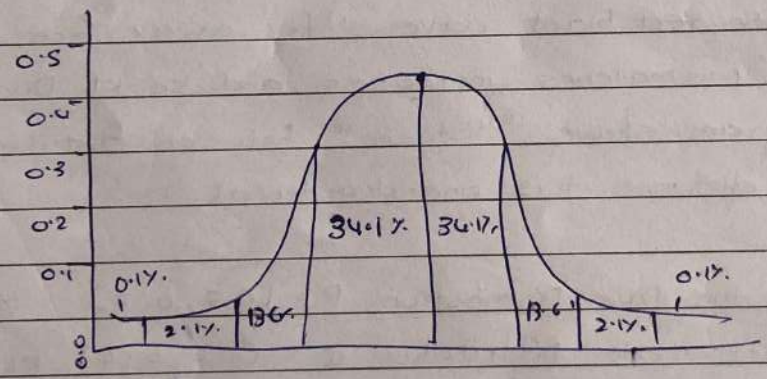
MACHINE LEARNING - U2

Q1. KL Divergence.

Ans KL Divergence stands for Kullback - Leibler Divergence. It is use to measure and compare two distribution with each other. It tells us how much information is lost when we approximate a true distribution "p" with another distribution "Q". If the two are the same, the KL Divergence is zero, the larger the difference the higher the divergence.

DIVERGENCE IN STATISTICS

The divergence between the two probability distribution quantifies how much the two differ from each other



KL-Divergence is an assymetric divergence which means that given probability distribution between P & Q will not be same as Q & P.

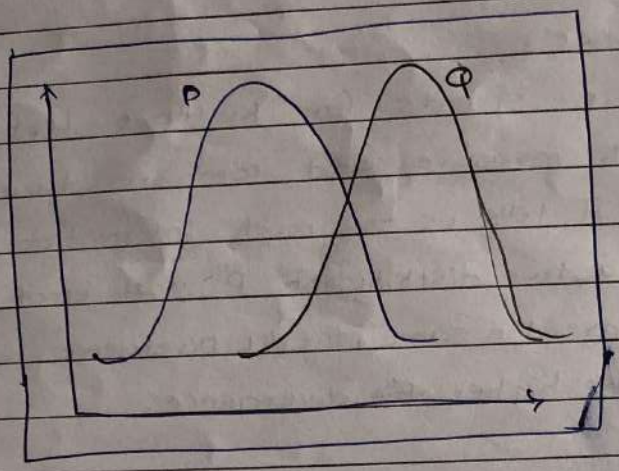
KL Divergence is defined as number of bits required to convert one distribution into another. The lower the bound value is zero and is achieved when the distribution under observation are identical It is often denoted with full notation . DKL(P || Q).

and the formula metric is $KL(P||Q) = \sum_x P(x) * \log\left(\frac{P(x)}{Q(x)}\right)$

Sundaram

FOR EDUCATIONAL USE

Ex:



Imagine 2-belled shaped curves (probability distributions) on the same graph, One curve is in blue the true distribution P and another in Black is the approximation Q. If the curves overlap perfectly that means the KL Divergence is zero. But if the ~~red~~ black curves shifts away from the blue one, the area of mismatches increases and so KL Divergence. Thus the the diagram shows "distance" between distribution but unlike regular distance it is one-directional.

Suppose the true Distribution $P = [0.8, 0.2)$ & the approximate Distribution $Q = (0.6, 0.4)$ KL Divergence is calculated as :

$$D_{KL}(P||Q) = 0.8 \log \frac{0.8}{0.6} + 0.2 \log \frac{0.2}{0.4}$$

This works out about 0.09 which is very small. That means our Q is quite close to P.

Sundaram
FOR EDUCATIONAL USE

Q. Decision Tree Algorithm.

Ans. Decision Tree are widely used in machine learning and can be applied to both classified and reggression tasks. These models works by splitting data into subsets based on features this process is known as decision making. Each leaf node provides a prediction and splits ~~the creation~~ to create a tree like structure. Decision tree are popular because they are easy to interpret and visualize making it easier to understand the decision making process.

WORKING of DT

It works by splitting data into smaller subsets based on conditions, forming a tree-like structure. At each node, the algorithm selects the feature that best seperates the data using certain criteria.

This spits continues untill the tree reaches the point where there is no further division is usefull, leading to leaf nodes,

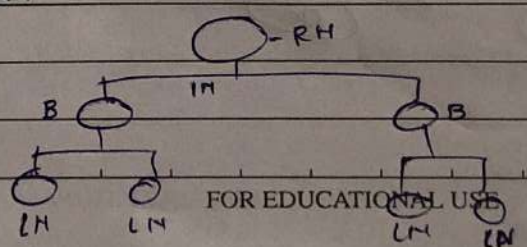
STRUCTURE INVOLVES

(i) ROOT NODE : Represents the entire dataset, which then splits into datasets

(ii) INTERNAL NODE : Represents a test or a condition on a specific attribute or the feature of data

(iii) BRANCHES : Represents the possible outcomes or values of the attribute ~~test~~ of an internal node

(iv) LEAF NODES : Represents the final decision, prediction or class label.



