Machine learning

Unit : 3

-Prof. Mayur Ghorpade

ML Unit: III
Supervised Learning: Regression

Q. Explain and Compare training error and generalization error.

→ **Training Error:**

- In machine learning, training a predictive model means finding a function which maps a set of value x to a value y.
- Training error is calculated as:

$$E_{train} = \frac{1}{n} \sum_{i=1}^{n} \text{error}\,(f_D(x_i), y_i)$$

where,

$n \rightarrow$ No. of training example.

$f_D(x_i) \rightarrow$ Predictive value.

$y_i \rightarrow$ Actual or True value.

$(f_D(x_i), y_i) \rightarrow$ These two values are same or not, if not then find difference.

Generalization Error:

- For Supervised learning applications in machine learning and statistical learning theory, generalization error is a measure of how accurately an algorithm is able to predict outcome values for previously unseen data.

$$E_{gen} = \int \text{error}\,(f_D(x_i), y_i)\, P(y, x)\, dX$$

where,

$f_D(x_i) \rightarrow$ Predictive value

$y_i \rightarrow$ Actual value.

$P(X, Y) \rightarrow$ How often we expect to see such X & Y.

Q. Explain terms Bias, Variance, Generalization.

→ **Bias:**

- Bias is a phenomenon that skews the result of an algorithm in favour or against an idea.
- Bias is considered a systematic error that occurs in the ML model itself due to incorrect assumption in the ML process.
- Low Bias :- A Low Bias model will make fewer assumption about the form of the target function.
  - Closely match the training data set.
- High Bias :
  - ~~It makes~~ high bias makes more assumptions.
  - It cannot perform well on new data.

Variance:

- Variance tells that how much a random variable is different from its expected value.
- Variance error are either low variance or high variance.
- Low variance :
  - Small variation in the prediction of target fun^n
- High Variance :
  - large variation in the prediction of target fun^n.
  - High variance model leads to overfitting.
  - Increases model complexities.

Different combination of Bias-Variance :

- Low Bias, low variance : Ideal ML Model (Not possible)
- Low Bias, High Variance : Occurs when the model learns with a large no. of parameters. (overfitting).
- High Bias, Low Variance : Occurs when few no. of parameter (underfitting)
- High Bias, High Variance : Prediction are inconsistant & also inaccurate on average.

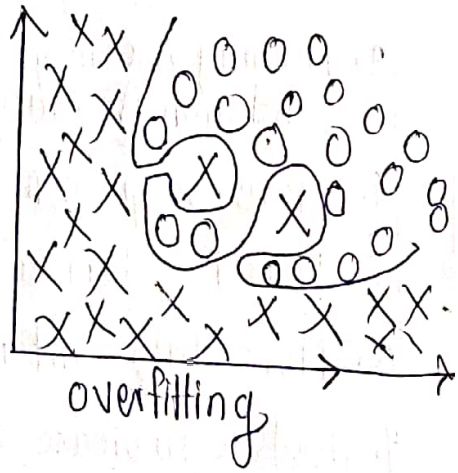
Generalization:

- Generalization is a term used to describe a model's ability to react to new data.
- After being trained on a training set, a model can digest new data and make accurate prediction.
- Model is accurately train that time generalization not possible.
- It will make inaccurate prediction when given new data, it making model useless even it able to give accurate prediction for the training data is called overfitting.
- Underfitting happens when a model has not been trained enough on the data.

Q. Explain Overfitting with the techniques to reduce the overfitting?

→ Overfitting:

- When a model gets trained with so much data, it starts learning noise & inaccurate data entries in our dataset.
- So High Variance in test data.
- It causes overfitting.
- Simply model does not make accurate prediction on testing data called overfitting.
- Reasons of overfitting:
  - High Variance & Low Bias
  - The model is too complex
  - Size of training data.



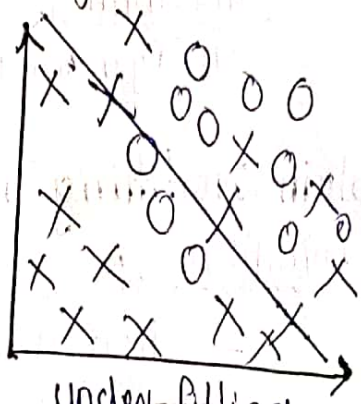
Techniques to reduce overfitting:

1. Increase Training Data.
2. Reduce model complexity.
3. Early stopping during the training phase.
4. Ridge Regularization & Lasso Regularization.
5. Use dropout for neural networks to tackle overfitting.
6. Ensembling

Q. Explain term Underfitting with techniques to reduce underfitting?

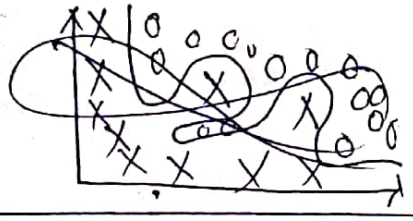
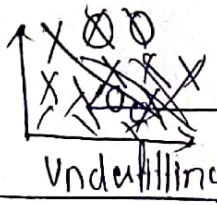
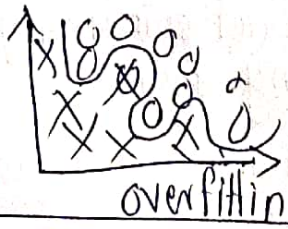
→ Underfitting:

- Machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
- It only perform well on training data but perform poorly on testing data.
- Underfitting destroy accuracy of ml model.
- Reasons for Underfitting:
  - High Bias and low variance.
  - Size of the training data is not enough.
  - The model is too simple.



Techniques to reduce underfitting:

1. Increase model complexity.
2. Remove noise from the data.
3. Increase the no. of features.
4. Increasing training time of model.

Underfitting	Overfitting
1. High Bias, Low Variance	1. High Variance, low Bias
2. Performs poorly on training data, well on testing data.	2. Performs poorly on testing data, well on training data.
3. Training Accuracy → Poor Validation Accuracy → Poor	3. Training Accuracy → Good Validation Accuracy → Poor
4. More complex model	4. More simple model.
5. Less Regularization	5. More Regularization
6. More data can't help.	6. More data can help.
7. Need to increase complexity.	7. Need to reduce complexity.
8.   Underfitting.	8.  overfitting.

Q. What is Regression and List the type of Regression.

→ **Regression :-**

- Regression is supervised learning techniques.
- ~~It is used to find~~
- Regression analysis is a statistical method to model the relationship between a dependent and independant variables with one or more variable.
- It is used for prediction, forcasting, time series modelling.
- example : Linear regression.

Types of Regression :

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression

Q. Explain Linear Regression.

→ **Linear Regression :-**

- Linear regression is statistical regression method which is used for predictive analysis.
- It is one of the very easy and simple algorithm which works on regression and shows the rein betn the continuous variables.
- It is used for solving the regression problem in ML.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis) hence, called linear regression.
- $\boxed{Y = aX + b}$

where:

Y → Dependent variables.

X → Independent variable

a & b → linear coefficient

Application:

- Analyzing trends & sales estimates.
- Salary forecasting.
- Real estate prediction.
- Arriving at ETAs in traffic.

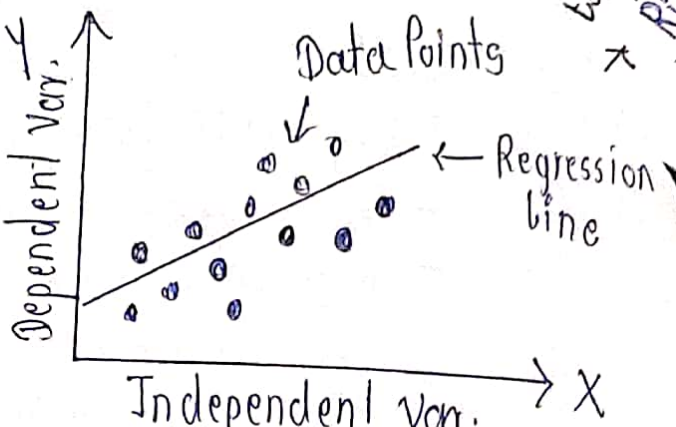


fig. Linear Regression.

Q. Explain Lasso Regression.

→ Lasso Regression:

- Lasso regression is regularization technique to reduce the complexity of the model.
- It is similar to Ridge regression except that penalty ~~that~~ term contains only the the absolute weights instead of the square of weights.
- Since it takes absolute value, hence it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called L1 Regularization.
- Equation

$$L(x,y) = \min \left(\sum_{i=1}^{n} (y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n} |w_i| \right)$$

- Application:

  Lasso has been applied in economics and finance.

