Machine Learning Techniques (KCS 055)		
Course Outcome (CO)		Bloom's Knowledge Level (KL)
At the end of course , the student will be able:		
CO 1	To understand the need for machine learning for various problem solving	K_1 , K_2
CO 2	To understand a wide variety of learning algorithms and how to evaluate models generated from data	K_3 , K_5
CO 3	To understand the latest trends in machine learning	K_2 , K_3
CO 4	To design appropriate machine learning algorithms and apply the algorithms to a real-world problems	K_4 , K_6
CO 5	To optimize the models learned and report on the expected accuracy that can be achieved by applying the models	K_4, K_5
DETAILED SYLLABUS		3-0-0
Unit	Topic	Proposed Lecture
I	**INTRODUCTION** – Learning, Types of Learning, Well defined learning problems, Designing a Learning System, History of ML, Introduction of Machine Learning Approaches – (Artificial Neural Network, Clustering, Reinforcement Learning, Decision Tree Learning, Bayesian networks, Support Vector Machine, Genetic Algorithm), Issues in Machine Learning and Data Science Vs Machine Learning;	08
II	**REGRESSION:** Linear Regression and Logistic Regression **BAYESIAN LEARNING** - Bayes theorem, Concept learning, Bayes Optimal Classifier, Naïve Bayes classifier, Bayesian belief networks, EM algorithm. **SUPPORT VECTOR MACHINE:** Introduction, Types of support vector kernel – (Linear kernel, polynomial kernel,and Gaussiankernel), Hyperplane – (Decision surface), Properties of SVM, and Issues in SVM.	08
III	**DECISION TREE LEARNING** - Decision tree learning algorithm, Inductive bias, Inductive inference with decision trees, Entropy and information theory, Information gain, ID-3 Algorithm, Issues in Decision tree learning. **INSTANCE-BASED LEARNING** – k-Nearest Neighbour Learning, Locally Weighted Regression, Radial basis function networks, Case-based learning.	08
IV	**ARTIFICIAL NEURAL NETWORKS** – Perceptron's, Multilayer perceptron, Gradient descent and the Delta rule, Multilayer networks, Derivation of Backpropagation Algorithm, Generalization, Unsupervised Learning – SOM Algorithm and its variant: **DEEP LEARNING** - Introduction,concept of convolutional neural network , Types of layers – (Convolutional Layers , Activation function , pooling , fully connected) , Concept of Convolution (1D and 2D) layers, Training of network, Case study of CNN for eg on Diabetic Retinopathy, Building a smart speaker, Self-deriving car etc.	08
V	**REINFORCEMENT LEARNING**–Introduction to Reinforcement Learning , Learning Task,Example of Reinforcement Learning in Practice, Learning Models for Reinforcement – (Markov Decision process , Q Learning - Q Learning function, Q Learning Algorithm), Application of Reinforcement Learning,Introduction to Deep Q Learning. **GENETIC ALGORITHMS:** Introduction, Components, GA cycle of reproduction, Crossover, Mutation, Genetic Programming, Models of Evolution and Learning, Applications.	08
Text books:		

27/09/22

UNIT-1

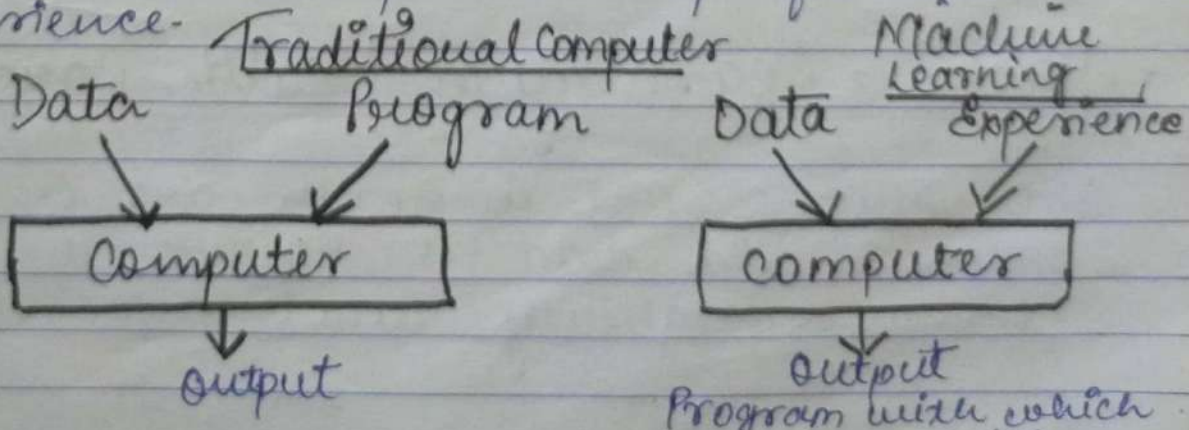
Introduction

Learning:- Learning is any process by which a system improve performance from experience. Learning denote changes in the system that enable a system to do the same task more efficiently next time.

We can say learning is a change in knowledge, skills and behaviour which outcomes is experience

Machine Learning:- It is a subfield of computer science that evolved from the study of computational learning theory in artificial intelligence (A.I)

According to Arthur Samuel in 1959 define machine learning as a field of study that gives computer ability to learn without being explicity program in machine learning for any task, a machine improve its performace with this experience.