Sundaram
FOR EDUCATIONAL USE

Understand this with an Example

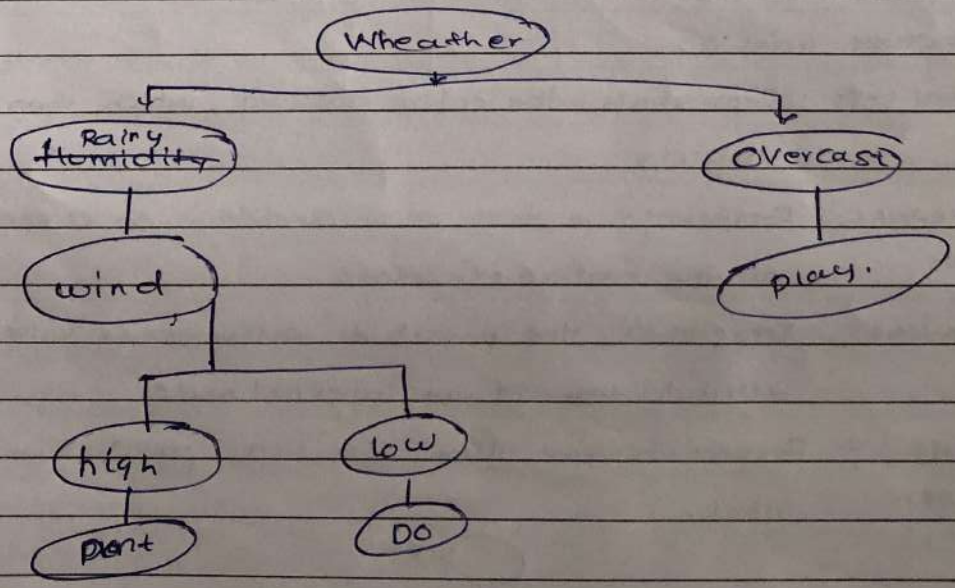
We are building an decision tree to decide wheather a person should play tennis or Not Play Tennis, based on wheather conditions. The tree makes decision based on simple question about the answer

① Root Node

Checks the wheather (Sunny, overcast or rainy.

- If overcast → Play (always good wheather)
- If Sunny → check Humidity
  - i) High - Don't Play
  - ii) Low → Play
- If Rain → check wind
  - i) Strong → Don't play
  - ii) Light → Play

This creates the simple decision tree as below



Sundaram
FOR EDUCATIONAL USE

Q. Decision Based Methods. - Non Linear & Instance Based.

ANS. Decision based methods are machine learning algorithm that make predictions by applying a series of rules or decisions. These rules split the data into groups based on feature values until a final decision is reached.

We have two types of decision based method

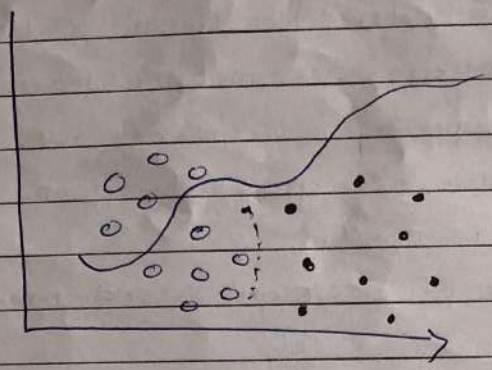
① NON LINER

Non Linear decision method works by splitting data into groups using ifelse Rules. Instead of drawing straight lines, they create flexible, curved or irregular boundaries that capture complex relationships. A decision tree is the simplest example, where each node checks a condition like, wheather = Sunny and then the branches.

Random forest combines many trees to reduce errors, while Gradient boosts also builds trees one after another to fix mistakes of the previous one. The main advantage of these methods is their ability to fix complex data very well, but they can be harder to interpret and may require more computational resources. In short, non-linear methods are choosen when simple straight-line models are not enough to solve a problem

Ex: A bank uses a decision tree model to decide if a loan application should be approved. Instead of drawing a straight-line rule, the model asks multiple conditions like "Is Credit score above 700?" and is monthly income higher than expenses?" to make non-linear decisions

