Intro to ML

> Machine Learning - A Computer Science Technology where a machine itself has a complex range of knowledge that allows it to take certain data input and use complex statistical analysis to create output value. "Machine learning" focuses on building a system that can access data and use it to learn for themselves.

> Why ML is important? -

ML and data mining, a component of ML are crucial tools to process data and gain insights from massive amount of data.

→ Increase In Data Generation -
→ Improved Decision Making -
→ Uncover patterns & trends in data -
→ Solve complex problems -

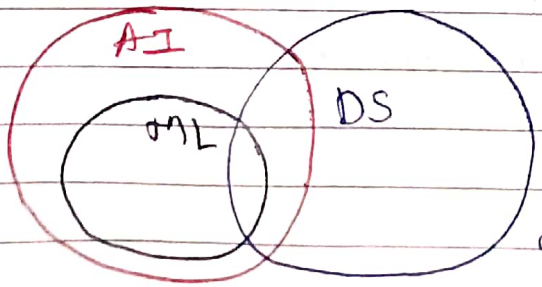
> Steps Involved.

1 Define the objective of problem Statement.
2 Data Gathering
3 Data preparation
4 Exploratory Data Analysis
5 Building Machine learning Model
6 Model Evaluation & optimization.

Comparison of ML with Traditional Programming.

	ML	TP
1)	Not a manual process Algorithm automatically predicts	Manual process. programmer Creates program without anyone has to manually Formulate
2)	Data → Result → Computation → Program	Data → Computation → Result. Result → Program
3)	As a Subset of AI ml is motivated by human learning behaviours.	We write down the exact steps required to solve problem
4)	ML algorithm takes an input and output and gives some logic which works on new Input	Traditional algo takes input and Some logic in the form of code and gives output.

> ML Data science and AI

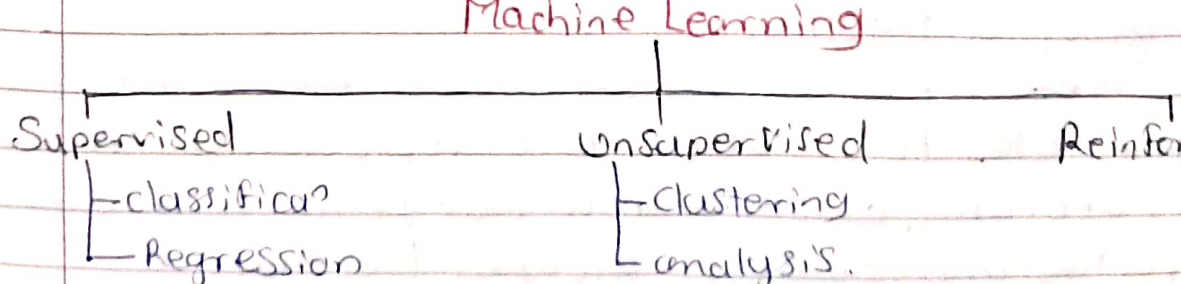


- ML is Subset of AI wherein Computer System learn from the environment.
- Data Science is processing, analysis and extraction of relevant assumptions from data.

- ML uses Statistical models. AI uses logic and decision Tree. Data Science deals with Structured data.

Machine Learning	AI	Data Sceence.
Focuses on learning from data and get experience to improve overtime.	Giving machines Cognitive & intellectual Abilities Similar to human	Focuses on extracting Information needles From data For decision making.
Uses Statistical model	uses logic & decision trees	Deals with Structured unStructured data.
A Form of Analytics in which program learns about data and find patterm.	Development of Comput- arized application that Simulate homan intelligest	Process of using advanced analytics to extract relevant info.
objective is to maximize accuracy	objective is to maximize the chance of success	objective is to extract actionable insights From data
ML can be done through Supervised, Unsupervised or reinforcement learning	Collection of intelligence concepts, including Predicates logic percep^n	Uses Statistics maths, big data, data worangling
Recommenda^n system Such as Spotify	Chatbots and voice assistant	Fraud detection and Healthcare

> Types of Learning



M T W T F S S
Page No.:
Date:
YOUVA

> Supervised Learning -

> It's a machine learning ~~pra~~ paradigm in which algo learns from labled training data to make predictions or decisions. The goal is to learn mapping from input data to the correct output lables so that the algorithm can make accurate predictions on neew unseen data.

> The algorithm uses the iterative optimization of an objective Function to predict the output that will be associated with neew inputs.

> Supervised learning algorithm uses classification and regression techniques to devlop predictive models. Regression technique predicts continuos responces ex. Changes in temperature.

> Supervised learning are trained using labelled dataset where the model learns about each type of data.

> Give examples - Adv & disadv

> UnSupervised Learning -

> It refers to learning from unlabled data. It is based on more Similiarities and differences than on anything else. In this learning Similar items are clustered together in a particular class.

> Many Unsupervised algorithm Create Similar hierarchical arrangements based on similarity based mapping.

> The task of heirarchical clustering is to arrange a set of objects into heirarchy such that similar objects are grouped together.

> A learner is fed with a set of scattered points, and it generates two clusters with representive centroid after learning. clusters show that points with Similar properties and closeness grouped together.

Types -
1) clustering
2) Association - Kmeans, KNN, Neural Network
Apriori Algo

	Supervised	Unsupervised
1)	Trained on labled data	trained on unlabled data
2)	model takes feedback to check if it's predicting correct or not	Does not take any feedback
3)	Predicts the output	Finds hidden pattern in data
4)	Data is provided to the model along with output	only Input data is provided.
5)	Goal is to train model that can predict when new data is inputed	Goal is to find pattern and useful insight from unknown dataset
6)	Needs Supervision to train	Does not need any Supervision
7)	Categorized into Classificam & regression	categorize into clustering & association problem.
8)	model produces accurate output	less accurate compared to Supervised learning
9)	linear regression, Logistic, SVM	clustering, KNN, Apriori Algorithm.

Page No.:
Date:
YOUVA

Distance models - Distance metrics aka similarity or dissimilarity metrics are mathematical functions used to quantify the similarity or dissimilarity between two points in a feature space. This metrics provide a way to measure the distance between data points which is fundamental concept in various ml algo.

→ Euclidian Distance - It's one of the well known distance metrics and is used in various applica? Calculates the straight line distance between 2 points in cartesian co-ordinate sys.

Formula $d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$

Euclidian distance is used in KNN and hierarchical clustering.

→ Manhattan Distance - Measures the distance between two points by summing the absolute differences between their coordinates:

Formula $d(p,q) = \sum_{i=1}^{n} |p_i - q_i|$

Manhattan used when movement between points can only occur along grid lines.

→ Cosine Similarity - Cosine Similarity measures the cosine of the angle between two vectors, indicating their similarity regardless of vector lengths. It is often used for comparing text documents, where each dimension corresponds to tf-IDF value.

Formula Cosine_similarity $(p,q) = \frac{p \cdot q}{|p|\,|q|}$

→ Hamming Distance - It is used to measure the difference between 2 strings of equal length often binary strings. It calculates the number of position at which the corresponding element are different. It is used in error detection and correction codes in digital communica?