Explain Ridge Regression.

→ **Ridge Regression:**

- Ridge Regression is the one the most robust versions of linear regression in which a small amount of bias is introduced so that we can get better long term predictions.
- The amount of bias added to the model is knows as Ridge Regression penalty.
- we can compute this penelty term by multiplying with the lambda to the squared weight of each individual features.
- Equation.

$$L(x,y) = \min\left(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n}(w_i)^2\right)$$

- Ridge regression is a regularization technique, which is used to reduce the complexity of the model.
- It is also called as L2 regularization.
- It helps to solve the problems if we have more parameters than samples.

Q. Explain Gradient Descent Algorithm.

→ **Gradient Descent Algorithm:**

- Gradient Descent is an optimization algorithm which is used to minimize the cost of function for many ML algorithms.
- Gradient Descent algorithm is used for updating the parameters of the learning models.
- Types:
  - Batch Gradient Descent:
    used for small training dataset.
  - Mini-Batch Gradient Descent:
    for large training set.
  - stochastic Gradient Descent: Process one training dataset set first

∴ Parameters will be updated after each iteration.

Q. Explain MAE, MSE, RMSE, R2.

→ MAE (Mean Absolute Error):

- MAE represents the average of the absolute difference between the actual & predictive value in the dataset.

$$MAE = \frac{1}{N} \sum |y - \hat{y}|$$

Where,

N → No. of data Points

y → Actual output

$\hat{y}$ → Predictive output

- Advantages:
  - MAE you get in the same unit as the output variable.
  - It is most Robust to outliers.
- Disadvantages:
  - Graph of MAE is not differentiable so we apply different optimizers like GDA.

MSE (Mean Squared Error):

- MSE represent the average of the squared difference bet'n the original and predicted values in the dataset.
- It measures the variance of the residuals.
- we perform squared to avoid the cancellation of negative terms. & it is the benefit of MSE.

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

- Advantages:
  - Graph of MSE is differentiable.

- Disadvantages:
  - The value you get after calculating MSE is a squared unit of output.
  - It is not robust to outliers.

RMSE (Root Mean Squared Error):

- RMSE is the square root of MSE.
- It measures standard deviation of residuals.

Advantages:
- Output value you get is in the same unit as the required output variable.

Disadvantages:
- Not robust to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$$

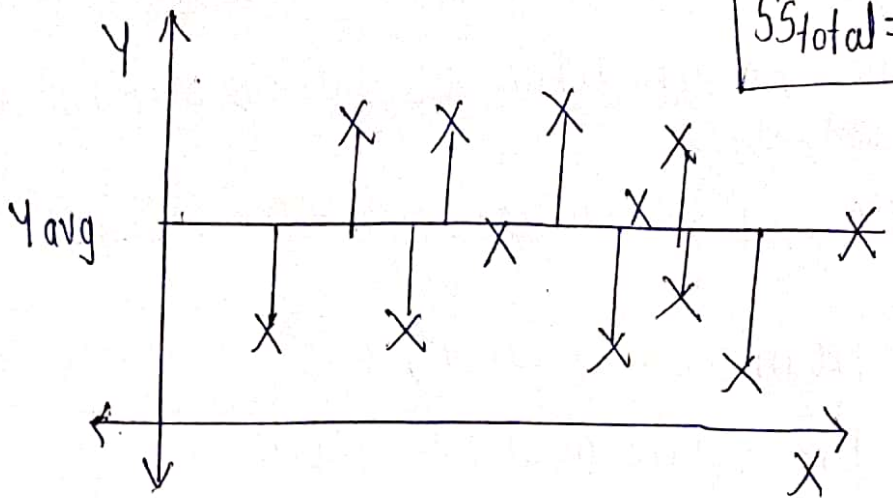
$$RMSE = \sqrt{MSE}$$

* R-Squared:-

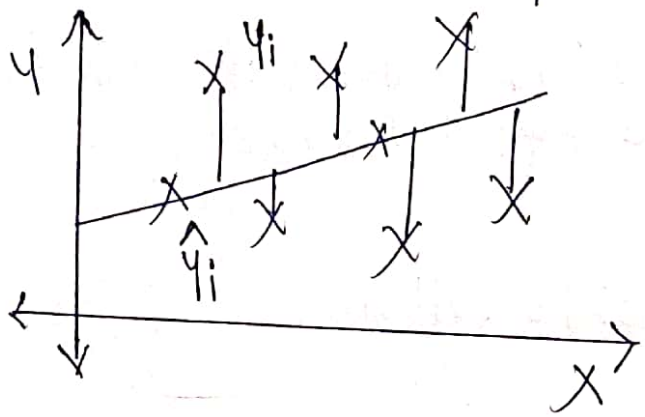
- R-squared is a statistical measure that represents the goodness of fit of a regression model.
- The ideal value for R-squared is 1.
- R squared is a comparison of the residual sum of squares (SSres) with the total sum of squares (SStot)
- The total sum of squares is calculated by summation of squares of perpendicular distance bet^n data points & the average line.

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)}$$

$SS_{total} = SUM(y_i - y_{avg})$



The residual sum of squares is calculated by the summation of squares of perpendicular distn betn data points & the best fitted line.



$$SS_{res} = SUM(y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res} → Residual Sum of squares
SS_{tot} → total sum of squares.

Q. Why we require evaluation metrics?

→ - ML model cannot have 100 per cent eff. otherwise the model is known as a biased model which further includes the concept of overfitting & underfitting.
- It is necessary to obtain the accuracy on training data, But it is also important to get a genuine & approximate result on unseen data.
- So to build & deploy a generalized model we require to evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, & obtain a better result.

Following table shows the midterm & final exam grades obtained for student in a database course, Use the method of least squares using regression to predict the final exam grade of a student who received 86 in the mid-term exam.

Midterm exa (X)	72	50	81	74	94	86	59	83	86	33	88	81
Final Exam (Y)	84	63	77	78	90	75	49	79	77	52	74	90

→

Sr No	X	Y	XY	X^2
1	72	84	6048	5184
2	50	53	2650	2500
3	81	77	6237	6561
4	74	78	5772	5476
5	94	90	8460	8836
6	86	75	6450	7396
7	69	49	2891	3481
8	83	79	6557	6889
9	86	77	6622	7396
10	33	82	1716	1089
11	88	74	6512	7744
12	81	90	7200	6561
Total	887	878	67209	69113

Eqn of Reg line is $Y' = aX + b$

$$a = \frac{n\Sigma XY - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = 0.65$$

$$b = \frac{1}{n}(\Sigma y - n X \Sigma x) = 25.12$$

$$Y' = 0.65X + 25.12$$

The final exa grade of a student who received 86 in the mid term exam.

$$Y' = 0.65 \times 86 + 25.12$$

$$Y' = 81.02$$

Q. Given the following data for the sales of car of an automobile company for six consecutive years. Predict the sales for next two consecutive years.

Years (X)	2013	2014	2015	2016	2017	2018
Sales (Y)	110	100	250	275	230	300

→ we will take $t = X - 2013$

Sr. No	t	Y	tY	t^2
0	0	110	0	0
1	1	100	100	1
2	2	250	500	4
3	3	275	825	9
4	4	230	920	16
5	5	300	1500	25
Total	15	1265	3845	55

Eqn. of regression line is

$$Y' = at + b \qquad a = 39,\ b = 113.33$$

eqn. of line become

$$Y' = 39t + 113.33$$

The sales of company for next two years,

$X = 2019, t = 6, \quad Y' = 39 \times 6 + 113.33 = 347.33$

$X = 2020, t = 7, \quad Y' = 39 \times 7 + 113.33 = 386.33$

Unit : 4

UNIT - 4] Supervised Learning - Classification.

Q. What is classification ? List the types of classification?

→ **Classification :-**

- classification algorithm is the supervised learning technique that is used to identify the category of new observations on the basis of training data.
- In classification, a ~~gro~~ program learns from the given dataset of observations & then classifies the new observations into no. of classes or groups.

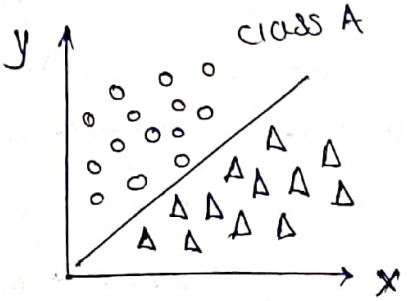


fig. classification.

Types :-

① Binary classification :-
- The problem has only two possible outcomes.
- e.g. yes or no, 0 or 1, etc.

② Multi-class classifier :-
- The problem has more than two possible outcome.
- e.g. classification of music.