Sundaram
FOR EDUCATIONAL USE



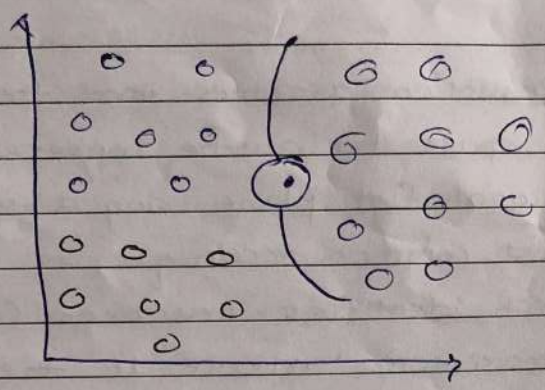
(ii) INSTANCE BASED

Instance based decision methods are different because they don't try to build a general model during training. Instead, they store the training data and use it directly to make prediction. When a new data point appears, the method ~~almost~~ look at most the similar examples from the stored data and bases the decision on them.

The common Example is K-Nearest Neighbors (K-NN), which finds the "K" closest neighbor and classifies the new point based on ~~its~~ their majority class. These methods are called "lazy learners" because they don't do much work until a prediction is needed. They rely on similarity measures to compare points. They are also sensitive to noise and irrelevant features which can mislead the decision process. Despite the drawbacks, instance-based methods are usefull when local pattern matters more than the global ones, such as in ~~global~~ recommendation systems or case-based reasoning.

Sundaram

FOR EDUCATIONAL USE



Ex: In Medical Diagnosis System, a new patient's symptoms are compared with the stored records of past patients. If most of the closest matched had "flu" then the system predicts the new patient also has flu.

Q. Dimesion Reduction - Linear & Non-Linear

Ans. Dimension Reduction means reducing the number of input features (variables) in a dataset while keeping as much as possible. In many real-world problems, datasets have hundreds or even thousands of features, which makes the training model slow, difficult and sometime inaccurate, this problem is called curse of dimensionality. By reducing dimensions, we make data easier to visualize, speed up computations and remove noise or redundant features. There are two type of main Dimension Reduction:

(i) Linear.

(ii) Non-Linear.

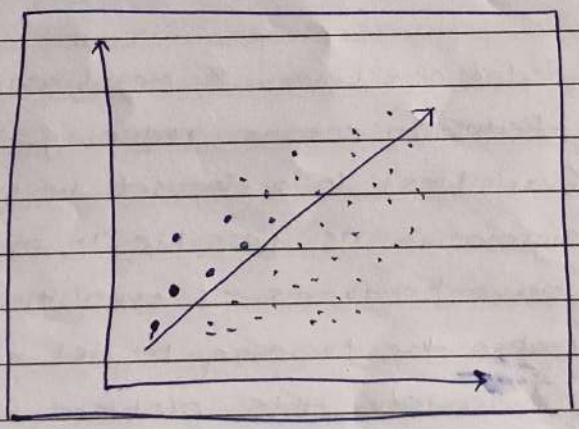
Sundaram®

FOR EDUCATIONAL USE

① Linear.

Linear Dimension Reduction assumes that the data lies in a high-dimensional space but can be represented effectively on a lower-dimensional flat plane using a straight line projections. The most common technique is Principal Compon Analysis (PCA), which identifies new axes (principal comp) that capture the maximum variance in the data.

Linear Methods are computationally efficient, easy to implement, and work well when the relationship between variables ~~and~~ are simple and mostly linear. However, they struggle when the data lies on a curved or complex surface

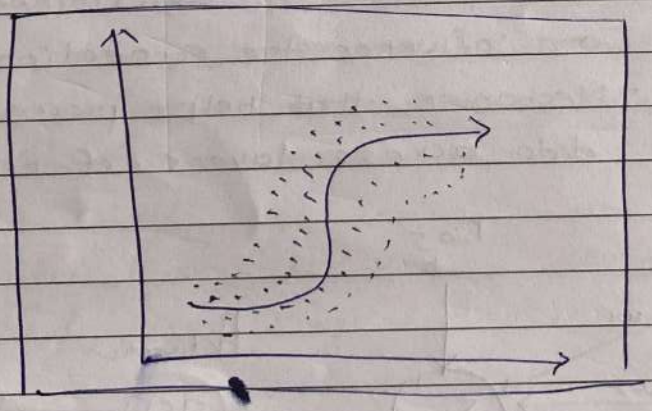


Ex: In Image Compression, PCA reduces thousands of pixels values into ~~larger~~ smaller number of components while still retaining most of the important details of the image. This allows storage transmission ~~to~~ of images with much less data.

Sundaram
FOR EDUCATIONAL USE

(ii) NON-LINEAR

Non-linear dimension reduction is used when data has complex, curved relationships that cannot be captured with the straight line. These methods assume that the data lies on a non-linear manifold - a curved hidden surface hidden in high dimension and aims to "unfold" into a simpler, low-dimensional representation. Popular techniques include t-SNE (t-Distribute Stochastic Neighbor Embedding), which preserves local neighborhood relationships and is excellent for visualization. Kernel PCA is another method which extends PCA with kernel functions to handle non-linear ~~programs~~ patterns. Non-linear methods provide powerfull insights for high-dimensional and complex datasets but require more computation and may be harder to interpret.