Ex:- Image Recognition, Speech Recognition, Medical, Social Networking, Gmail.

Scope of Machine Learning:-

Machine Learning in Search Engine:-

Search engine rank page based on what you are most likely to click on search engine depends on machine learning to improve this service such as:- voice recognition, image search and many more.

Machine Learning in Digital Marketing:-

Machine Learning can help significantly in digital marketing it allow more relevant personal, thus, company can interact and engage with customer Thus, company have information which can be used to learn customer behaviour and send message and display ad according to their interest to increase company sell.

Machine Learning in Health Care:-

The value of M.L in health care is its ability to process used data set beyond the scope of human capability and then reliable convert,

analysis of that data into clinical insights that did physician in planning and providing care, ultimately leading to better outcomes lower cost of care and increase patient satisfaction.

Spam Detector :- our mailing agent like G-mail does a lot of hardwork for us in classifying the mail and moving the spam (folder) mail in spam folder this is achieved with the help of M.L.

Traffic Alert :- Google maps application is used to give assistance in direction and traffic also based on M.L. Everyone using app is providing their location, average speed, the route in the travelling. help google to collect massive data about the traffic which makes them predict the upcoming traffic and adjust your route according to it

Social Media :- Social media like facebook, face recognition, face detection to automatically find the face of person which matches its database and hence suggest to try that person, this is also possible with the help of M.L.

8/9/22

Limitation of Machine Learning :-

Accuracy depends on training and learning which is not always available.

Have large data requirement to learn about various topics which may be time taking and require various resources.

Performance can not be audited or guranteed

Machine can not learn if there is no data available.

A machine needs to have hetrogenity in data sets to learn meaningful insights

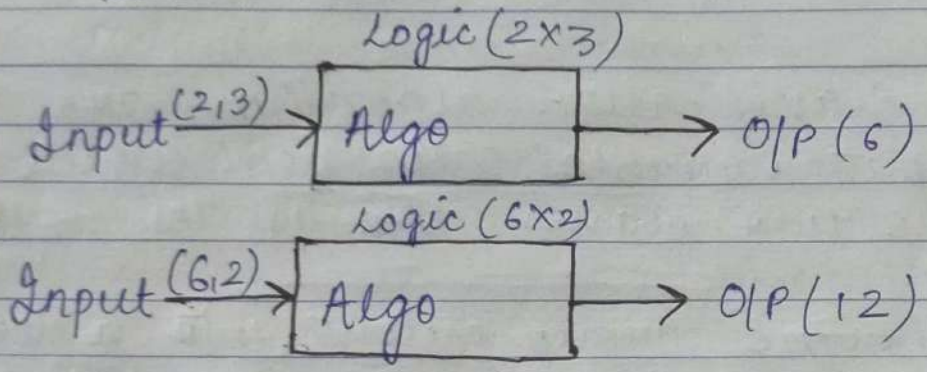
Lack of data or the variation in data sets leads to poor evaluation and prediction.

Machine Learning

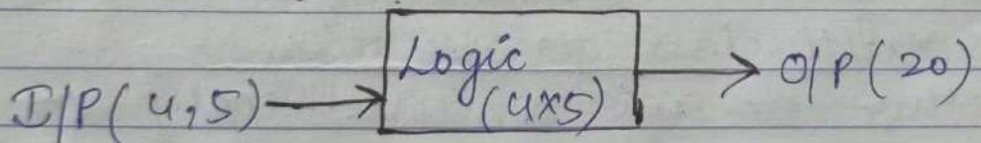
- Supervised Learning
  - → classification
  - → Regression
- Unsupervised Learning
  - → clustering
  - → Association
- Reinforcement Learning

Supervised Learning :- Supervised Learning is a type of machine Learning in which machines are trained using well labeled training data, on the basis of that data machine predicts output.

In this learning, given training example of input and output machine can predict the output for new input.
It is widely used for data where there is a precise mapping between input - output data. We can assume a data set as a teacher or supervisor and its role is to train the machine model.
9/9/22
Once the machine gets trained it can start making prediction or decision when new data given to it.



Training data set



Prediction with new data

Supervised Learning upper when a system is given input and output variable with an intentions of learning how they are mapped together.
The goal is to produce an accurate enough mapping functions that when new input is given, the algorithm can predict the output. This is an iterative process and each time the

algo makes a prediction, it is corrected or given feedback until it achieve an acceptable level of performance

Types of Supervised Learning:-

(i) classification
(ii) Regression

Classification:- classification algorithm are used when output variable categorical which means there are two classes like as Yes or No True/False etc.

classification technique group output into a class if the algorithm try to i/p input into distinct classes, it is called Binary classification.

classification selecting between more than two classes is known as multi class classification.

- Determining whether or not someone will apply for loan. Yes/No.
- The mail is spam or not.

Supervised Learning Algo for Classification

→ Support Vector Machine (SVM)
→ Naive Bayes Algo
→ Nearest Neighbor Algo
→ Neural Network
→ Decision Tree

→ Random forest

Regression:- Regression algorithm are used if there is a relationship between the input variable and output variable. It is used for the prediction of continuous variable such as weather forecasting, market trend etc.