If 10001001 and 10110001 are 2 codewords, then the corresponding bits differ in these two codewords is 3 bits. The number of bit positions in which two codewords differ is called hamming distance.

Equa^n $2^H \geq M + H + 1$ M = No. of bits in msg
H = Hamming bits.

⚠ Reinforcement Learning – RL is a subset of ml that deals with how agents should take actions in an environment to maximize a cumultative reward. RL involves learning through interaction with an environment without explicit supervision.

Key Components of RL –

1) Agent - Agent interacts with its environment. The agent takes actions, observes the enviroment, and learns to make better decisions over time.
2) Environment - External System with which the agent interacts. Representing the current situa^n and it can change based on agent's actions.
3) States - Representation of the current configura^n or situa^n of the environment at a given time. State gives essential info for the agent to make decisions.
4) Actions - Set of all possible moves or decisions that an agent can make in the environment.

Reinforcement learning widely used in various applicatio robotics, game playing, autonomous vehicles, recommenda^n systems, NLP and more.

Decision Tree - It's a Supervised ml algo that is used for both Classificaⁿ and regression tasks. It is a versatile and interpretable model that is particularly useful when you need to make decision based on a set of conditions or features.

key characteristics & concepts -

1) Tree Structure - Tree like structure composed of nodes. The tree begins with a root node, which represents the entire dataset and branches out into child nodes.
2) Nodes - Root Node - Representing entire dataset.
   Internal Nodes - Repre. Feature conditions and lead to child nodes.
   Leaf Nodes - Terminal nodes that make predictions.
3) Edges - Connectin between nodes represent the outcomes of feature conditions.
4) Feature Selection - At each internal node, the algo selects a feature and a threshold value that best splits the data into a more homogeneous subsets.
5) Decision Rules - Each path from root node to leaf node represents decision rule. For classificaⁿ tasks, these rules lead to class lables they lead to predicted values.
6) Impurity Measures -
   → Gini impurity - measures probability of misclassifying a randomly chosen element.
   → Entropy -
   → Mean Squared Error (MSE) - measures the variance of the target variable.

	Parametric method	Non Parametric method.
1)	Use a fixed number of parameters to build the model	Uses Flexible number of parameters to build model.
2)	Parametric Analysis is for testing group means.	For testing medians.
3)	applicable only for variables	applicable for both variable & attribute.
4)	Always considers Strong assumptions about data.	Generally considers fewer assumptions and data.
5)	Require lesser data than Non parametric	Requires much more data than parametric method
6)	Handles intervals data or relation data	handles original data.
7)	Follow normal distribⁿ	There's no assumed distribⁿ
8)	Output generated can be easily affected by outliers.	output cannot be easily affected by outliers.
9)	Parametric methods have more Statistical power than Non Parametric.	have less statistical power than parametric methods.
10)	Ex. Logistic Regression, Naive Bayes	Ex. KNN, Decision Tree, etc.

ML UII

— Feature Extraction —

★ Feature - Features are individual independent variables that act like input in the system. Feature is an attribute of a data set and used in ml process.

> Feature extracⁿ is the process of Selecting, transforming or creating relevant features from raw data. It aims to reduce the dimensionality of data.

> Feature Engineering - Process of creating new features or modifying existing one to improve performance.

> Feature Selection is a process that chooses a subset of features from the original feature space is optimally reduced according to a certain criterian.

★ Subset Selection -

It is a technique in feature selecⁿ where you choose a subset of most relevant features for use in a ml model. The goal is to reduce dimensionality of dataset by retaining the most relevant features

> There are 2 Types of subset selecⁿ -

1) Forward selecⁿ - It starts with an empty set of features and iteratively adds one feature at a time, selecting the one that improves model performance

2) Backward selecⁿ - It starts with all variables and remove them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error significantly.

3) Sequential Forward Search - SFS is the simplest greedy search algo. It starts from empty set, sequentially add the feature x^+. SFS performs best when optimal subset is small.

> The main disadv of SFS is that it is unable to remove features that become obsolete after adding new feature

4) Sequential Backward selecⁿ - Starts with all variables and remove them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error significantly.
> SBS works best when the optimal feature subset is large, since SBS spends most of its time visiting large subsets.

Preprocessing of Data -

Data preprocessing is a data mining technique that involves transforming raw data into understandable format.
> Data which capture from various source is not pure It contains some noise. known as dirty data.
> Incomplete, noisy and incosistent data are commonplace properties of large real world database.
> Steps during Pre-Processing -
1) Data cleaning - It is cleaned through processes such as filling in missing values, smoothing the noisy data or resolving inconsistencies in the data.
2) Data integraⁿ - Data with different representaⁿ are put together and conflicts within the data are resolved.
3) Data transformaⁿ - normalizaⁿ, Scaling.
4) Data Reducⁿ - To present duced representaⁿ of dataset.

> Normalizaⁿ & Scaling -

> Normalizaⁿ is a data preparation technique frequently used in ml. The process of transforming the columns in a dataset to the same scale is reffered to as normalizaⁿ.

> Normalizaⁿ makes features more consistent with each other, which allows the model outputs more accurately
> Normalizaⁿ refers to rescaling real valued numerical attributes into a 0 to 1 range. KNN, SVM uses normalizaⁿ
> Most widely used normalizaⁿ

1) Min-max Scaling -

$$X_{normalize} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2) Standardizaⁿ (Z-Score) -

$$X_{std} = \frac{X - \mu}{\sigma} \qquad \sigma = \sqrt{Var(X)}$$

$$\sigma^2 = \frac{(x - \mu)^2}{N}$$

Dimensionality Reduction -

Two components - Feature Selectⁿ & Extracⁿ.
→ Missing value, low variance, Decision Tree, Random Forest

Principle Component Analysis -

> It's Dimensionality reduction ~~techni~~ technique commonly used in machine learning. Primary goal is to reduce the number of features in a dataset while preserving as much of original info as possible.
> PCA is Unsupervised learning algo technique used to examine interrelations among a set of variables. PCA reduce the dimensionality of a dataset by finding new variables
> Variance Capture - PCA identifies a set of orthogonal axes called Principal Components, that captures the maximum variance in the data
> Orthogonal Transformaⁿ - PCA performs an orthogonal linear transformation of original features ensuring

that new components are uncorrelated with each other. PCA often uses data visualizaⁿ

> STEPS Involved -

Step 1 - Data Pre-Processing

Start with dataset that contain n data points and m features. Standardized or normalize the data to ensure that features with different scales do not dominate analysis.

Step 2 - Calculate Covariance matrix.

measures the strength of joint variability between 2 or more variables, indicating how much they change in relation to each other.

$$Cov(x_1, x_2) = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n-1}$$

Step 3 - Calculate eigen vectors and Eigen values.

> Performs eigen value decomposiⁿ on Covariance matrix Σ to calculate the eigen vectors & values.

> eigenvectors represent the directions in which the data varies the most, and eigenvalue indicates how much variance is explained along with principle component.

> Given square matrix A find eigen value λ and eigenvector X such that $AX = \lambda X$

Rearrange eqⁿ $Ax - \lambda X = 0$.

> $X(A - \lambda I) = 0$ I is identity matrix.