Types of ML classification Algorithm :-

- KNN
- SVM
- Naive Bias
- Decision tree
- random forest classification.

Q. Explain KNN Algorithm ?

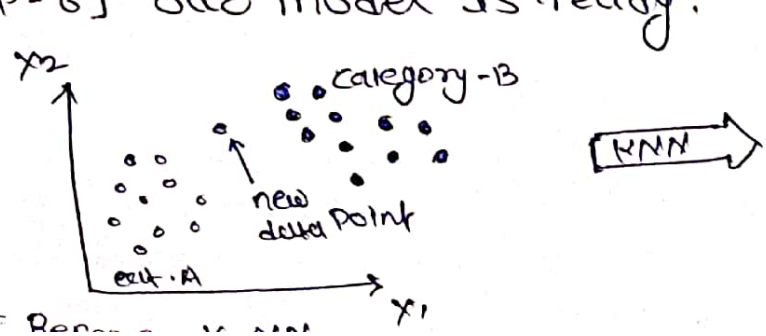
→ K-nearest neighbour :-

- KNN is one of the simplest machine learning algorithm based on supervised learning technique.
- KNN algorithm assumes the similarity between the new data & available data & put the new data into category that most similar to the available categories.
- KNN algorithm stores all the available data and ~~the~~ classifies a new data point based on the similarity.
- This means when new data appears then it can be easily classified into a well suite category by KNN algorithm.
- It is used for both classification as well as regression.
- It is also called as a lazy learner algorithm.

Steps :-

- Step-1] : select the number K of neighbors.
- Step-2] calculate the euclidean distance of K number of neighbors.

  Euclidean distance betn. A & B = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$
- Step-3] Take the K nearest neighbors as per the calculated Euclidean distance.
- Step-4] Among these K neighbors, count the number of the data points in each category.
- Step-5] Assign the new data points to that category for which the number of neighbors is maximum.
- Step-6] our model is ready.



Advantages of K-NN algorithm :-

- It is simple to implement.
- It is robust to the noisy training data.
- More effective on large training data.

Disadvantages :-

- Always needs to determine the value of K, which may be complex some time.
- High computation cost.

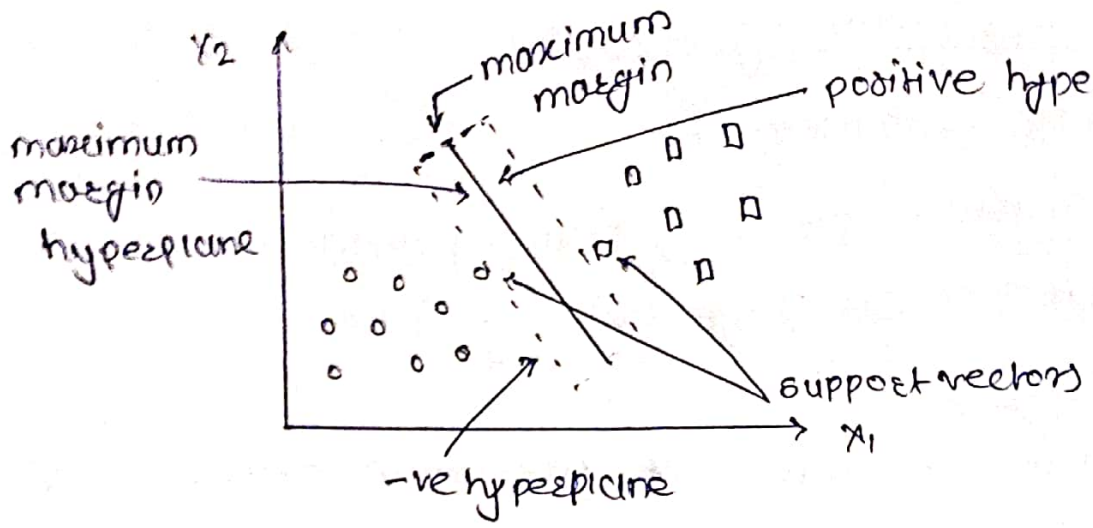
Applications :-

- Finance - credit rating, loan management, stock market.
- medicine

Q. Define SVM - Explain how margin is computed & optimal hyper plane is decided ?

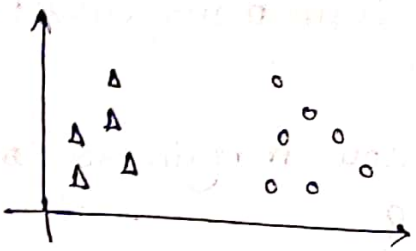
→ SVM -

- Support vector machine is one of the most popular supervised learning algorithm.
- used in classification as well as Regression problems. but primarily used for classification.
- the goal of the SVM algorithm is to create the best line or dicision boundary that can segregate n-dimensional space into classes.
- so that we can easily put new data point in the correct category in future.
- this best decision boundary is called a hyper plane.
- SVM chooses the extreme points that help in creating the hyperplane.
- these extreme are called as support vectors.
- the boundary which devide classes & helps to find the support vector in example is called Optimal Dicision Boundary

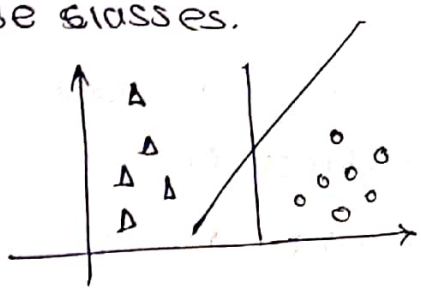


Example :-

Suppose we have a dataset has two tags (green & blue), & then classify this by using SVM.

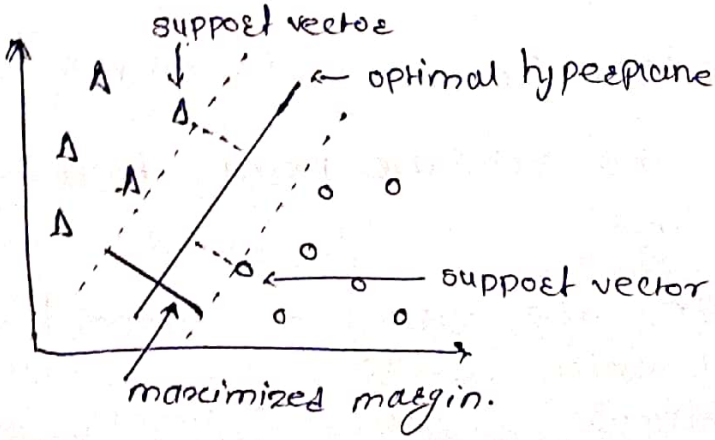


here it is 2-d space so using straight line, we can easily separate these two classes. but it has multiple lines that can separate, these classes.



Hence the SVM helps to find the best line or decision boundary, this best boundary is called hyperplane.

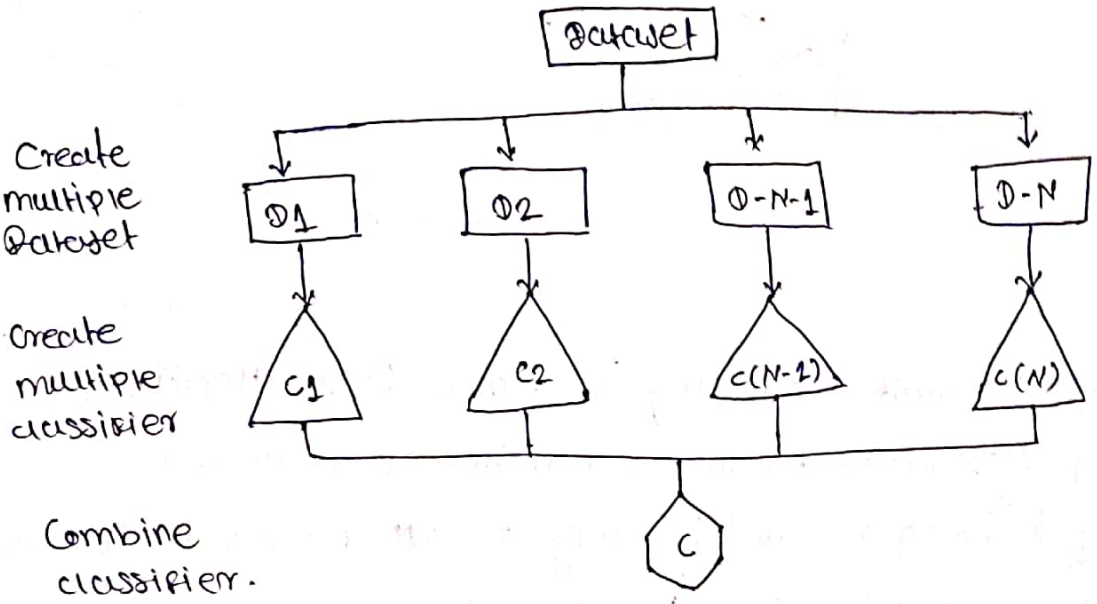
The data points which are closest to the line is called support vector.



