UNIT 2 → Supervised Learning

ML Algo ::

1. Supervised :: Based on feature and target values, build a model for predicting the target given new instance.  
   a. Regression - target is continuous value.
   1. Simple
      1. Multiple
      2. Polynimial
      3. Regularization
      4. Ridge-lasso b. Classification - target is categorical variable.
      5. Logistic Regression
      6. KNN
      7. DT
      8. SVM
      9. Naive Bayes
2. Unsupervised :: We take only feature values.  
   The dataset only has features x1,x2...xn.  
   We take only X and NO Y vector.  
   There is no target variable.  
   Use only X matrix during training and build a model. THe objective is model is used to find ::
   1. Patterns in data
      1. Groups in data - Clusters  
         | Clustering ::The datapoints are grouped called clusters.  
         Eg : K-means, Hierarchial clustering.
      2. Dimensionality Reduction :: Each feature is called a dimension in a dataset. If the dataset has too many features, we are going to reduce the no. of features, because::
         1. There may be irrelevant features.
         2. Duplicate features.
         3. Features may be combined to reduce the no. of features.
         4. Less no of features, Reduce the training time and processing time. Algorithms for DR is LCA nad PCA

Dimensionality Reduction ::

Reduce the no of features.

If a dataset has N features, reduce it to K where (K<N).

1. Feature Selection :: Select some of the features and others are discarded. No. of methods ::
   1. Lasso Regression
2. Dim Reduction :: Create few new features from the set of features.  
   With the help of Eigen Vectors - principal components of data are derived by retaining the variations in the data. Eg ::
   1. PCA Algorithm
   2. Auto Encoders(Neural network ie, DL Method)

# **Clustering ::**

1.Given the feature matrix, divide the datapoints into groups. The groups are called clusters.

1. It is an unsupervised methods.
2. Dataset has only x matrix and no Y vector.  
   |**Application** ::
   1. FB users can be clustered based on the communication among the users.
   2. News articles are clustered into sports, political news, business news etc..
   3. Mails can also be grouped based on the business, personal etc...

Algorithms can be broadly classified into ::

1. Partitional Clustering Algorithms
2. Hierarchial CLustering
3. Density based clustering

**1. Partitional Clustering Algorithms** :: The clusters are non-overlapping. ie, any datapoint belong to only one cluster.

Eg :: 1. K means, k-medoids.

**2. Hierarchial CLustering** :: Clusters form hierarchy. Eg :: Algomotive

* Bottom up Approach : Each datapoint is a cluster and then go on merging until desired no. of clusters are formed.
* Top down approach :: Strat the entire datset as one cluster ant then start dividing into smaller clusters until desired no. of clusters are formed or each cluster has one datapoint.

**3. Density based clustering** :: Clusters are formed based on how close the datapoints are.  
Eg :: DBScan, Optics

Clustering is given datapoints, divide the datapoints into groups called clusters.

It is an **unsupervised learning method**.

We consider only feature matrix 'x' and no target vector.

**K-Means** is the Clustering algorithms based on Distance measure to form the clusters.

It uses Euclidian distance.

Datapoint is a tupple with all feature values.

Every cluster will have cluster header also called as **Cluster Centroid**.

Procedure Steps ::

1. Arbitarily take 'n' centroids and calculate the distance of each centroid from every datapoint.
2. Consider the least distance and assign it to that centroid.
3. Calculate the

**|Algorithms ::**

1. Choose the k value(no. of clusters).
2. Choose initial K no. of centroids.
3. Repeat ;
   1. Calculate the distance for each datapoint to k centroids.
   2. Assign the datapoint to that cluster that is closest.
   3. Calculate the average distance of the datapoints belonging to each cluster.
   4. Do the (3) for all the clusters.
   5. These average distance become new centroinds.
4. Until convergence.

**|Convergence Criteria ::**

1. Until datapoints do not change the clusters.
2. After spcified no. of iterations.
3. Centroids are not changing.