This eqⁿ will only hold if $(A - \lambda I)$ is singular matrix

Calculate determinant of $(A - \lambda I)$ Set it to equal to zero.

$$\det(A - \lambda I) = 0$$

> Solve λ to obtain eigenvalue of matrix A

for each eigenvalue λ_i, you can find corresponding X_i by solving eqⁿ $AX_i = \lambda_i X_i$.

Step 4 - Projec onto Principal Component -

> Project the original data onto selected k principle components to obtain new reduced dimension dataset.

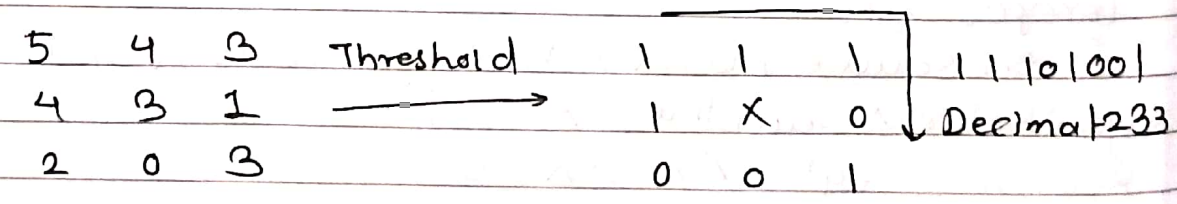
> Projecn is performed by multiplying the Standardized data by the k- Selected eigen vectors.

✱ Local Binary Pattern -

> It is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighbourhood of each pixel and consider the result as binary number.

> LBP operator is expressed by a sequence of binary values obtained by comparing central pixed with its neighbours in Circular manner.

> Compare the intensity val value of Centtral pixed with the intensity value of neighbor. For each neighbor if intensity value is greater than or equal to that of central pixel assign a binary '1' othercwise '0'



> The Decimal values obtained from LBP are used as texture descriptors for each Pixel and they can be further processed for various CV tasks.

> Texture classificaⁿ
Face recognitⁿ
object Detecⁿ
Image retrieval

Feature Selection Techniques -

Feature Selecⁿ

Filter method | Wrapper | Embedded

> Filter method -

Set all Features → Selecting the Best Subset → Learning Algo → Performance.

→ Filter method are generally used as preprocessing step. The Selecⁿ of features is independent of any ml algo. Features are selected on the basis of their scores in various statistical tests.

Some common techniques.

→ Infomaⁿ Gain, Chi Square Test, Anova

→ Correlaⁿ based Feature Sedecⁿ is a simpler filter algo that ranks features subsets according to a correlation based on heuristic evaluaⁿ Fcnⁿ.

> Wrapper method -

Set of Features.

↓

→ Generate Subset.

↓

Algo + performance.

↓

Performance.

In wrapper method, Selecⁿ of features is done by considering it as search problem, in which different combinaⁿ are made, evaluated and compared with other combinaⁿ

It trains algorithm by using subset of features iteratively.

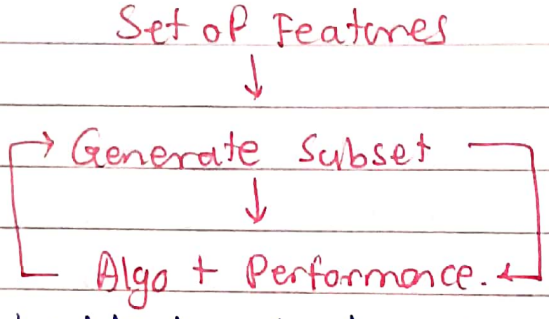
methods of wraper technique are.

→ forward Sedecⁿ

→ Backward Eliminaⁿ

Embedded method –

It is combined with the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter but more accurate than the filter method.



In embedded method feature selecⁿ algo is blended as part of learning algo, thus having its own built in feature selection method

Some techniques are -

→ Regularizaⁿ - This method adds penalty to different parameters of ml model to avoid overfitting of model.

→ Tree based methods - These methods such as Random Forest, Gradient Boosting provides us feature importance as a way to select feature as well.

UNIT

ML Supervised learning

Page No.
Date

Underfitting –

> Underfitting occurs when model is too simple to capture underlying patterns in training data.
> Underfitting destroys the accuracy of machine learning model, model or Algorithm does not fit the data well enough.
> It usually happens when model is trained on less data
> Techniques to Reduce underfitting.
1) Increase model complexity
2) Increasing no. of features perform feature engineering
3) Remove noise from the data.
4) Increase no. of epochs

Overfitting –

> A Statistical model is said to be overfitted, when we model is trained on too much data, it starts learning from the noise & inaccurate data in data set.
> Then the model does not categorized the dat correctly, because of too many details & noise. The cause of overfitting are non-parametric & non-linear methods because these models have freedom in building the model on dataset.
> Techniques –
1) Cross validation
2) Feature selectn
3) Regularizaⁿ – L1 or L2
4) Ensemble methods
5) ~~Simplyf~~ Simplify model.

Bias variance Tradeoff -

> Bias - Refers to the error introduced by approximating real world Problem too Simplistically. It is the inability of model because of that there is some difference or error occuring between model's predicted value and actual value.

> ways to reduce -

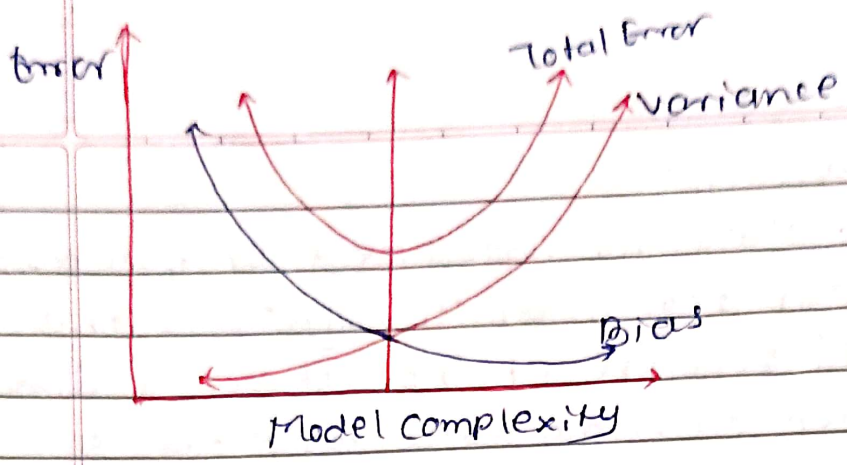
1) Use complex model
2) Increase number of features
3) Reduce Regularizon
4) Increase the size of training data.

> Variance - Variance is measure of spread in data from its mean position. Variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of ~~training~~ training data.

> ways to reduce

1) cross valida^n
2) Feature seleco^n
3) Regularization - L_1 & L_2
4) Ensemble methods.
5) Simplifying model

> If the algorithm is too simple then it may be on high bias and low variance conditions. If algorithm fit too complex ~~bias and~~ then it may be on high variance & low bias. There's something between both of these conditions known as Trade-off or Bias variance trade-off. The goal is to find the right level of complexity in a model minimize both bias & variance



Page No.
Date

Linear Regression - Linear regression is a type of Supervised learning algo which computes the linear relationship bet^n a dependent & more independent variables. The goal of algorithm is to find best linear equation that can predict the value of dependent value based on independent values.

eq^n $y = \beta_0 + \beta_1 X + \epsilon$

y - dependent variable. β_0 - y intercept.