Explain: Ensemble learning & explain the term bagging & boosting?

→ Ensemble Learning :-

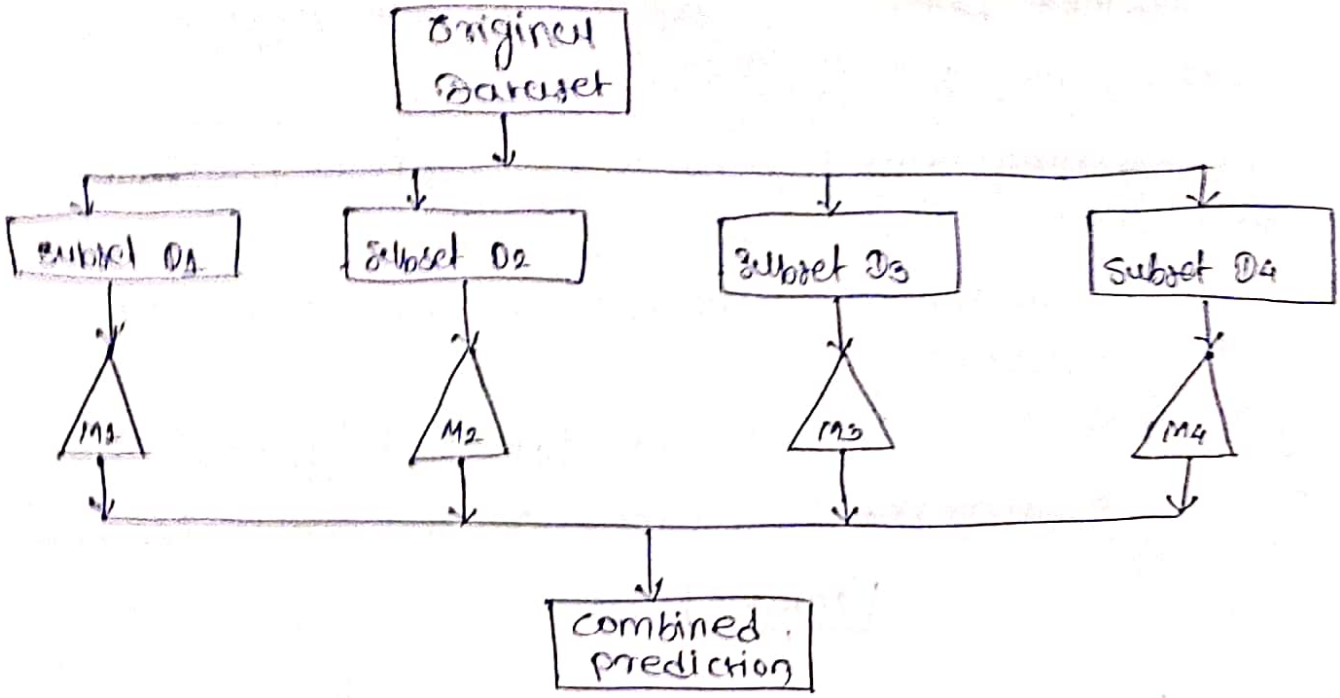
- Ensemble learning helps improve ML results by combining several models.
- The model that contribute to the ensemble are called as ansemble members.
- They may or same type or & of different types.
- They may or & may not be trained on the same trained data.



Bagging :-

- Bagging is the combine approach of Bootstrap & aggregation.
- It is used to reduce the variance of a dicision tree.
- Implementation steps of Bagging :

1. multiple subset are created from the original data set with equal tuples, selecting observations with replacement.
2. A base model is created on each of these subsets.
3. Each model is learned in parallel from each training set and independent of each other.
4. The final predictions are determined by combining the predictions from all the models.



Boosting :-

- Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifier.
- It is done by building a model by using a weak models in series.
- Firstly, model is created by using training data.
- Then second model built which tries to correct the errors present in the first model.
- This procedure is continued & models are added until either the complete training data set is predicted correctly or the maximum no. of models are added.
- Adaboost was the first really successful boosting algorithm developed for the purpose of binary classification.

Explain random forest algorithm in detail?

Random forest :-

- Random forest is a supervised ML algorithm that is used for classification & regression
- It is based on ensemble learning.
- Random forest is a classifier that contains a number of decision tree on various subsets of the given datasets & takes the average to improve the predictive accuracy of that dataset.

- Steps :-

1. Select random K data points from the training set.
2. Build the decision trees associated with the selected data points.
3. Choose the number N for decision trees that you want to build.
4. Repeat step 1 & 2.
5. For new data points, find the predictions of each decision tree, & design the new data points to the category that wins the ~~majority~~ majority votes.

Advantages :-

- Random forest is capable of performing both classification & regression.
- It is capable of handling large datasets with high dimensionality.
- prevent the overfitting problem.

Disadvantages :-

- It is not more suitable for regression tasks.

Applications :-

1) Banking.
2) medicine.
3) Land use.
4) marketing.

Q. Difference bet? Binary classification vs multiclass classif?

→

Parameters	Binary classification	multi-class classification
No. of classes	It is a classification of 2 groups : i.e. classifies objects in at most two classes.	There can be any no. of classes in it, i.e. classifies the object into more than two classes.
Algorithms used	• Logistic regression • K-nearest neighbor • decision tree • SVM • naive bias	• KNN • decision tree • naive bias • Random forest • gradient boosting.
Example	• email spam detection • churn prediction	• face classification • optical character recognition.

Q. explain varient of multiclass classification : one vs one one vs all ?

→ **one vs one :-**

- In one vs one reduction, one trains $\frac{K(K-1)}{2}$ binary classifier for a K-way multiclass problem.
- each recieves the samples of a pair of classes from the original training set and learns to distinguish these two classes.
- At predictive time, a voting scheme is applied :

all $\frac{K(K-1)}{2}$ classifiers are applied to an unseen sample & the class that gets the highest number of +1 predictions, gets the predicted by combined classifier.

- Like one vs all one suffers from ambiguites in that some regions of its input space may recieve the same number of votes.

one vs all :- (one vs rest)

- one vs all strategy involves training single classifier per class with the samples of that class as positive samples & all other samples are negative

- Inputs :
  1) L, a learner.
  2) samples X.
  3) lables y where $y_i \in \{1, \ldots K\}$ is the label for the sample X.

output :
  1) a list of classifiers f_k for $k \in \{1, \ldots K\}$

Procedure :
  1) construct a new label vector Z where $z_i = y_i$ if $y_i = k$ & $z_i = 0$ otherwise.
  2) Apply L to X, Z to obtain f_k.

Q. Explain terms : Accuracy, Precision, Recall, Fscore?

→ confusion matrix :

	actual value: positive	negative
positive	TP	FP
negative	FN	TN

Accuracy :-

- Accuracy is the fraction of predictions our model got right.

$$\text{Accuracy} = \frac{\text{No. of correct prediction}}{\text{Total number of prediction.}}$$

$$\text{Accuracy} = \frac{TP + TN}{Tp + TN + Fp + FN}$$

Precision :-

- It is the ratio of correctly positive samples (True positive) to a total number of classified positions samples (either correct or incorrect).

$$\text{Precision} = \frac{Tp}{Tp + FP}$$

Recall :- (sensitivity).

- The recall is calculated as the ratio bet? the number of positive samples correctly classified as positive to the total number of positive samples.

$$\text{Recall} = \frac{TP}{TP + FN.}$$

Fscore :-

- precision and Recall are combined into a single measure to form F measure (F-score).

$$\text{Fscore} = \frac{2 * \text{precision} * \text{Recall}}{\text{precision} + \text{Recall.}}$$

what is cross validation in ML?

- Cross Validation :-

cross validation is the technique for validating the model efficiency.

- It is a technique how a statistical model generalizes to an independent dataset.

- Methods used for cross validation :-

1. validation set approach
2. K-Fold cross validation
3. leave-p out cross validation
4. Leave one out cross validation

Q. Explain terms Micro Average - precision, Recall & Fscore.

→ Micro-Average precision :-

- micro precision measure the precision of the aggregated contributions of all classess.
- It is micro average precision.
- micro averaging is used when a problem has more than 2 labels that can be true

$$\text{micro precision} = \frac{(TP)\ sum}{(TP)sum + (FP)sum}$$

Micro average Recall :-

- Ratio bet? the no. of +ve samples correctly classified as positive to the total number of positive samples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Micro F1 Score :-

- Micro F1 score is used to assess the quality of multilabel binary problems
- It measures the f1-score of the aggregated contibutions of all classes.