Regression technique predict a single output value using training data.

We can use regression to predict the house price from training data. The input variable will be locality, size of house etc.

Algorithm used in Regression:-

→ Support Vector Machine (SVM)
→ Linear Regression
→ Bayesian Linear Regression
→ Neural Network

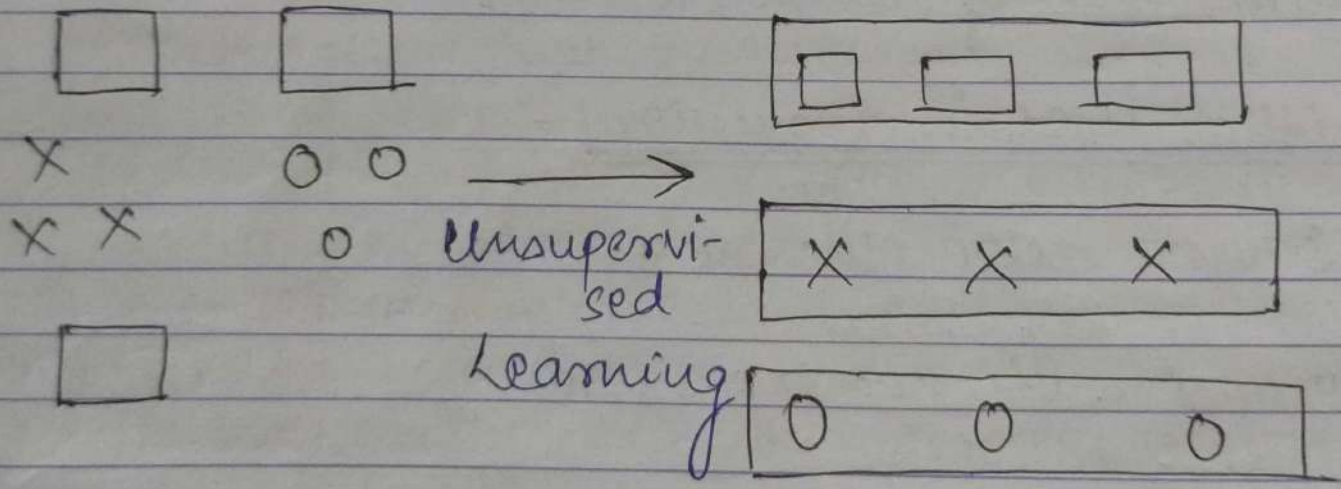
12/9

UnSupervised Learning:- Unsupervised learning deals with unlabelled data, No teacher (training data set) is provided that means no training will be given to the machine therefore machine works on its own to discover information.

Here, the task of machine is to groove unsorted information according to the similarities, pattern and differences without any prior training of data

Unsupervised learning algorithm allow to perform more complex processing task compare to supervised learning. Unsupervised learning algo try to use technique on the input data to make for rules, detect and summarize and groove the data point which helps in deriving meaningful inside and describe the data better to the users

Unlabelled data might includes, photos, video, audios etc. There is no explanation for each piece of unlabelled data, it just contains the data nothing else



Types Of Unsupervised Learning:-

(i) clustering
(ii) Association

19/9

UNIT-1

(Machine Learning Technique) Introduction

Machine learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

Machine learning is programming computers to optimize the performance criteria using an example data and past experience. We have a model defined upto some parameters and learning is a execution of a computer program to optimize the parameters of model using the training data or past experience. The model may be predictive to make predictions in the future or descriptive to gain knowledge from the data.

The field of study known as Machine learning is concerned with the question of how to construct computer programs that automatically improve experience.

Learning:- A computer program is said to learn from experiences 'E' with respect to some class of task 'T' and performance measure 'P', if its performance at task 'T', as measured by 'P', improves with experience 'E'

Ex:- (1) Handwriting recognition learning problem

Task T = Recognizing and classify handwritten words within image

Performance P = percent of words correctly classified.

Experience E = A data set of handwritten words with given classification

Ex. ② A robot driving learning problem

Task T = driving on highways using vision sensors

Performance P = Average distance travelled before an error.

Experience E = A sequence of images and steering command recorded while observing a human driver

2019

Classification of Machine Learning:-

Machine learning implementations are classified into four major categories. depending on the nature of learning "signal" or "Respons" available to learning system which are as follows.

① **Supervised Learning:-** Supervised learning is the machine learning task of learning a function that maps an input to an output based on example of input output pairs. The given data is labelled. Both classification and Regression problems are supervised learning problems

for ex:- Consider the following data regarding patients entering a clinic. The data consist of the gender and the age of the patients and each patient is labelled as healthy or sick

patient Entry:-

Gender	Age	label
Male	20	Sick
female	25	Healthy
female	19	Healthy
Male	24	Healthy
Female	32	Sick

② **Unsupervised Learning:-** Unsupervised learning is a type of machine learning algorithm used to draw interface from data sets consisting of input data without labelled responses. In unsupervised learning algorithms, classification or categorization is not included in the observations.

ex:- consider the following data regarding patients entering a clinic. The data consist of the gender and age of the patients.