X - independent β_1 - Slope

ϵ - error term.

$y = ax + b$.

$$a = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} \qquad b = \frac{1}{n}(\Sigma y - a * \Sigma x)$$

Regularization - It is a technique used in machine learning to a prevent overfitting and improve the generalization of model. It introduces a penalty term to the function that the model is trying to minimize during training. The magnitude of the coefficients is reduced by using different types of regulariza^n techniques like

1) Lasso Regression / L_1 regulariza^n
2) Ridge Regression. / L_2 regulariza^n

> Ridge Regression - It's a linear regression technique that adds penalty term of sum of squared values of model's coefficients. Goal is to prevent overfitting.

eqn Ridge R = Loss + $\lambda \sum_{i=1}^{n} (w_i)^2$

1) Loss Term represents the standard linear regression, loss often MSE
2) λ is regularizaⁿ Strength.
3) Ridge regression shrinks the coefficient towards zero but does not force them to exact zero
4) Larger value of λ results more shrinkage
5) It discourages model from relying too much on any single variable when multiple variables are highly correlated.

> LASSO Regression (Least Absolute shrinkage & Selecⁿ operator

It is a form of linear regression that incorporates use of regularization to avoid overfitting. Term lasso reflects it ability to shrink coefficeint towards zero and also acheives exact zero value

eqn Lasso R = Loss + $\lambda \sum_{i=1}^{n} w_i$

1) Adds the sum of ~~Squared~~ absolute values of model's coefficients as penalty term
2) Encourage Sparcity in model ie it tends to drive some coefficient to exactly zero
3) Useful when there's belief that many features are irrelevant or redundant.

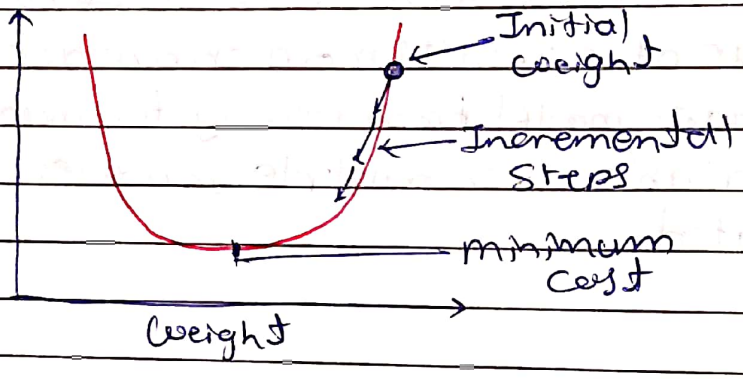
Date

Gradient Descent Algo -

Gradient Descent is an optimizn algorithm used to minimize a fnc^n iteratively by adjusting its parameters. It helps in finding local minimum of a function.

> If we move towards negative gradient or away from the gradient of the Function at current point, it will give local minimum.

> Move towards positive Gradient we will get local max.



> The main goal is to minimize cost Fnc^n using iter^n
To acheive this goal it perform 2 consecutive steps.

1) Calculates 1st order derivative of the Function to compute the gradient or slope of that Fnc^n

2) Move away from the direct^n of gradient which means slope increased from current point by alpha time alpha is the learning rate

3) Gradient Descent requires function to follow certain conditions like 1) function should be differentiable & 2) Function should be convex.

> Steps involved —

1) Initialization
2) Compute Gradient
3) Update parameters
4) Repeat
5) Convergence

> Types of GD -

(BGD) 1) Batch GD - It is used to find the error for each point in training set and update the model after evaluating all training examples. This is known as training epoch.

(SGD) 2) Stochastic GD - SGD is a type of GD that runs one training example per iteration. It processes training epoch for each example within dataset and updates each training example's parameters one at a time.

3) Mini Batch GD - Combinⁿ of both BGD & SGD divides the training dataset into small batch sizes then perform the updates on those batches separately.

★ Evaluation Metrics - Evaluation metrics are measures used to assess the performance of a machine learning model. These metrics provide insights into how well the model is performing on a given task, such as classificaⁿ, regression, clustering.

> classificaⁿ metrics

1) Accuracy = $\frac{\text{No. of correct predicⁿ}}{\text{Total No. of predicⁿ}}$

2) Precision = $\frac{TP}{TP + FP}$

3) Recall = $\frac{TP}{TP + FN}$

4) F1 Score = $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

> Confusion matrix - Performance measurement for machine learning classificaⁿ problems consists of True positive, True Negative, False positive, False negative

> AUC ROC - Areaunder the curve
ROC (Receiver operating characteristic) curve is the plot between False positive rate and True positive rate.

> Gini Coeff. - Used for classificaⁿ problems. Can be derived straight away from AUC ROC number. It is ratio between the ROC curve & diagonal line & the area above triangle.

$$Gini = 2 * AUC - 1$$

> Mean Absolute Error (MAE) -
Diff. between Actual & predicted values

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

(y: Actual, $\hat{y}$: Predicted.)

Aim is to get minimum MAE

> Root MEAN Squared Error (RMSE) -

$$RMSE = \sqrt{MSE}$$

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

(y_i: Actual, $\hat{y}_i$: Predicted.)

Lower RMSE expressed indicates better model performance
RMSE is Sensetive to outliers

> R Squared – R^2 is independent of context.
R squared is also known as Coeff. of Determination or Goodness of fit.

$$R^2 \text{ Squared} = 1 - \frac{SSr}{SSm} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (\bar{y}_i - \hat{y}_i)^2}$$

SSr - Squared sum error of regression line.
SSm - Squared sum error of mean line.

R-Squared measures the strength of relationship between ~~your~~ model and the dependant variable on a convenient ~~of 0 100% Scale~~ 0-1 scale.

> 1 when model perfectly fits the data.
> 0 when model does not predict any variability in the model. and it doesn't learn any relationship between the dependant & independant variable.

UNIT

Supervised learning : classificaⁿ

Page No.
Date

KNN Algorithm (K-Nearest Neighbor)

The KNN algorithm is a simple, intuitive & versatile supervised machine learning algo for both classificaⁿ & regression tasks. KNN is Non-parametric and instance based learning algo means it doesn't make assumpⁿ about underlying data distribaⁿ & memorizes training dataset.

> The algorithm stores entire training dataset in memory.

> To make a prediction for a new data point, the algorithm finds the k-nearest neighbors in the training dataset based on similarity metric

Q Consider following data to predict student pass or fail using KNN for value Phy = 6 mark & chem = 8 marks.