- Micro F_1 score is defined as the harmonic mean of the precision & recall.

$$\text{Micro } F_1\text{-Score} = 2\left[\frac{(\text{micro precision}) \cdot (\text{micro Recall})}{(\text{micro precision}) + (\text{micro Recall})}\right]$$

Q. Explain terms macro average precision, Recall & F-score?

→ Macro Average precision :

- we calculate the precision for each class separately in an one vs all way.
- Then we take average of all precision value.

e.g. 3 classes a, b, c.

we calculate P_a, P_b, P_c

macro average will be, $\frac{P_a + P_b + P_c}{3}$

Macro Average Recall :-

- macro recall measures the average recall per class.
- Recall is a metric used in binary classification problem

$$\text{Recall} = \frac{TP}{Tp + FN}$$

Macro - Average F_1 score :-

- The macro F_1 score is computed using the arithmetic mean of all the per class P_1 scores.
- It is used for overall performance.
- macro average f score is the harmonic mean.

$$F_1 \text{ Score} = 2\left[\frac{\text{Precision} \times \text{Recall}}{\text{Precision} \times \text{Recall}}\right]$$

Unit: 5

Unit V: Unsupervised Learning

Q. Define Clustering. Explain k-means clustering algorithm.

→ **Clustering :-**

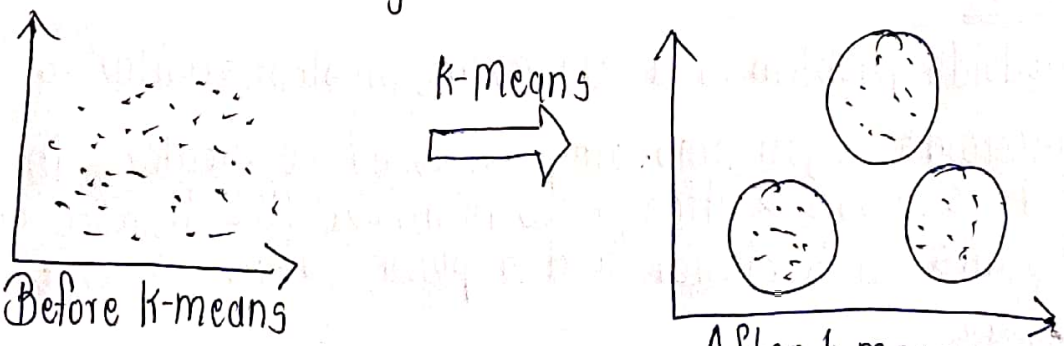
- Clustering is a technique in which the data points are arranged in similar group dynamically without any pre-assignment of groups.
- Cluster means group of the similar things.

* **k-means Clustering Algorithm :-**

- It is unsupervised learning algorithm.
- k-means is one of the popular clustering algorithm.
- It helps to form clusters from the given dataset.
- Here k → defines the number predefined cluster.

-Working :

1. Determine the number of k to decide the no. of cluster.
2. Select random k points or centroids.
3. Assign each data points to nearest centroid.
4. Calculate the varience & place new centroid of each cluster
5. Repeat step 3.
6. If any reassignment occurs, then go to step4 else go to END. or FINISH.
7. The model is ready.



Advantages :- -Relatively efficient
- Adopt easily to new example.

Disadvantages :

- Difficult to predict k-value.
- With Global Cluster didn't work well.

Problem: Apply k-means algorithm on given data for k=3 Use Explain
C1(2), C2(16) & C3(38) as initial cluster centres.
Data: 2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 23, 25, 30.

→ $C_1 = 2$, $C_2 = 16$ $C_3 = 38$

The numbers which are close to mean are grouped into resp. clusters.

$K_1 = \{2, 4, 6, 3\}$, $K_2 = \{12, 15, 16, 14, 21, 23, 25\}$, $K_3 = \{31, 35, 30\}$

Again calculate new mean for new clusters group.

$C_1 = 3.75$ $C_2 = 18$, $C_3 = 32$

New clusters

$K_1 = \{2, 4, 6, 3\}$, $K_2 = \{12, 15, 16, 14, 21, 23, 25\}$ $K_3 = \{31, 35, 30\}$

$C_1 = 3.75$, $C_2 = 18$, $C_3 = 32$

clusters remains unchanged.

Final clusters.

$K_1 = \{2, 4, 6, 3\}$, $K_2 = \{12, 15, 16, 14, 21, 23, 25\}$, $K_3 = \{31, 35, 30\}$

Q. Explain the concept of k-medoids.

→ **K-medoids :-**

- k-medoids problem is a clustering problem similar to k-means.
- k-medoids algorithms are partitioned i.e breaking the dataset up into group & attempt to minimize the distance between the points in a cluster and a point which is center of that cluster.
- k-medoids chooses actual data points as center.
- It is also called Partitioning Around Medoid.)

$$E = |P_i - C_i|$$

methods, medoids

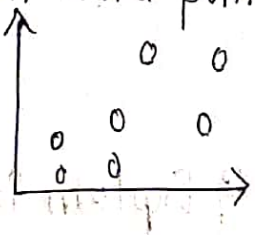
Explain Hierarchical Clustering Algorithm.

Hierarchical Clustering :-

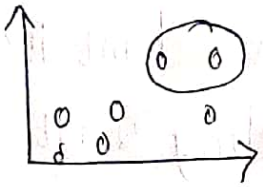
- It is unsupervised machine learning algorithm.
- Hierarchical clustering is also called as Hierarchical Cluster Analysis (HCA).
- Group of similar objects into groups is called cluster.
- Two types of HC:
  1. Divisive: (Top-down) Also called as DIANA.
  2. Agglomerative (Bottom-up): Also called as AGNES.
- We represents the hierarchy of cluster in the form of tree called as dendogram.

Steps :

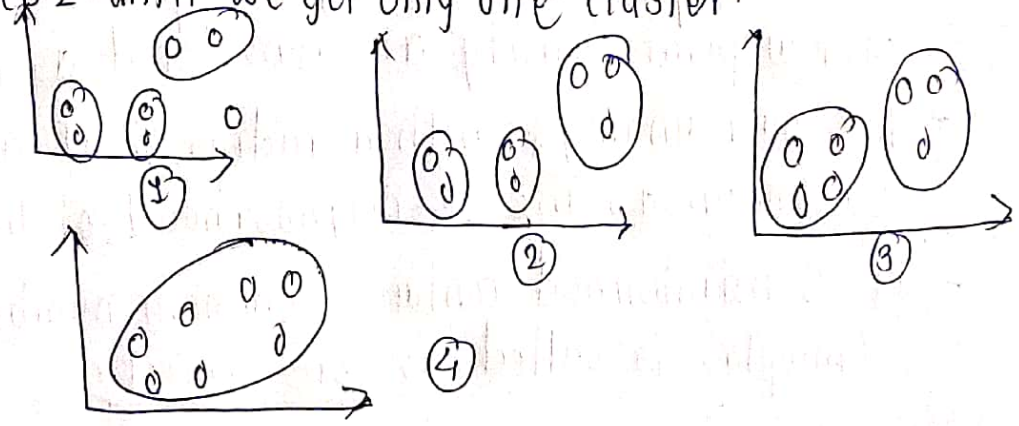
1. Create each data point as a single cluster.



2. Merge two nearest clusters.



3. Repeat step 2 until we get only one cluster.



4. Once all cluster are combined into one big cluster. then develop the dendogram to divide the cluster as per problem.

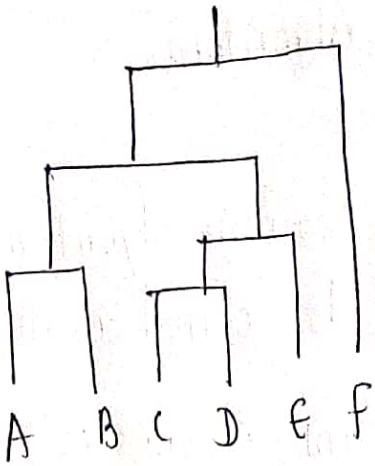


fig Dendogram

Advantages :

- Easy to Implement.
- Clear path for advancement.

Disadvantages :

- Poor flexibility
- communication banier.