Ex: In NLP, t-SNE can handle or can take high dimension words embedding (like 300-dimensional vectors from word2vec) and reduce them to 2D, showing clusters of similar word. Eg: like "king" "queen" "palace" close together.

FOR EDUCATIONAL USE

Sundaram®

Q. Explain Neural Networks

ANS. A Neural Network is an machine learning model inspired by the working of human brain. Just like the brain has neurons connected togetther, a neural network also has artificial neurons called as nodes, arranged in layers. These layers process data step by step to recognize patterns, make prediction / decision or classify information.

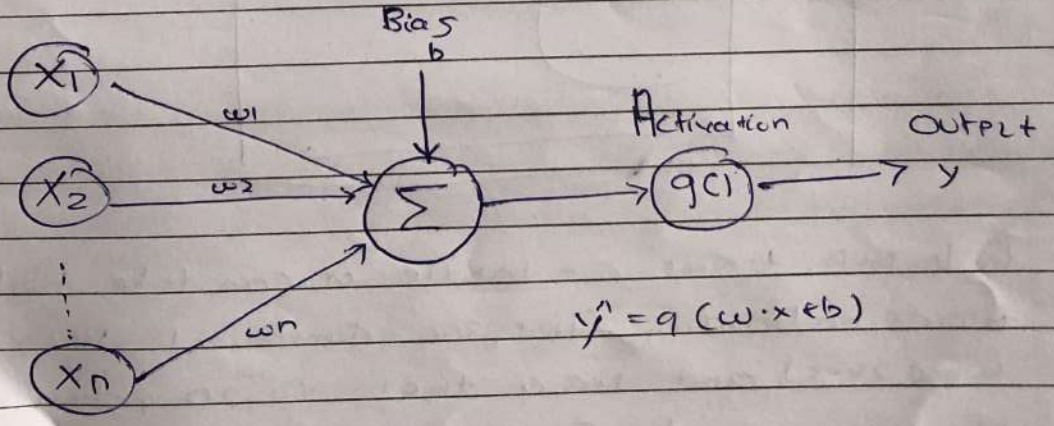
Neural Networks are capable of learning and identifying patterns and directly from data without pre-defined rules. These networks are built from several key components.

Neurons : The basic unit that recieves the input

Connections : Links between neurons that carry information regulated by weight and biases.

Wieght & Biases : These parameters determines the strength and influence the connections.

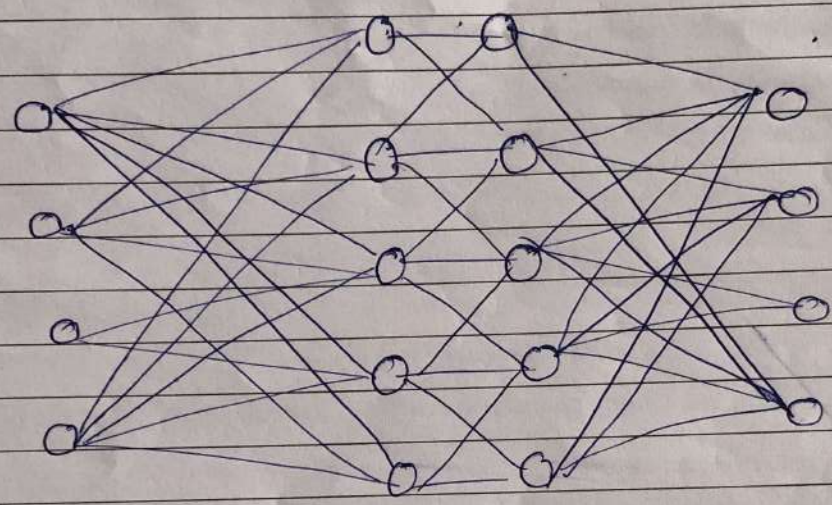
Propogation Function : Mechanism that helps process and transfer data across layers of neurons



FOR EDUCATIONAL USE

Sundaram®

Structure & working



Input layer is the first layer where the raw data enters the network, such as pixel values from image, words from text or numerical features from dataset.
Then Hidden layer come, where these are the intermediate layer where the actual learning happens. Each neuron in these layers take inputs, applies weight, adds a bias and passes it through an activation function to introduce non-linearity. Multiple hidden layers allow the network to capture very complex patterns.
Output layer, this is the final layer that produces the result, such as predicting wheather an image is a "Cat" or "dog" or giving numerical value in regression problems.

Sundaram
FOR EDUCATIONAL USE