Gender	Age
Male	54
female	60
Male	64
female	93
Male	10
female	29

③ **Reinforcement Learning:-** Reinforcement learning is a (Hit n trial method) problem of getting an agent to act in the world so as to maximize its rewards. A learner is not told what actions to take as in most forms of machine learning but discover which actions yield the most rewards by trying them.

for example:- Consider Teaching a dog a new trick we cannot tell it what to do. But we can reward / punish it if it does the right / wrong thing

(4) Semi Supervised Learning :- Where an incomplete training signal is given, a training set with some of the target output missing. There is a special case of this principle known as transduction where the entire set of problem instances is known as learning time, except that part of the targets are missing. Semisupervised learning is an approach to machine learning that combines small labelled data with a large amount of unlabelled data during training. Semisupervised learning falls between unsupervised learning and supervised learning.

21/9 Categorizing Based on Required Output :-

Classification :- When inputs are divided into two or more classes, the learner must produce a model that assign unseen inputs to one or more of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification where the inputs are email messages and the classes are spam and not spam.

Regression :- Regression is also a supervised problem. A case when the outputs are continuous rather than discrete. Continuous

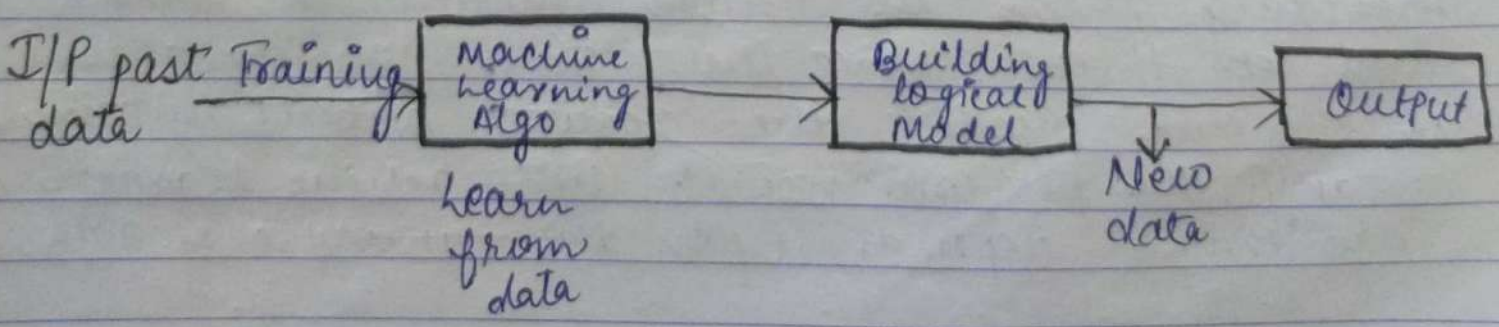
variables are the variables that have infinite number of possibilities. for ex:- A persons weight. So someone could be 180 kg or they could be 180.10 kg or 180.110 kg.

Clustering:- when a set of inputs is to be divided into groups. Unlike in classification, the groups are not known before, making this typically and unsupervised task.

How Does Machine Learning Work:-

A machine learning system learns from historical data builts the prediction model and whenever it receives new data, predicts the output for it The accuracy of predicted output depends upon the amount of data, as a huge amount of data helps to build a better model which predicts the output more (correctly) accurately.

Machine learning has change our way of thinking about the problem. The below block diagram explains the working of machine learning algorithm.



Features of Machine Learning:-

- → Machine learning uses data to detect various patterns in a given data set.
- → It can learn from past data and improves automatically.
- → It is a data driven technology.
- → Machine learning is much similar to data mining as it also deals with the huge amount of data.

Need of Machine Learning:-

The need of machine learning is increasing day-by-day. The reason behind the need for machine learning is that it is capable of doing task that are too complex for a person to implement directly. As a human we have some limitations as we cannot access huge amount of data manually, so for this, we need some computer system and here come the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the model and predict the required output automatically. The performance of the machine learning algorithms depends upon the amount of data.

and it can be determined by the cost function. With the help of machine learning we can save both the time and money.

The importance of machine learning can be easily understood by its uses, currently machine learning is used in self driving cars, cyber fraud detection, face recognition and friend suggestion by facebook etc. Various top companies such as Netflix and amazon have build machine learning models that are using a vast amount of data to analyse the user interest and recommend product accordingly.

Machine learning at Present:-

Now machine learning has got a great advancement in its research and it is present everywhere around us such as amazon, alexa, recommender system, self driving cars and many more. It include supervised, unsupervised and reinforcement learning with clustering, classification, decision tree, ~~svm~~ algorithms etc.

Modern machine learning models can be used for making various predictions including weather prediction, disease prediction, stock market analysis etc.

Machine learning Life Cycle:-

Machine learning has given the computer systems the ability to automatically learn without being

explicitly programmed. But how does a machine learning system work it can be describe using the life cycle of machine learning. Machine learning life cycle is acyclic process to build an efficient machine learning project. The main purpose of the lifecycle is to find a solution to the problem or project

Machine learning life cycle involves seven major steps which are given below:-

1. Gathering data
2. Data preparation
3. Data wrangling
4. Analyse data
5. Train the Model
6. Test the Model
7. Deployment

The most important thing in the complete process is to understand the problem and to know the purpose of the problem. Therefore before starting the life cycle we need to understand the problem because good result depends on the better understanding of the problem.