Q. What is Density-Based Clustering & Explain its working ?

→ Density-Based Clustering :

- It is most popular unsupervised ML methodologies.
- It is used for model building & ML algorithms.
- The data points in the region separated by two clusters of low point density are considered as noise.
- The surroundings with a radius ϵ of a given object are known as the ϵ-neighborhood of the object.
- If ϵ-neighborhood contains minimum number of ~~points~~ objects (minpts) is called as core object.

Working :-

1. Let set of object denoted by D, object i is directly reachable from the object J only if it is located within the ϵ-neighborhood of j and j is a core object.

2. An object i is density reachable from the object J with respect o & minpts in a given set of objects.

$i_1 \ldots\ldots i_n$

$i_j = J$

$p_n = i$

such that i_{j+1} is directly reachable from i_j → ϵ & minpts.

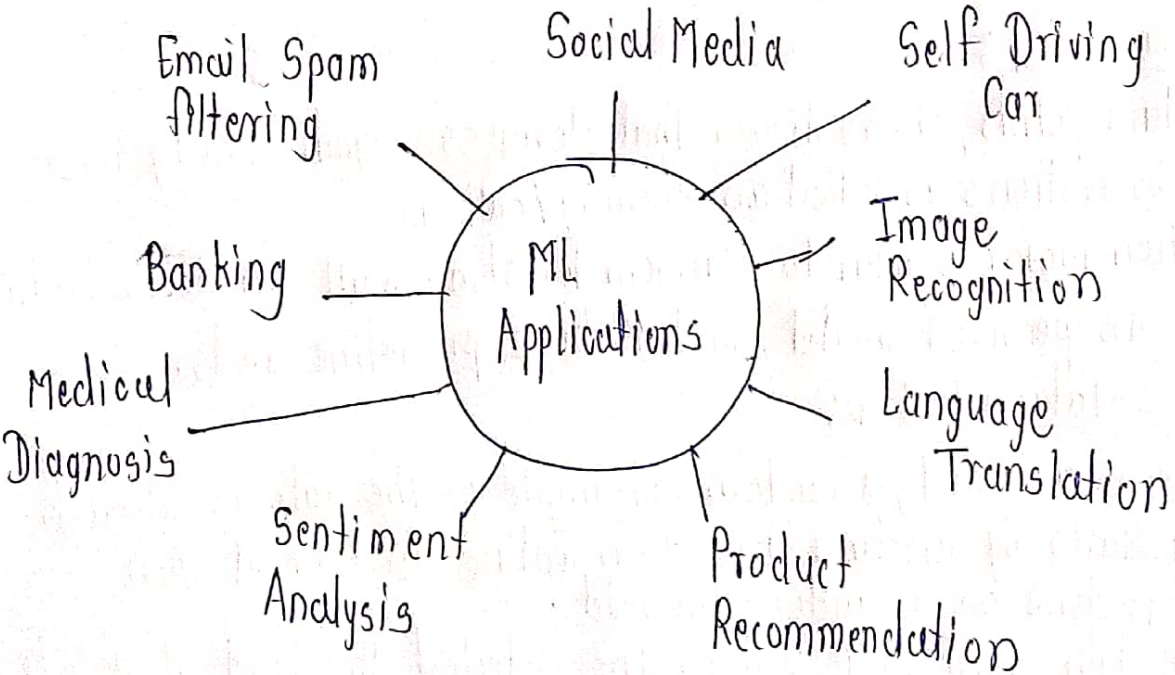
3. O is belonging to D so i & j are density reachable from o with respect to ϵ and minpts.

Features of Density Based Clustering:

- It is scan method.
- It required density parameters as a termination condn,
- It is used to manage noise in data clusters.
- DBC is used to identify clusters of arbitary size.

Q. Write short note on ML applications.

→



Q. What is outlier analysis.

→
- Outlier analysis is a fundamental issue in data mining.
- It is used to detect & remove anomalous object from data mining.
- The approach to detect outlier includes three methods.
  - (i) Clustering:
    - k-means used
    - Partitions the data set into given number of clusters.
  - (ii) Pruning:
    - It is based of on distance measure.
  - (iii) Computing Outlier score:
    - for unprune points this method used.
    - It is based on LDOF (local distance based outlier factor)
    - LDOF tells how much a point is ~~deriving~~ deviating from its neighbors.

Q. What are Isolation factor.

→ Isolation factor :-
- Any data points / observations that deviates significantally from other observations is called an Anomoly / outlier.
- Isolation factor similar to Random forest are built on decision tree.
- It is unsupervised model, so don't have pre-define labels.

Working of Isolation factor:

1. when given a dataset, a random subsample of the data is selected.
2. Branching starts by selecting a random feature first. And then branching is done on a random threshold.
3. Value of data point is less than the selected threshold, it goes to the left branch otherwise to the right. And thus a node is split into left & right branches.
4. This process from step 2 is continued recursively till each data point is completely isolated.

Explain local Outlier factor. Mention its advantages & disadvantages.

→ **Local Outlier factor :-**

- Local outlier factor (LOF) is an algorithm used for unsupervised outlier detection.
- It produce an anomaly score that represents data points which are outlier in the dataset.
- It does this by measuring the local density deviation.

Advantages :-

- Sometime it is difficult to find outliers. because of lack of training data. But for LOF is not required prior example.
- LOF can be applied in many other fields to solve problems. like video streams, geographic data, etc.
- LOF work well for anomoly detection algorithm for multimodel distribution.

Disadvantages :-

- LOF score not always same it might very for different datasets.
- In higher dimensions, accuracy of LOF algorithm get effected.

Q. Explain different evaluation metrics & score.

→ 1 **Confusion matrix :-**

PV \ AV	P	N
P	TP	FP
N	FN	TN

- Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
- Precision = (Positive Predictive value) $\frac{TP}{TP + FP}$

- Recall or sensitivity $= \frac{TP}{TP+FN}$

2. F1 Score :

F1 score is harmonic mean of precision & recall values.

$$F1 = 2\left[\frac{(Precision)\cdot(recall)}{(Precision)+(Recall)}\right]$$

3. RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$$

$$RMSE = \sqrt{MSE}$$

4. R-squared:

$$R^2 = 1 - \frac{MSE(model)}{MSE(base\ line)}$$

Q. Explain Elbow Method.

→ Elbow method :-

- In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a dataset.

Elbow Method Calculation :

1. Compute clustering algorithm for different values of k.
2. For each k calculate the total WSS.
3. Plot the curve of WSS according to the no. of clusters k.

- Elbow method does not always work well, if data is not very much clustered. Dataset ~~need not~~ did not have a clear Elbow.
- we can have a fairly smooth curve, it is unclear what is the ~~best value of k~~ choose best value of k.

Difference between Intrinsic & extrinsic methods

Inhinsic Motivation	Extrinsic Motivation
1. Known as internal motivation	1. known as external motivation
2. keep you productive	2. Just play with your mind.
3. Useful for whole life	3. can only be used for temporary purpose.
4. makes your decision making strong	4. makes your instant decision may be bad.
5. Hard to find	5. Easy to get
6. Example: your family Happiness	6. Example: Youtube channel likes.

Unit : 6

Unit VI: Introduction to Neural Networks.

Q. Explain Artificial Neural Networks?

→ **Artificial Neural Networks:-**

- ANN is a massively distributed processor made up of simple processing units which store experiential knowledge & make it available for use.

- Important features of ANN:-
  1. Knowledge is required from its environment through a learning process.
  2. Interneuron connections store the acquired knowledge.
  3. Artificial Neural Networks learns by examples.

- Advantages:-
  1. Parallel processing capability.
  2. Storing data on entire network.
  3. Capability to work with incomplete knowledge.
  4. Having memory distribution.
  5. Having a fault tolerance.

Q. Explain Single layer neural networks.

→ **Single layer Neural Networks:**

- Single layer neural network contains input & output layer.
- The input layer receives the input signals and the output layer generates the output signals accordingly.

