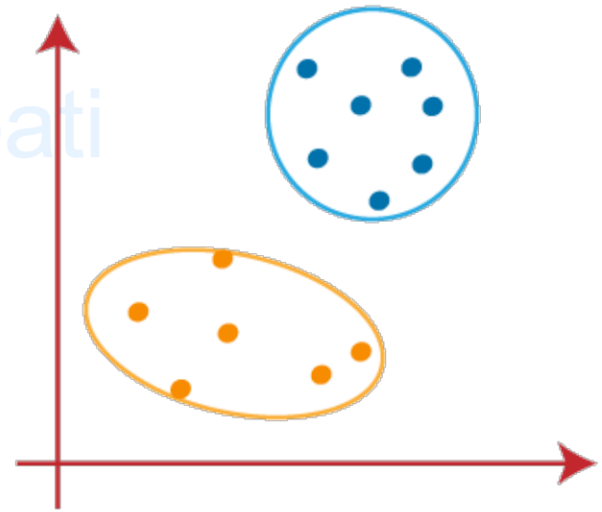


Figure10

Prepared by : Ritesh Prajapati

Unsupervised Machine Learning -Types

Association:

- Identifying associations(relationship) between data elements
- It determines the set of items that occurs together in the dataset.

Example

- Market Basket Analysis
- Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item.

Prepared by : Ritesh Prajapati

Algorithms

- Apriori Algorithm
- Eclat Algorithm-Equivalence Class Transformation.
- F-P Growth Algorithm-Frequent Pattern

Association

- Checks for the dependency of one data item on another data item and maps accordingly
- It tries to find some interesting relations or associations among the variables of dataset.
- **example of a supermarket**
- 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.

Prepared by : Ritesh Prajapati

Transaction Number	Purchased Items
1	{Bread, Milk, Egg, Butter, Salt, Apple}
2	{Bread, Milk, Egg, Apple}
3	{Bread, Milk, Butter, Apple}
4	{Milk, Egg, Butter, Apple}
5	{Bread, Egg, Salt}
6	{Bread, Milk, Egg, Apple}

Association Rule

- Association rule learning works on the concept of If and Else Statement

- If element is called **antecedent**, and then statement is called as **Consequent**.
- Relationships where we can find out some association or relation between two items is known *as* ***single cardinality*****.**
- It is **all about creating rules.**
- If the **number of items increases, then cardinality also increases** accordingly.
- The associations between thousands of data items are measured by several metrics (Support,Confidence,Lift).

Prepared by : Ritesh Prajapati

Association Rule

- Itemset:
  - One or more items are grouped together and are surrounded by brackets to indicate that they form a set called as itemset.
  - an itemset is called **k-itemset if it contains k number of items**.
  - the itemset {Bread, Milk, Egg} is a three-itemset.
- Support :
  - the number of transactions in which a particular itemset is present.
  - It denotes the frequency of occurrence for the itemset.
  - Ex: itemset {Bread, Milk, Egg} occurs together three times and thus have a support of 3.
  - If there are X datasets, then for transactions T, Support is

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

Prepared by : Ritesh Prajapati

Association Rule

- Confidence:
  - Confidence indicates how often the rule has been found to be true.
  - How often the items X and Y occur together in the dataset when the occurrence of X is already given.
  - Ratio of the transaction that contains X and Y to the number of records that contain X.
  - if we consider the association rule {Bread, Milk} → {Egg}

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

$$\begin{aligned}\text{Confidence } c(\{\text{Bread, Milk}\} \rightarrow \{\text{Egg}\}) &= \frac{\text{support count of } \{\text{Bread, Milk, Egg}\}}{\text{support count of } \{\text{Bread, Milk}\}} \\ &= \frac{3}{4} \\ &= 0.75\end{aligned}$$

Prepared by : Ritesh Prajapati

  - A low support may indicate that the rule has occurred by chance.

Association Rule

- Lift:

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

  - It is the strength of any rule.
  - Ratio of the observed support and expected support if X and Y are independent of each other.
  - 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 : Ritesh Prajapati

A-Priori Algorithm

- The Apriori algorithm uses **frequent itemsets** to **generate association rules** from the transactions.
- Using this association rule, it determines how strongly or how weakly two objects are connected.
- Frequent itemsets - items whose support is greater than the **threshold value**.
- Consider two transactions: A= {1,2,3,4,5} and B= {2,3,7}, here 2 and 3 are the frequent itemsets.

Prepared by : Ritesh Prajapati

A-Priori Algorithm Steps

- **Step-1:** Determine the support of itemsets in the transactional database, and select the minimum support and confidence.
- **Step-2:** Take all supports in the transaction with higher support value than the minimum or selected support value.
- **Step-3:** Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
- **Step-4:** Sort the rules as the decreasing order of lift.

Prepared by : Ritesh Prajapati

A-Priori Algorithm Example

- Consider this data for transaction.
  - Assume threshold(min) support value is 2.
  - Assume threshold(min) confidence value is **60%.**

TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

- **Iteration 1:**
  - Create the item sets of size of 1 and calculate their support values.
  - item 4 has a support value of 1 which is less than the min support value. So we **discard {4}.**
  - We have the final Table F1.

TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

C1

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

C1

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

F1

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

Prepared by : Ritesh Prajapati

A-Priori Algorithm Example

TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

Only Items present in F1

C2

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

F2

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

- **Iteration 2:**
  - We create itemsets of size 2 and calculate their support values.
  - All the combinations of items set in F1 are used in this iteration.
  - Itemsets having Support less than 2 are eliminated again. In this case **{1,2}.**
  - We have the final Table F2.
- **Iteration 3:**
  - We create itemsets of size 3 and calculate their support values.
  - All the combinations of items set in F2 are used in this iteration.
  - Itemsets having Support less than 2 are eliminated again. In this case **{1,2,3}, {1,3,5}.**
  - We have the final Table F3.

TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

F3

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

Prepared by : Ritesh Prajapati

A-Priori Algorithm Example

TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

F3

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

C3

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

- **Iteration 4:**
  - Using sets of F3 we will create C4.
  - Support of this itemset is less than 2, So stop here and the final itemset will be F3.

- With F3 we get the following itemsets:
- **For Itemset = {1,3,5}**, subsets are {1,3}, {1,5}, {3,5}, {1}, {3}, {5}
  For Itemset = {2,3,5}, subsets are {2,3}, {2,5}, {3,5}, {2}, {3}, {5}
- **Applying Rules:** We will create rules and apply them on itemset F3.
  - For every subsets S of **Itemset I**, output the rule
  - S –> (I-S) (means S recommends I-S)
  - if **support(I) / support(S) >= threshold confidence value ➔ Rule selected**

Prepared by : Ritesh Prajapati

A-Priori Algorithm Example

- **{1,3,5}**

- **Rule 1:** {1,3} –> ({1,3,5} – {1,3}) means 1 & 3 –> 5
- Confidence = support(1,3,5)/support(1,3) = 2/3 = **66.66% > 60%**
- Hence Rule 1 is **Selected**

- **Rule 2:** {1,5} –> ({1,3,5} – {1,5}) means 1 & 5 –> 3
- Confidence = support(1,3,5)/support(1,5) = 2/2 = **100% > 60%**
- Rule 2 is **Selected**

- **Rule 3:** {3,5} –> ({1,3,5} – {3,5}) means 3 & 5 –> 1
- Confidence = support(1,3,5)/support(3,5) = 2/3 = **66.66% > 60%**
- Rule 3 is **Selected**

Prepared by : Ritesh Prajapati

A-Priori Algorithm Example

- **Rule 4:** {1} –> ({1,3,5} – {1}) means 1 –> 3 & 5
- Confidence = support(1,3,5)/support(1) = 2/3 = **66.66% > 60%**
- Rule 4 is **Selected**

- **Rule 5:** {3} –> ({1,3,5} – {3}) means 3 –> 1 & 5
- Confidence = support(1,3,5)/support(3) = 2/4 = **50% <60%**
- Rule 5 is **Rejected**

- **Rule 6:** {5} –> ({1,3,5} – {5}) means 5 –> 1 & 3
- Confidence = support(1,3,5)/support(5) = 2/4 = **50% < 60%**
- Rule 6 is **Rejected**

Prepared by : Ritesh Prajapati

A-Priori Algorithm Question

- Find the frequent itemsets and generate the association rules using the Apriori algorithm

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Prepared by : Ritesh Prajapati

Given: Minimum Support= 2, Minimum Confidence= 50%

A-Priori Algorithm Question

Itemset	Support_Count
A	6
B	7
C	5
D	2
E	1

Itemset	Support_Count
A	6
B	7
C	5
D	2

Itemset	Support_Count
{A, B}	4
{A,C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

Itemset	Support_Count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2

Prepared by : Ritesh Prajapati

A, B, C, D

Itemset	Support_Count
{A, B, C}	2
{B, C, D}	1
{A, C, D}	0
{A, B, D}	0

only one combination, {A, B, C}.

A-Priori Algorithm Question

Rules	Support	Confidence
A ^B → C	2	Sup{(A ^B) ^C}/sup(A ^B)= 2/4=0.5=50%
B^C → A	2	Sup{(B^C) ^A}/sup(B ^C)= 2/4=0.5=50%
A^C → B	2	Sup{(A ^C) ^B}/sup(A ^C)= 2/4=0.5=50%
C→ A ^B	2	Sup{(C^(A ^B)}/sup(C)= 2/5=0.4=40%
A→ B^C	2	Sup{(A^(B ^C)}/sup(A)= 2/6=0.33=33.33%
B→ B^C	2	Sup{(B^(B ^C)}/sup(B)= 2/7=0.28=28%

Prepared by : Ritesh Prajapati

the first three rules A ^B → C, B^C → A, and A^C → B can be considered as the strong association rules

Supervised Learning	Unsupervised Learning
• Supervised learning algorithms are trained using labeled data.	• Unsupervised learning algorithms are trained using unlabeled data.
• Supervised learning model takes direct feedback to check if it is predicting correct output or not.	• Unsupervised learning model does not take any feedback.
• Supervised learning model predicts the output.	• Unsupervised learning model finds the hidden patterns in data.
• In supervised learning, input data is provided to the model along with the output.	• In unsupervised learning, only input data is provided to the model.
• The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	• The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
• Supervised learning needs supervision to train the model.	• Unsupervised learning does not need any supervision to train the model.

Prepared by : Ritesh Prajapati

Supervised Learning	Unsupervised Learning
Supervised learning can be categorized in Classification and Regression problems.	Unsupervised Learning can be classified in Clustering and Associations problems.
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

Prepared by : Ritesh Prajapati

THANK YOU

Prepared by : Ritesh Prajapati