In the complete life cycle process to solve a problem we create a machine learning model and this model is created by providing training. But to train a model we need data. Hence life cycle start by collecting data.

① **Gathering data:-** Data gathering is a first step of the machine learning life cycle. The goal of this step is to identify

and obtain all data related problem. In this step we need to identify the various different data sources as data can be collected from various sources such as files, database, internet or mobile devices. It is one of the most important step of the life cycle. The quantity and quality of collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the production. The steps includes are given below

1. Identify various data sources.
2. Collect data
3. Integrate the data obtain from different sources.

Data Preparation:- After collecting the data we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training. In this step first we put all data together and then randomise the ordering of data. This step can be further divided into sub processes.

1. **Data Exploration:-** It is used to understand the nature of the data that we have to work with. We need to understand the characteristics, format and quality of data.
2. **Data Preprocessing:-** This steps is for preprocessing of data to analyse it completely.

Data Wrangling:- This is the process of cleaning and converting raw data into a useable format. Cleaning of the data, selecting the variable to use and transforming the data in a proper format to make it more suitable for analysis is a purpose of this step. It is one of the most important step of the complete process. In real world applications, collected data may have various issues including missing values, duplicate data, invalid data etc so we use various filtering techniques to clean the data

Data Analysis:- Now the cleaned and prepared data is used to analysis.

This step involves:-

1. Selection of analytical techniques.
2. Building Models.
3. Review the result

The aim of this step is to build the machine model to analyse the data using various analytical techniques and review the outcomes. It starts with the determination of the type of the problems where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association etc then build the model using prepared data and evaluate the model.

Train Model:- Now the next step is to train the model. In this step we train our model to improve its performance for better outcome of the problem. We use data sets to train the model using various machine learning algorithm. Training a model is required so that it can understand the various pattern, rules and features.

Test Model:- Once our machine model has been trained on a given data set, then we test the model. In this step we check for the accuracy of our model by providing a test data set to it. Testing the model determines the percentage of accuracy of the model as per the requirement of project or problem.

Deployment:- The last step of machine learning life cycle is deployment, where we deploy the model in the real world system.

Advantages of Supervised Learning:-

- With the help of SL the model predict the output on the basis of previous experience.

- In supervised learning, we can have an exact idea about the classes of object.
- SL model helps us to solve various real world problems such as fraud detection, spam filtering etc.

Disadvantages of Supervised Learning:-

- SL models are not suitable for handling complex task
- SL cannot predict the correct output, if the test data is different from the training data set.
- Training required lots of computational time
- In SL we need enough knowledge about the classes of objects

Advantages of Unsupervised Learning:-

- USL is used for more complex task as compared to SL because in unsupervised learning we don't have labelled input data
- USL is preferrable as it is easy to get unlabelled data in comparison to labelled data

Disadvantages of Unsupervised Learning:-

- USL is more difficult than SL as it does not have corresponding output

- The result of USL algorithm might be less accurate as input data is not labelled, and algorithms do not know the exact output in advance.

Difference between Supervised Learning and Unsupervised Learning:-

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

• SL can be categorized in classification and Regression problems	• USL can be classified in clustering and Association problems
• SL can be used for those cases where we know the input as well as (output) corresponding outputs	• USL can be used for those cases where we have only input data and no correspon--ding output data.
• SL model produces an accurate result.	• USL model may give less accurate result as compare to SL
• SL is not close to true AI as in this we first train the model for each data and then only it can predict the correct output.	• USL is more close to true AI as it learns simila--rly as a child learns daily routine things by his experiences.
• It includes various algorithm such as linear regression, logistic regression, support vector machine, multi class classification, Decision tree, Bayesian logic etc	• It includes various algorithm such as clustering, KNN, apriory algorithm.

26/9

Well Defined Learning Problems :-

A computer program is said to learn from experience E in context to some task T and some performance major P, If its performance T as measured P upgrade with experience E.

① To better filter emails as spam or not.
Task, T = classify emails as spam or not.
Performance, P = The fraction of emails accurately classified as spam or not spam.
Experience, E = observing your label emails as spam or not spam.

② checker's learning problem.
T = playingchecker's game
P = percent of game won against opposer.
E = playing implementation games against itself.

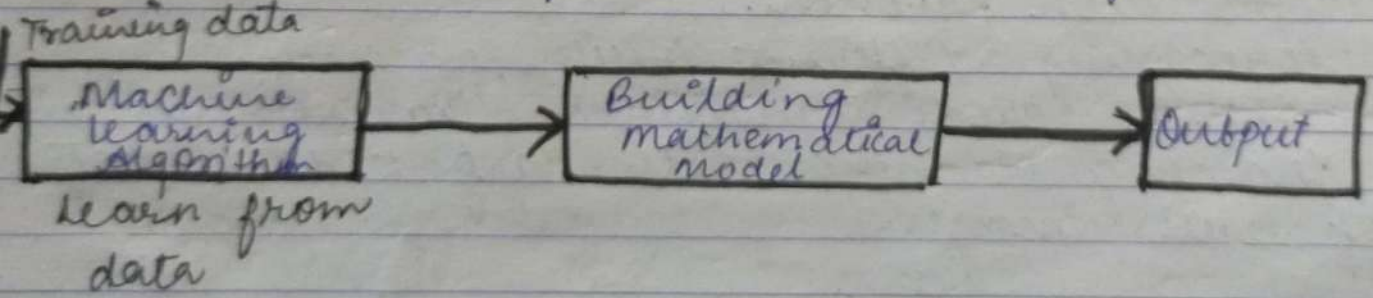
③ fruit prediction problem.
T = forecasting different fruits for recognition.
P = Able to predict maximum variety of fruit.
E = training machine with largest data sets of fruits images.