Let us consider a Single Layer Network

Now, the separating line is given by,

$$b + x_1 w_1 + x_2 w_2 = 0$$

if $w_2 \neq 0$ then:

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2}$$

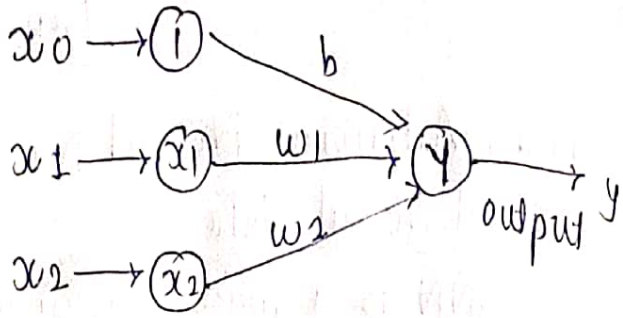


fig. SLNN

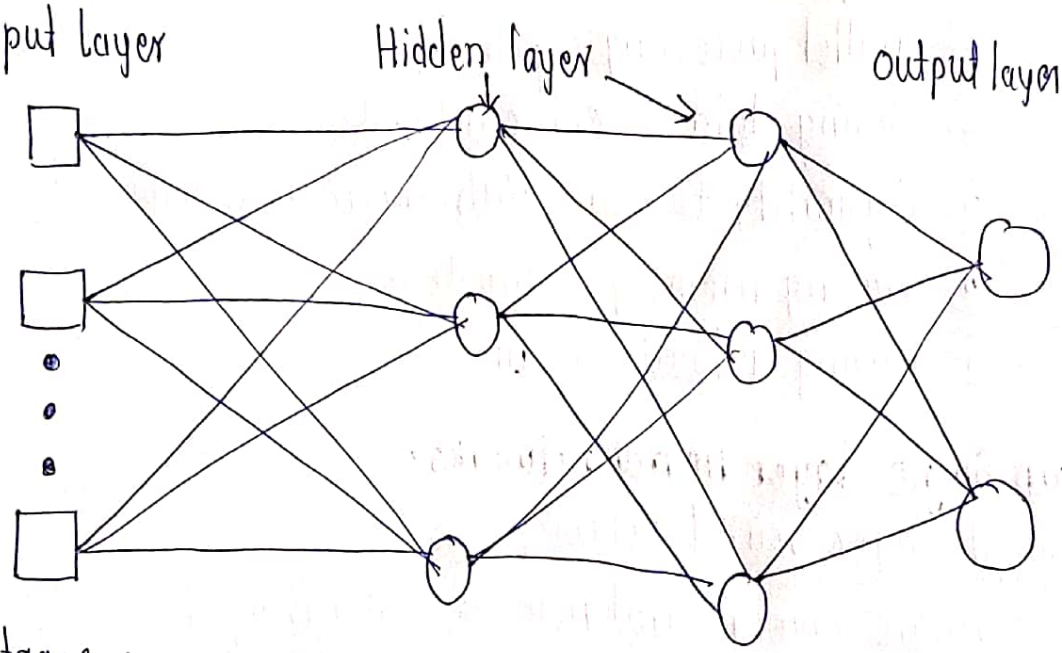
Thus the requirement for the positive response is:

$$b + x_1 w_1 + x_2 w_2 > 0.$$

Q. Explain multilayer perceptron.

→ Multilayer Perceptron:

- This network is made up of multiple layers.
- The architecture of this class consists of one or more intermediate layers called as hidden layers, besides having input & output layer.



Advantages:

1. Can be applied to non-linear complex problem.
2. Works well with large Input data.
3. Provides quick predictions after training.

Disadvantages:

1. Computations are difficult & time consuming.
2. Functioning of model depends on training data.

Q. Explain Back Propogation Network (BPN) / Learning.

→ BPN:

- Backpropogation is a widely used algorithm for ~~trainig~~ training feedforward neural network.
- The network connected to back-propogation learning algorithm are also called as back-propogation network (BPN).

Working of BPN:

① It compares generated output to the desired output & generate error report if the result does not match the generated output vector.

② Then it adjust the weights according to bug report then get desired output.

Algorithm:

Steps:

1. Inputs X, arrive through the perconnected path.
2. The input is modeled using true weights, weights chosen randomly.
3. Calculate the output of each neuron.
4. Calculate error in output.

Backpropogation Error = Actual Output − Desired Output.

5. From the output layer, go back to the hidden layer to adjust the weights to reduce the error.
6. Repeat the process until desired output is achieved.

Types of Backpropogation:

1. Static Propogation: Designed to map static input for static output.
2. Recurrent Backpropogation: It is also called as recursive backpropogation, used for fixed point learning.

Advantages :-

- It is simple, easy, fast to program.
- Flexible and efficient.

Disadvantage :

- Performance depends on input data.
- Spend too much time to training.

Q. Explain function link Artificial Neural Network (FLANN)?

→ ## FLANN :-

- Functional link artificial neural network (FLANN) is a single layer ANN.
- It has low computational complexity, which is used for different field of applications, such as system identification, pattern recognition, etc.
- It is used to model the reln betn input & output variables.

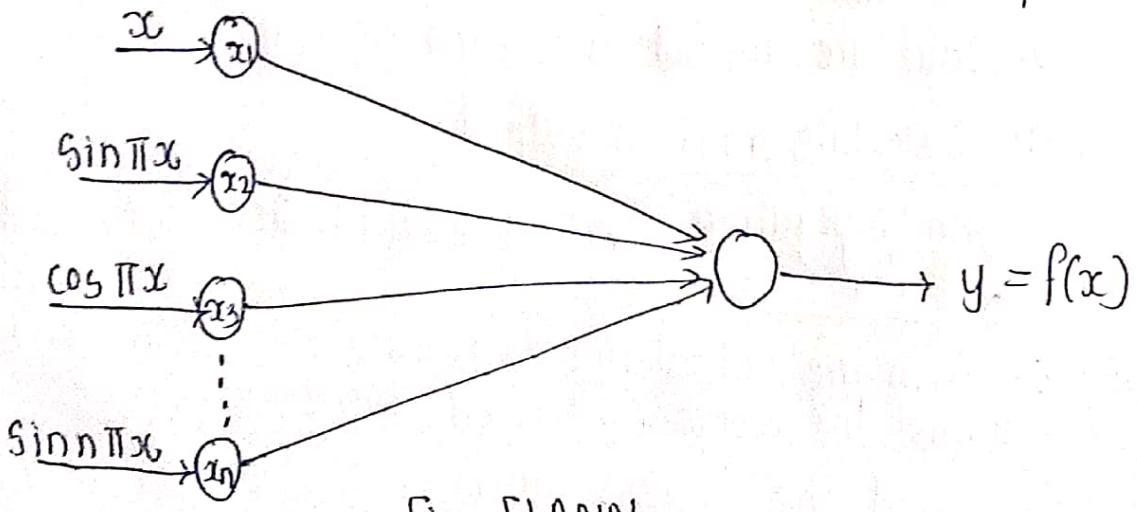
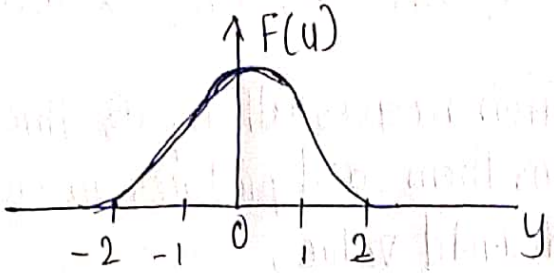


Fig. FLANN.

Q. Explain Radial Basis Function (RBF) Network ?

→ - The Radial Basis function is a functional approximation neural network.

- It was developed by Powell.
- The network uses sigmoidal & Gaussian kernel function.
- The response of Gaussian function is positive for all values of y & response decreases as $|y| \to 0$.
- Gaussian function $F(y) = e^{-y^2}$



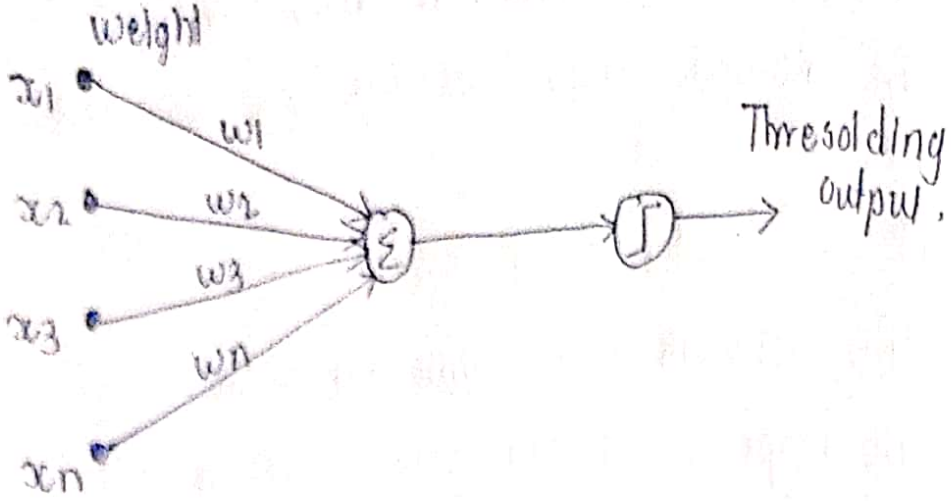
Advantages :

- Easy design.
- Good Generalization.
- Faster Training
- Only one hidden layer.