Phy	chem	Result
4	3	Fa
6	7	P
7	8	P
5	5	F
8	8	P

— Distance $= \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

$(x_2, y_2) = (8, 6)$

For (4,3) dist $= \sqrt{(8-4)^2 + (6-3)^2} = 5$

:

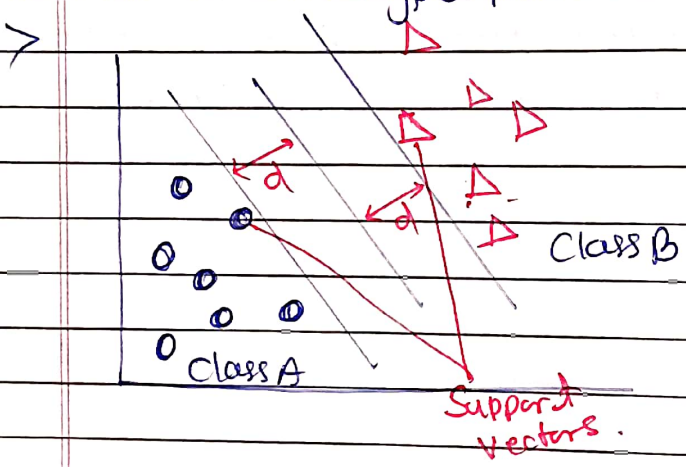
Phy	chem	Res	dist.
4	3	F	5
6	7	P	1
7	8	P	2.23
5	5	F	3.16
8	8	P	2.

3 Smallest dist are 1, 2.23 & 2 The result of these distances is pass so res of resul of new data point is pass

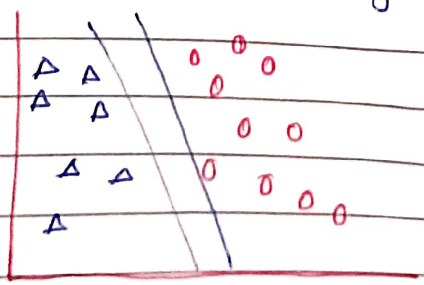
Page No.
Date

SVM (Support vector Machine) -

> SVM are set of Supervised learning algorithm which learns from the dataset
> SVM used for classification & regression tasks It performs both linear & non-linear classificon
> SVM finds a hyperplane to seperate the inputs into ~~sequence~~ seperate groups. There can be many hyperplanes that successfully seperates input vectors.
> Points closest to the hyperplane are known as Support vectors. These are the points lie closest to the decision boundary influence the position & orientaⁿ of hyperplane.



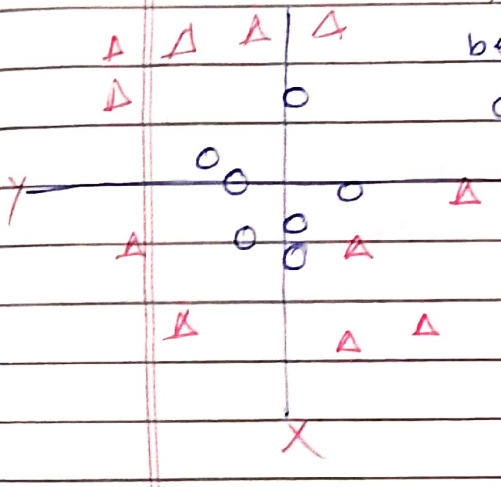
> Linear SVM -> Linear SVM aims to find a hyperplane in the input space that best seperates data points of different da classes.
> The decision boundary is straight line that maximizes the margin between classes.



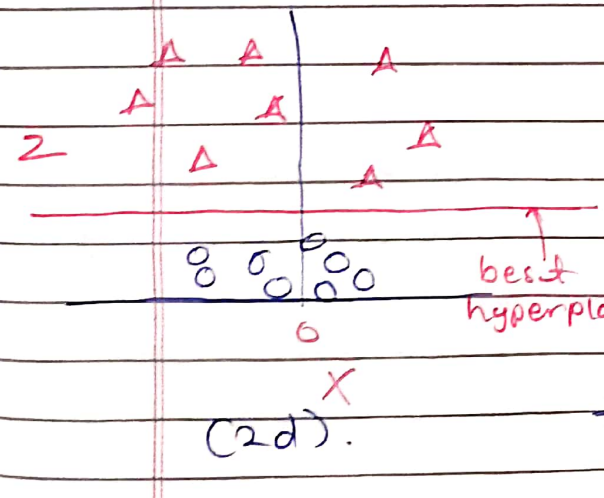
> Non Linear SVM - Non linear SVM used for non linearity see seperated data, which means if a dataset cannot be classified by using a straight line.

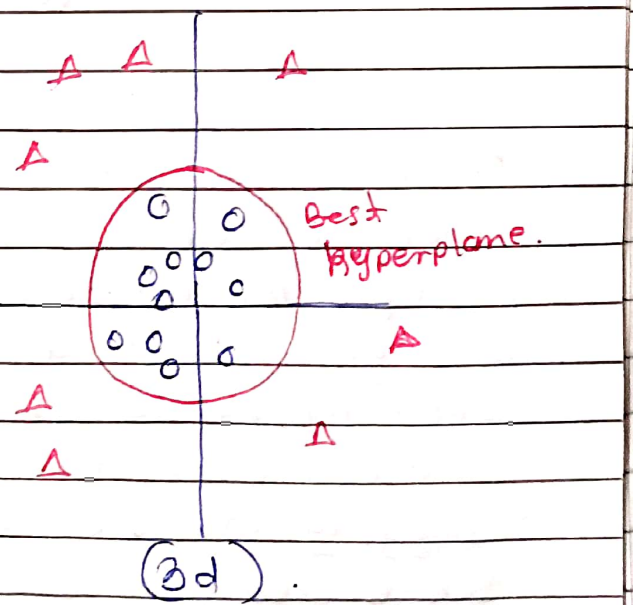
> Non linear uses kernel trick to implicitly map the input features into higher dimensional space where a linear decision boundary can be effective.

> It is suitable for dataset with complex & non-linear relationships between features.



> There are 2 classes that can't be seperated by straight line. But a ~~line~~ circular hyperplane can ~~Seperate~~ Seperate them with the help of x & y where $z = x^2 + y^2$ Now introducing 3rd dimension in graph.





> Seperating hyperplane – used in binary classificaⁿ which semaximally seperates 2 classes of data points in feature space Goal is to Search optimally Seperating hyperplane

$w^T x + b = 0$ w is weight.
x is feature vector.
b is bias

$w^T x + b > 0$ corresponds to one class
$w^T x + b < 0$ corresponds to another class.

> Margin – Distance between seperating hyperplanes and nearest datapoint from both classes. It represents safety buffer or width of corridor in which data points can reside without affecting Ga classificaⁿ result. Larger margin model is robust if can tolerate noisy data. Smaller margin may lead to more accuracy but may overfit

$$\text{Margin} = \frac{1}{|w|} |w^T x_i + b|$$

★ Ensemble Learning – Ensemble learning is a machine learning technique that involves combining multiple models to improve overall performance and generalizaⁿ. basic idea is to combine multiple weak models to create a strong & more accurate model.

> Bagging (Bootstrap Aggregating) –

> Bagging involves training multiple instances of the same learning algo on different subsets of training data.

> Each model in the ensemble is exposed to a different subset of the data introducing diversity among the models.

> Random Forest is a popular ensemble method that uses bagging with decision trees.
> Bagging can reduce overfitting, Increase stability, Improves accuracy & generalizaⁿ.