④ face recognition problem
T = Predicting different types of faces.
P = Able to predict maximum typis of faces.
E = training machine with maximum amount of data set of different face images.

5. Automatic Translation document
T= Translating one type of language used in a document to other language.
P= Able to convert one language to other efficiently.
E= Training machine with a large data set of different types of languages.

Design a learning System :-

When we fed the training data to machine learning algorithm, this algorithm will produce a mathematical model and with the help of that mathematical problem. The machine will make a prediction and take a decision without being explicitly program. Also during training data, the more machine will work with it, the more it will get experience and the more efficient result is produced.



Steps for Designing Learning Systems are:-

Step ① :- Choosing the training experience :-

The very important and first task is to choose the training data which will be fed to the machine learning algorithm. It is important to note that the data or experience that we fed to algorithm must have a significant impact on the success or failure of the model. So, training data should be choosen wisely. Below are the attributes which will impact

- On the success and failure of the data.
- The training experience will be able to provide direct or indirect feedback regarding choices
- Second important attribute is the degree to which the learner will control the sequences of training examples.
- Third important attribute is how it will represent the distribution of examples over which performance will be measured. Machine learning algorithm will get more and more experience by passing through more and more examples and hence its performance will increase.

Step ②:- choosing the target function:-

The next important step is choosing the target function it means according to the knowledge fed to the algorithm the ML will choose next move function which will describe what type of legal moves should be taken.

Step ③:-choosing representation for target function:

When a ML algorithm will know all the possible legal moves. The next step is to choose the optimised moves go using any representation that is using linear equations, hierarchical graph representation tabular form etc.

Step ④ :- Choosing function approximation algorithm:-

An optimised move can not be choosen just with the data set. The training data had to go through with set of examples and through these examples the training data will approximates which steps are choosen an after that machine will provide feedback on it.

Step ⑤ :- final design :-

The final design is created at last when system goes from number of examples, failure and succes, correct and incorrect decision and what will be the next step etc.

22/9

Artificial Neural Network:-

Artificial Neural Network is robust to errors in the training data and has been successfully applied for learning real valued functions

containing problems such as interpreting visual senses, speech recognition, learning robot control (~~strategies~~) strategids.

The study of ANN has been inspired by the observation of biological learning system. These biological systems are build of various complex webs of interconnected neurons in the brain. ANN systems are motivated to capture this kind of highly parallel computation based on distributed representation.

ANN are build of densely interconnected set of simple units, where each units takes a number of real valued inputs and produces a single real valued output. But ANN are less motivated by biological neural ~~networks~~ systems, as There are many complex neural systems that are not modelled by ANN.

Difference between Biological Neurons & Artificial Neurons:-

Biological Neurons	Artificial Neurons
• Major component in biolog-ical neuron → Axion → Dendrites → Synapse	• Major component in artifi-cial neuron → Nodes → Input → Output → Weights

- Information from other neurons, in the form of electrical impulses, enters the dendrites at connection points called synapse. The information flows from the dendrites to the cell where it is processed. The output signals, a train of impulses is then sent down the axion to the synapse of other neurons

- A synapse is able to increase or decrease the strength of the connection. This is where information is stored.
- Approx 10^{11} neurons

- The arrangements & connections of the neurons made up the network and have three layers. The first layer is called the input layer and is the only layer exposed to external signal. The input layer transmits signals to the neurons, in the next layer which is called a hidden layer. The hidden layer extract relevant features or patterns from the receive signals. Those features or pattern that are considered important are then directed to the output layer, which is the final layer of the network.
- The artificial signals can be changed by weights in a manner similar to the physical changes that occur in the synapses.
- (Approx) 10^2 - 10^4 neurons with current technology.

Difference between Human Brain & Computers in terms of how information is processed

Human Brain	Computers
• Biological Neuron Network	• Artificial Neural Network (ANN)
• The human brain works asynchronously.	• Computer work synchronously
• Biological Neurons compute slowly.	• Artificial Neuron compute fast (less than 1nano second per computation
• The brain represents information in a distributed way because neurons are unreliable and could die anytime.	• In computer programs every bit has to function as intended otherwise these program would crash
• Our brain changes their connectivity over time to represents new information and requirements imposed on us.	• The connectivity between the electronic components in a computer never change unless we replace its components.
• Biological Neural Network have complicated topologies	• ANN are often in a tree structure.
• Researchers are still to find out how the brain actually learns.	• ANN use gradient learning.