Q. Explain common activation functions used in neural network.

→ Activation function :-

- The activation function decides whether a neuron should be activated or not calculating the weighted sum & further adding bias to it.
- The purpose of the activation function is to introduce non-linearity into the output of a neuron.



- Here, $x_1, x_2, x_3, \ldots, x_n$ → n-inputs.
  $w_1, w_2, w_3, \ldots, w_n$ → weights.

- Biological neuron receives all inputs through the dendrites, sums them and produces an output.
  if sum > threshold value,
  then, Input signal pass onto cell body through synapse.

- $I = w_1x_1 + w_2x_2 + \cdots + w_nx_n$

$$I = \sum_{i=1}^{n} w_i x_i$$

- output.

$$y = f(I)$$

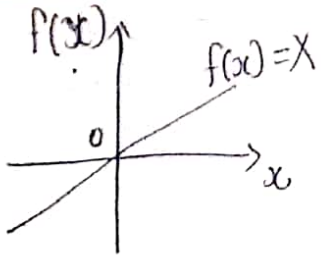
called Activation function, or Transfer fun^n or Squared fun^n.

Different Activation function used in neural network;

(1) Linear function:

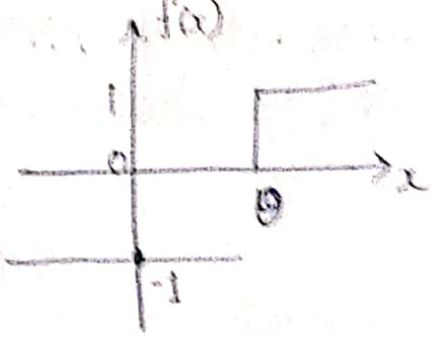
$$f(x) = X \quad \text{for all } x$$

~~Here Output = Inp~~ Input = output.



② Bipolar Step function

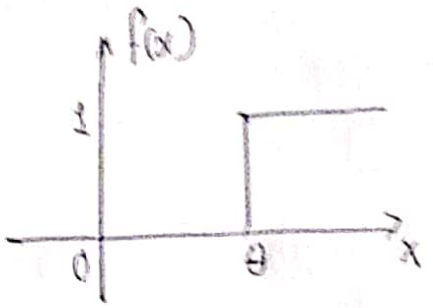
$$f(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ -1 & \text{if } x < \theta \end{cases}$$



$\theta \rightarrow$ threshold value

③ Binary step function.

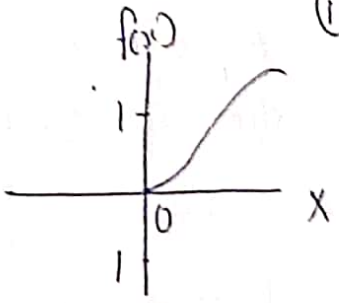
$$f(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta \end{cases}$$



④ Sigmoidal function

$$S(x) = \frac{1}{1+e^{-x}}$$

- Used for back-propogation nets.
- Two types:-

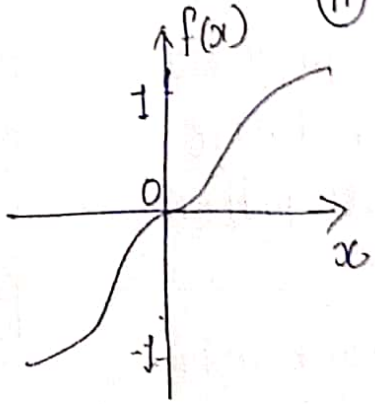


(i) Binary Sigmoidal function. (unipolar, logistic).

$$f(x) = \frac{1}{1+e^{-\lambda x}}$$

$\lambda \rightarrow$ Steepness parameter

Range $\rightarrow$ 0 to 1.



(ii) Bipolar sigmoid function.

$$f(x) = \frac{1-e^{-\lambda x}}{1+e^{-\lambda x}}$$

Range $\rightarrow$ -1 & +1

9. Difference between ANN, RNN (Recurrent Neural Network) & CNN (convulation neural Network).

→

	ANN	RNN	CNN
1. Type of Data	Tabular data, Text data.	Sequence data	Image data.
2. Parameter sharing	No	Yes	Yes
3. Fixed length input	Yes	No	Yes
4. Recurrent connections	No	Yes	No
5. Vanished & exploring gradient	Yes	Yes	Yes.
6. Spatial reln	No	No	Yes
7. Performance	less powerful than CNN, RNN	less feature than CNN	more powerful than ANN, RNN
8. Application	Facial recognition	Text to Speech	Facial recognition.
9. Advantage	Having Fault tolerance, ability to work with incomplete knowledge.	~~Remberes~~ remembers time series prediction	High accuracy in image recognition
10 Disadvantage	Hardware Dependency.	Gradient vanished. exploding gradient	Large training data needed.

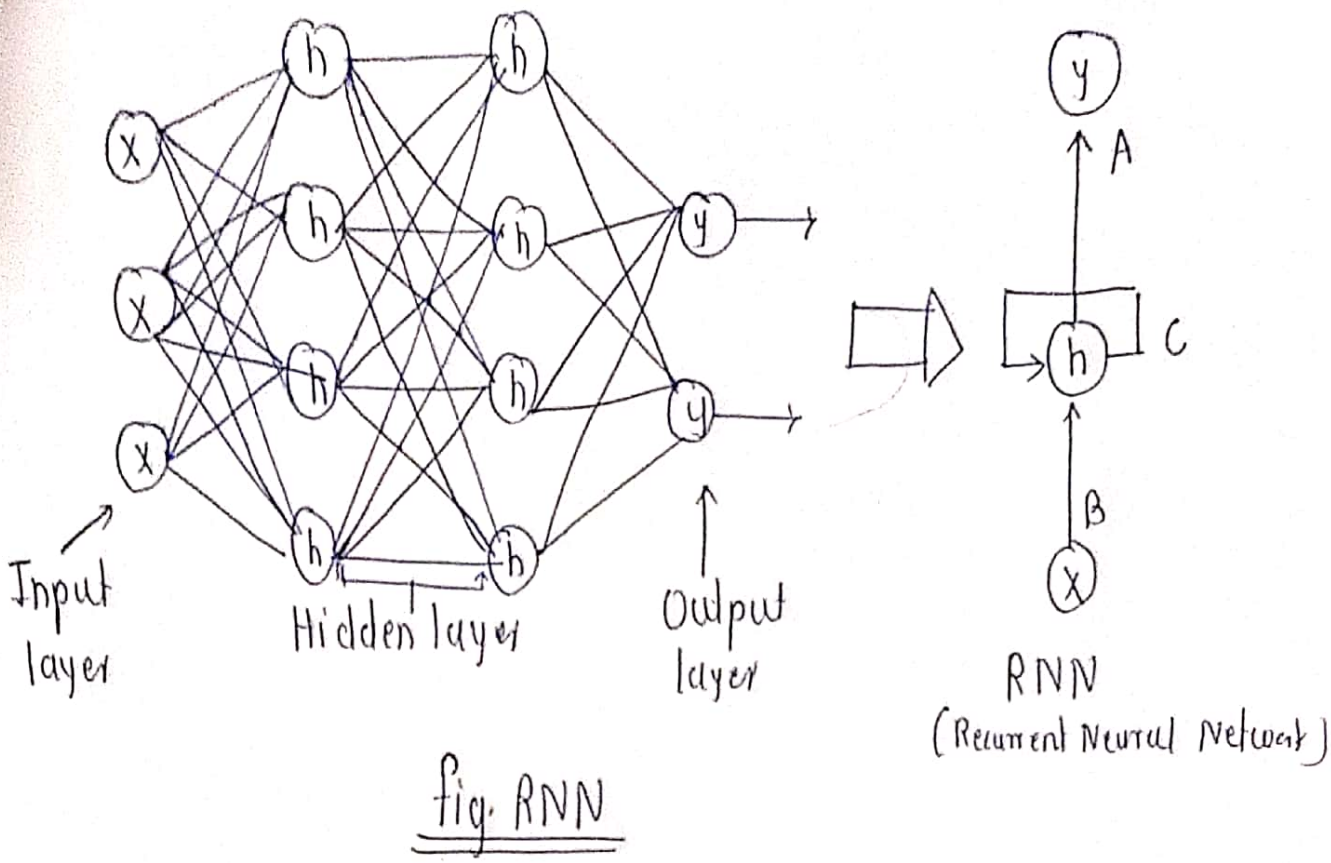


fig. RNN

Thank you

~Prof. Mayur Ghorpade
~Prof. Suraj Sonawane