> Boosting –
> Boosting aims to sequentially train a series of weak models, each focusing on correcting the errors made by predecessor.
> Instances that are misclassified by one model are given higher weights in the training set of next model.
> Predictions from all weak models are combined with different weights, and the final predicⁿ is made by taking a weighted majority ~~weight~~ vote.
> Ex - AdaBoost, Gradient Boosting, XGBoost.
> It can acheive high accuracy, Effectively ~~hande~~ handles imbalanced datasets, Adapts more complex relationships in data.

* Random Forest - Random Forest is an ~~em~~ ensemble learning technique that belongs to family of bagging. It builds multiple decision trees during training and merges them together to obtain a more accurate and stable prediction.

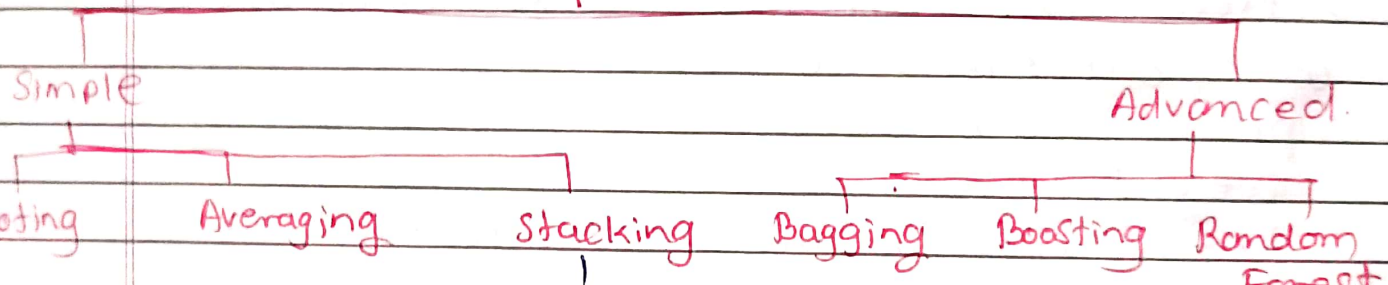
Steps -
1) Random Forest starts creating multiple random samples with replacement from original training dataset. Each sample is used to train a seperate decision tree.

2) At each node Random forest randomly selects a subset of features. To include ~~randomness~~ diversity.

3) Decision tree are grown for each bootstrap sample using randomly selected features. Each tree expands until a certain condition is met.

4) Vote for classificaⁿ task the predicⁿ from individual trees are combined through majority vote for regression task prediction ~~& av~~ are averaged.

* Ensemble Learning methods -



Stacking combines predictions from ~~multiple models~~ multiple models using meta model.
Meta model is trained to make final predictions based ~~on~~ on base model.

* Confusion Matrix - It is a table that is often used to evaluate the performance of a classificaⁿ algo. It provides a summary of the classificaⁿ results by showing the number of true positive, ~~fr~~ true -ve, false +ve, false -ve predicⁿ.

> True Positive (TP) - No. of instances correctly predicted as +ve
> True-ve - No. of instances correctly predicted as -ve
> False +ve - No. of instances incorrectly predicted as +ve (Type I error)
> False-ve - No. of instances incorrectly predicted as -ve (Type II error).
> The importance of a confusion matrix lies in its ability to provide a detailed breakdown of a classifier's performance various performance metrics can be calculated including.

1) Accuracy $\Rightarrow \frac{TP+TN}{TP+FP+FN+TN}$

2) Precision $\Rightarrow \frac{TP}{TP+FP}$

3) Recall $\Rightarrow \frac{TP}{TP+FN}$

4) F1 Score = ~~2xPr~~ $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

> These metrics help in assessing different aspects of a classifier's performance, such as ability to correctly classify positive instances, avoid false positives & overall accuracy.

> F1 Score - It is a metrics that combines precision and recall into a single value. It is particularly useful in situations where there is an uneven class distribu^n
> F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the worst possible performance.
> F1 score is harmonic mean of precision & recall, giving more weight to lower values. This makes it's a suitable metric when there's an imbalance bet^n the classes.

Unit V

Page No.
Date

Unsupervised Learning

* Clustering – clustering is a process of partitioning a set of data in a set of meaningful subclasses called as clusters.

* K-Means clustering – It is a popular machine learning algo used for partitioning a dataset into k distinct, non overlapping subgroups or clusters. The goal of kmeans is to group data points into clusters such that points within the same cluster are more similar to each other.

Steps

1) Initializaⁿ – choose number of clusters k (centroids)
2) Assignment - Assign each data point to the nearest cluster centroid. Usually done by using Eudlidean distance.
3) Update centroids - Recalculate the centroid of each cluster by taking mean of all data points assigned to that cluster.
4) Repeat step 2 & 3 until convergence. Convergence occurs when assignment of data points to clusters no longer changes significantly

> Algo takes the unlabeled dataset as input, divides the dataset into k-number of clusters and repeats the process until it doesn't find best clusters.

> Find value of K

* Elbow method - This method uses the concept of WCSS value (Within cluster sum of squares) which defines the total variations within cluster.

> It executes the kmeans clustering on a given dataset for different k value from 1-10
> For each value of K, calculates the WCSS value.

> Plots a curve between calculated WCSS values and the number of clusters K.
> The sharp point of bend or a point of the plot looks like an arm, then the point is considered as the best value of K

Sum Squared Error = $\sum_{i=1}^{n} \sum_{j=1}^{k} ||x_i - c_j||^2$

n - no. of data pt.
K - no. of clusters
x_i - ie i^{th} data pt.
c_j - j^{th} cluster.

K-MEDOIDS - It is a clustering algo similar to K-means but has key difference in how it defines cluster centres. While Kmeans uses the mean of the data points in a cluster as the cluster center, K medoids uses actual data point that is most centrally located in a cluster as the medoid. The medoid is the data point with minimum average dissimilarity to all other data points in cluster.

> Steps

1) Initializaⁿ - choose K initial data points ~~in dataset~~.
2) Assignment - Assign each data point to the nearest medoid by using any common dist. metric (Euclidean dist. or Manhattan dist).
3) (Swap) while the cost of configuraⁿ decreases.
   > for each medoid 'm' and each data pt 'o'
      a) consider the swap of m and o and compute cost change.
      b) If the cost change is ~~cur~~ current best, remember this m and o combinaⁿ
   2) Perform the best swap mbest & Obest if it decreases the cost function

Hierarchical clustering – It is a unsupervised learning clustering algo builds a hierarchy of clusters where each node in the tree represents a clusters. classified into 2 types
a Agglomerative (bottom up) & divisive (top-down)

> Agglomerative – It is known as bottom up approach or hierarchical agglomerative clustering. It starts clustering by treating the individual data points as a single cluster then it is merged continuously based on similarity until it forms one big cluster containing all objects.

Steps –

1) Create each data point as Single cluster.
2) Take two closest data points or clusters and merge them to form one cluster.
3) Again take 2 closest clusters and merge them together to form one cluster.
4) Repeat Step 3 until one cluster left.
5) Once all clusters are combined into one big cluster develop the dendgram & divide the clusters as per the problem.