Advantages of Artificial Neural Network

- Problem in ANNs can have instances that are represented by many attribute-value pairs.
- ANNs used for problems having the target function output may be discrete-valued, real-valued or a vector of several real- or discrete-valued attributes.
- ANNs can bear long training times depending on factors such as the no. of weights in the network, the no. of training examples considered and the settings of various learning algorithm parameters.
- It is used generally used where the fast evaluation of the learned target function may be required.

Disadvantages of ANNs:-

- ANNs require lots of computational power.
- Neural Network models are hard to explain.
- Neural Network training requires lots of data.
- Data preparation for neural network models needs careful attention.
- Optimizing neural network models for production can be challenging.

29/9

Clustering:- (Clustering is the most important type of unsupervised learning) The task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar to each other then to those in other cluster

Clustering is useful for many applications. for example:- It can be used to automatically organized form of data.

eg:- The plant species or animal species or news document or book.

It can be also used for understanding hidden structure in data and some times clustering is used as preprocessing for further analysis of the data.

Applications of Clustering:-

1. **Biology:-** Classification of plants and animal kingdom given their feature.
2. **Marketing:-** Customer segmentation based on a database of customer data containing their properties and past buying records.
3. **Web Log:-** clustering web log data to discover groups of similar access (data) patterns.
4. Recognize communities is in social Network

(Major Measure) Classification of Clustering Technique:-

① **Partitioning:-** Construct various partition and then evaluate them by some criteria

② **Hierarchical:-** create a hierarical decomposition of the set of object using some criteria.

③ **Model Based:-** Hypothise a model for each cluster and find best fit of models to data.

④ **Density Based:-** Guided by connectivity and density functions.

⑤ **Graph theortic Clustering:-** Based on the under construction of a graph and looping at some graph theortic measures like dnd so on.

Difference b/w Classification & Clustering

Both classification and clustering used for

categorization of objects into one more classes based on the features. In case of classification there are pre defined labels assigned to each input instance according to their properties whereas in clustering these labels are missing.

(Type) Parameter	Classification	Clustering
(1) Types	• Used for SL	• Used for USL
(2) Basic	• process of classify-ing the input instances based on their correspo-uding class labels.	• Grouping the instan-ces based on their similarity without the help of class labels.
(3) Need	• It has labels so there is need of training & testing data set for verifying the model created	• There is no need of training & testing data sets.
(4) Complexity	• More complex as compared to clustering	• Less complex
(5) Examples	• Logistic Regression Naive Bayes classif-ier, Support Vector Machine etc.	• K-means clustering algorithm, fuzzy C means clustering algor-ithm, Gaussian clusteri-ng algorithm

Quality of Clustering:- The quality of a clustering result depends on the algorithm, the distance function & the application.

Decision Tree:-

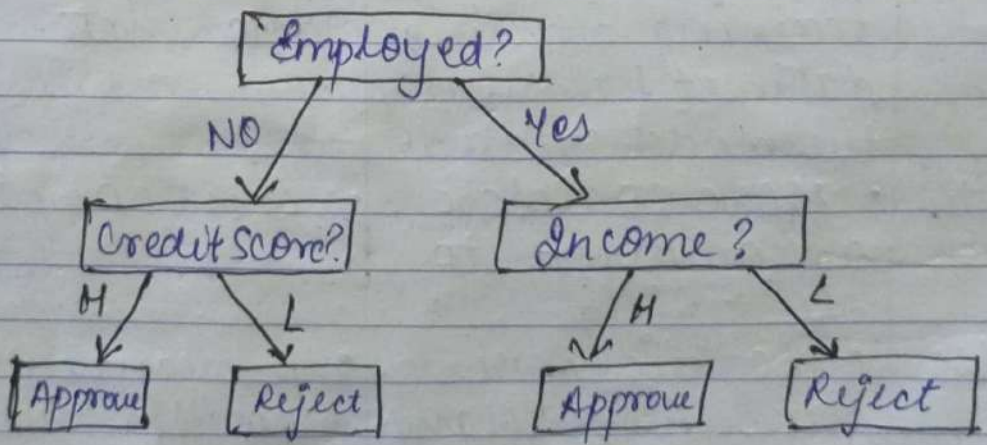
A tree has nodes and branches. A rooted tree has root node and children and then leaves which have no children. A decision tree is a tree having two types of nodes