> Divisive – This works just apposite of agglomerative clustering. It starts by considering all the data points into big Single cluster and later on Splitting them into smaller heterogeneous clusters continuously until all data points are in their own cluster. It follows a top down approach and is more efficient than Agglomerative clustering.

Steps -

1) Split into clusters using any flat-clustering method say K-means.
2) choose the best cluster among the clusters to split further, choose the one that has the largest Sum of Squared Error.
3) Repeat Steps 2 & 3 until a single cluster is formed.

> Measure For the distance betⁿ 2 clusters -

1) Single Linkage - Shortest distance betⁿ the closest points of clusters.
2) Complete Linkage - Farthest distance between the 2 points of two diff. clusters.
3) Average Linkage - Distance between each pair of datasets is added up and then divided by the total number of datasets to calculate avg dist. betⁿ 2 clusters.
4) Centroid linkage - It is the linkage method in which the dist. betⁿ centroid of the clusters is calculated.

> Dendogram - A dendogram is a tree-like diagram that illustrates the hierarchical structure of clusters in hierarchical clustering. It's a visual representaⁿ of relationships and similarities betⁿ data pt.

Key Features.

1) Vertical Lines - Represents data pt. or clusters.
2) Horizonatal Lines - merging or splitting of clusters.
3) Height of Fusion - heigh of each fusion or division represents the dissimilarity betⁿ merged or split clusters.

4) Nodes - Points where vertical & horizontal lines meet are nodes.
5) Leaves - The individual data pt. or small clusters at the bottom of dendogram.

> Steps -
1) Start by treating each ~~node~~ data pt. as Singleton cluster
2) Identify the two most Similar cluster and merged them into new cluster.
3) Repeat the process, identifying & merging the next most similar clusters until all data pts. belong to a single cluster

DBSCAN (Density Based Spatial Clustering of Applications with Noise)

> It groups densely grouped data points into a single cluster. It can identify cluster in large Spatial datasets by looking at the local density of data points. It is robust to outliers.
> It also does not require the number of clusters to be told beforehand unlike k-Means.
> DBSCAN requires only two parameters
I) epsilon - radius of the circle to be created around each data point to check the density
II) minpoints - min. no. of data points required inside that circle for that data point to be classified as a core point.
> DBSCAN creates a circle of epsilon radius around every data point and classifies them into core point, Border point & Noise

> If the no. of points is less than min points, then it is classified as Border point. and if there are no other data points around any data point within epsilon epsilon radius, then it is treated as Noise.

> Direct density Rechable - A point A is directly density reachable from another point "B" if A is in ϵ-neighborhood of "B" and "B" is core point.

> Density Reachable - A point "A" is density reachable from B if there are a set of core points leading from "B" to "A".

> Density Connected - A and B are density connected if there are a core point "C" such that both A and B are density reachable from "C"

Local Outlier Factor (LOF) -

> LOF is an algorithm that identifies the outliers present in the dataset. Outlier ~~distribu~~ detecn method can be distribun based, clustering based and density based. LOF allows to define outliers by doing density based scoring

> LOF algo is an unsupervised anomaly detecn method which computes the local density deviatn of given dt data pt. with respect to its neighbors.

> The LOF algo can be used for outlier detection & ~~move~~ novelty detecn. The diff. betn outlier & novelty detecn lies in training dataset. Outlier detecn indudes outliers in training dataset. The algorithm fits the area with high density data & ignores the outliers & anomalies.

* OPTICS (Ordering Points To Identify clustering Structure)
It is Density based clustering algo that extends the concepts of DBSCAN to identify clusters of varying shapes and sizes in a dataset.
> The core concept in OPTICS is the reachability distance, which measures how easily one data point can be reached from another.
> The minimum ~~requirement~~ reachability distance required for a point to be considered as core point
Core points have atleast specified no. of ~~neg~~ neighbors within specified radius.
> OPTICS constructs a reachability ~~m~~ plot. which is a sorted list of points and their corresponding reachability distances.
> Local maxima in reachability plot correspond to clusters in dataset. These maxima indicate points where the density changes suggesting presence of cluster

* DENCLUE (Density based Clustering) -
It is a Density based clustering is a density baked clustering algo focusing on modeling clusters as attractors in feature space. Aims to discover clusters with arbitrary shapes, handle varying cluster densities & adapt Noisy data.

1) It has a Solid mathematical founda^{n}
2) It is definetly good for datasets with large Noise
It allows compact mathematical descrip^{n} of arbitrarily shaped clusters in high dimensional datasets.

3) It uses grid cells but only keeps info about grid cells that do actually contain data points and manages these cells in a tree based access structure.
4) The influence Func^n is central to the gradient ascent process, where DENCLUE iteratively moves towards the maximum density points known as attractors or density peaks.
5) The clusters can be determined mathematically by identifying density attractors. Density attractors are local maxima of the overall density Func^n.

UNIT - ML

Intro to NN

Page No.
Date

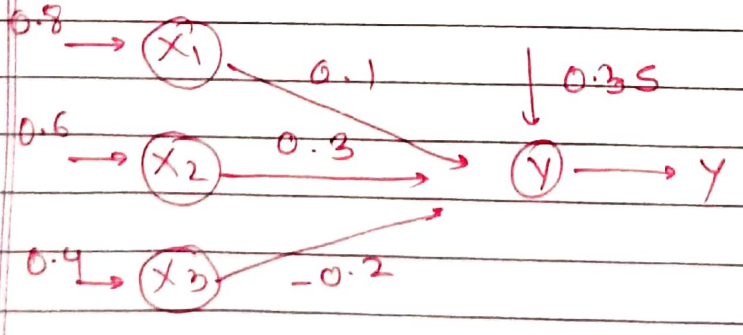
Artificial Neural Network –

An Ann is a computational model inspired by the structure and functioning of the brain. It consists of interconnected nodes also known as artificial neurons or perceptrons, organized into layers. ANNs are used for machine learning tasks, including pattern recogn^{n}, classifica^{n}, regression.

> characteristics

1) ANNs consist of layers of Nodes, including an input layer, one or more hidden layer and an output layer. The connections between nodes have associated weights that are adjusted during learning process.
2) Each node in an ANN typically applies an activation function to its input, determining the Node's output.
3) Feed Forward phase, input data is propagated through the network to produce an output.
Backpropagation phase calculates the weight and errors and adjust it to minimize errors.

Q Obtain output of the neuron Y for the network in figure using activation function as
> binary sigmoidal & > bipolar sigmoidal.



$x_1 = 0.8 \quad x_2 = 0.6 \quad x_3 = 0.4$

$w_1 = 0.1 \quad w_2 = 0.3 \quad w_3 = -0.2$

$$y_{in} = b + \sum_{i=1}^{3} x_i w_i$$

$$= 0.35 + 0.08 + 0.18 - 0.08 = 0.53$$

$y_{in} = 0.53$

1) Binary Sigmoidal activation Function

$$y = f(y_{in}) = \frac{1}{1+e^{-y_{in}}} = \frac{1}{1+e^{-0.53}} = 0.625$$

2) Bipolar Sigmoidal activation Function

$$y = f(y_{in}) = \frac{2}{1+e^{-y_{in}}} - 1 \Rightarrow \frac{2}{1+e^{-0.53}} - 1 = 0.259$$

Back Propagation Learning -

Backward Propagation of Errors is a supervised learning algorithm used to train ANNs. It is a key component of the learning process in neural network is employed to minimize the error between the predicted output and actual output.

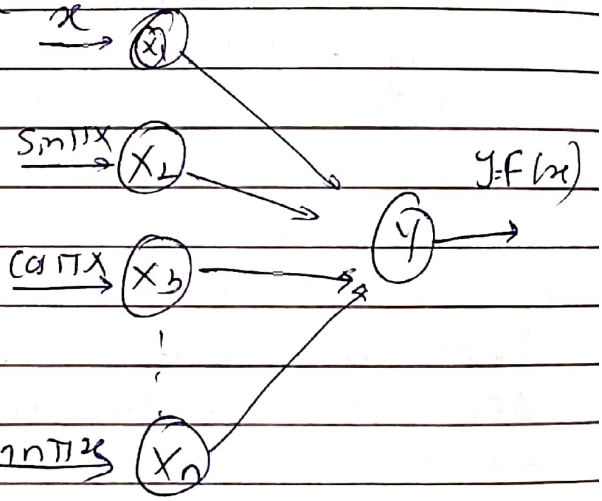
> Propagation Steps -

1) Initialize weights and biases of Neural Net randomly
2) Input data is propagated Forward through the network to compute the predicted output. Apply activation Func^n at each layer.
3) ~~Computing the error with respect to the weights and biases using the chain rule of calculus.~~
3) Computing error or difference bet^n predicted output & acctu actual target using loss Function.

4) Propagate the error backward through the network to update the weights and biases
5) Adjust weight and biases by using backpropaga using an optimiza algo like Gradient descent.
6) Repeat steps for multiple itera or epochs.

Functional LINK ANN –

> FLANN is a Single layer ANN with less computational complexity which is used in different fields of applica such as System identifica, pattern recognition
> FLANN architecture uses a single layer Feed Forward network.
> Using functionally expanded features FLANN overcomes the non-linearity nature of problems and encounters in Single layer network.
> FLANN eliminates the hidden layers. Due to its Single layer construction, the FLANN structure offers less computational complexity and Faster Convergence than MLP.



Activation Function -

1) Activaⁿ Fᵘⁿᶜⁿ also known as transfer function is used to map input nodes in certain fashion.
2) Id decides whether to activate the neuron or not and transferred to the next layer.
3) Activaⁿ Fᵘⁿᶜⁿ help in normalizing the output betⁿ 0 to 1 or -1 to 1. It helps in the process of backpropagaⁿ. During backpropagaⁿ loss Fᵘⁿᶜⁿ gets updated and activaⁿ Fᵘⁿᶜⁿ helps the gradient descent curves to acheive their local minima.
4) Activaⁿ Fᵘⁿᶜⁿ decides in any neural net to that given input or recieving info is relevant or not.
5) Sigmoid Activaⁿ Fᵘⁿᶜⁿ -

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad \text{Range } (0,1)$$

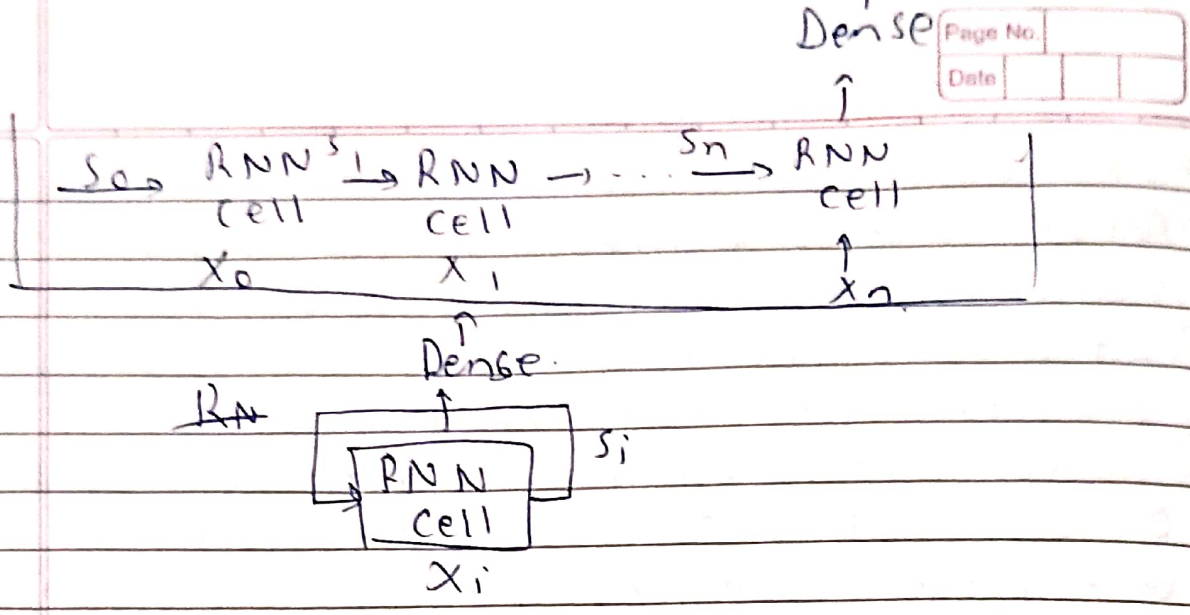
6) Tanh activaⁿ Fᵘⁿᶜⁿ -

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1} \qquad \text{Range } (-1,1)$$

7) Rectified Linear Unit (ReLU)

$$f(x) = \max(0,x) \qquad \text{Range } [0,\infty]$$

RNN - A RNN is a type of ANN designed for processing sequences of data. Unlike feed forward neural Net, RNNs have connections that form directed cycles, allowing them to maintain a hidden state that captures info about previous input. Ability to capture temporal dependencies makes RNNs well suited for tasks involving sequences such as NLP, time series predicⁿ. & Speech recognⁿ. RNNs have hidden state which acts as a memory, capturing info from previous time steps.



Formula for calculating current state:

$$h_t = F(h_{t-1}, X_t)$$

Example LSTM

CNN - ~~In CNN, the~~ It is a class of deep neural network specifically designed for processing structured grid data such as images CNNs are highly effective in image recogni^n.

Key components.

1) convolutional layer :- it is designed to perform convolution on input data. It is a mathematical opera^n that combines input data with a filter to extract local patterns or features.
   - > filters are ~~learnable~~ learnable matrices are convolved with input data it detects features like textures, edges.
   - > Convolution operation slides across input to capture local patterns & create feature maps.
   - > Activation Func^n.

2) Hidden layer - 1

1) Hidden layer refer to the layers between the convolu^{n} layer & output layer.

2) Hidden layers play a crucial role in learning hierarchical features from the ~~input~~ input data

3) Hidden layers also have learnable parameters including weights & biases which are updated during training to optimize network.

4) Activa^{n} Func^{n}. - To introduce non linearity