→ Decision node

→ Leaf node

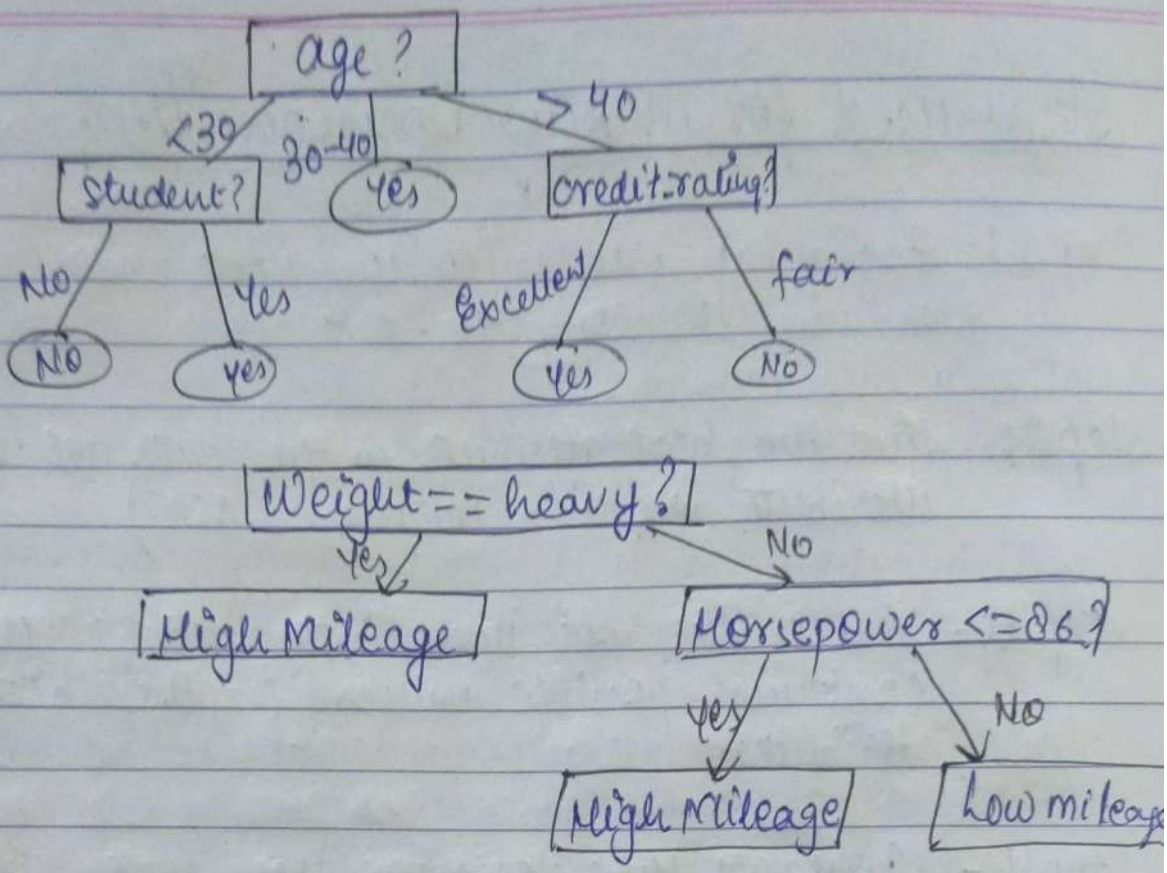
Decision node specify a choice or test based on this you can decide which direction you can go. In decision tree we test something and that test may have more than one result and based on the value of this test you follow the particular branch.

Ex:- Decision tree about whether to approve a loan.



Ex:- (2) Whether the person likely to buy a computer

(3) Decision tree for car milage prediction



Decision tree algorithm falls under the category of supervised learning. They can be used to solve both Regression and Classification problem.

Decision tree used tree representation to solve the problem in which each leaf node correspond to a class label and attributes are represented on the internal node of the tree. Internal nodes represent the features of a data set, branches represent the decision rules. and each leaf node represent the outcome.

In decision tree, there are two nodes used to make any decision and have multiple branches whereas leaf nodes are the output of those decision and do not contain any further branches. The decisions or the test are perform on the basis of features of the given data set.

Steps used for Making Decision Tree:-

Step 1:- Began the tree with the root nodes which contains the complete data set.

Step 2:- find the best attribute in the data set using attribute selection measure (ASM)

Step 3:- Divide the root node into subsets that contains possible values for the best attributes.

Step 4:- Generate the decision tree node, which contains the best attribute.

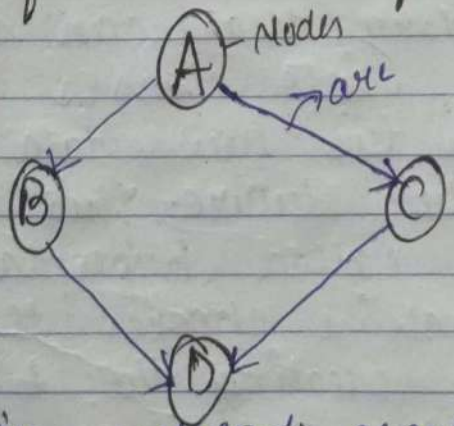
Step 5:- Recursively make new decision tree using the subset of the data set created in the step 3. Continue this process until a stage is reached where you can not further classify the nodes and called final node as a leaf node.

Bayesian Network:- Bayesian belief network is key computer technology for dealing with probabilistic events and to solve a problem which has uncertain-ity. A Bayesian network is a probabilistic graphical model which represent a set of

(problems) variables and their conditional dependency using directed Acyclic graph
It is also called a Bayes Network, Belief Network, Decision Network or Bayesian Model.
Bayesian Networks are probabilistic because these networks are built from a probability distribution. An also use probability theory for prediction and detection.

It can also be used in various task including Prediction, Anamoly detection, Diagnostic, Automated insight, reasoning, time series prediction and Decision making under uncertainity.
Bayesian Network can be used for building models from data and experts opinions and it consist of two part.

1. Directed Acyclic Graph
2. Table of conditional probabilities



A Bayesian network graph is made up of nodes and arcs where each node corresponds to the random variables and a variable can be continuous or discrete. Arc or directed arrows represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.

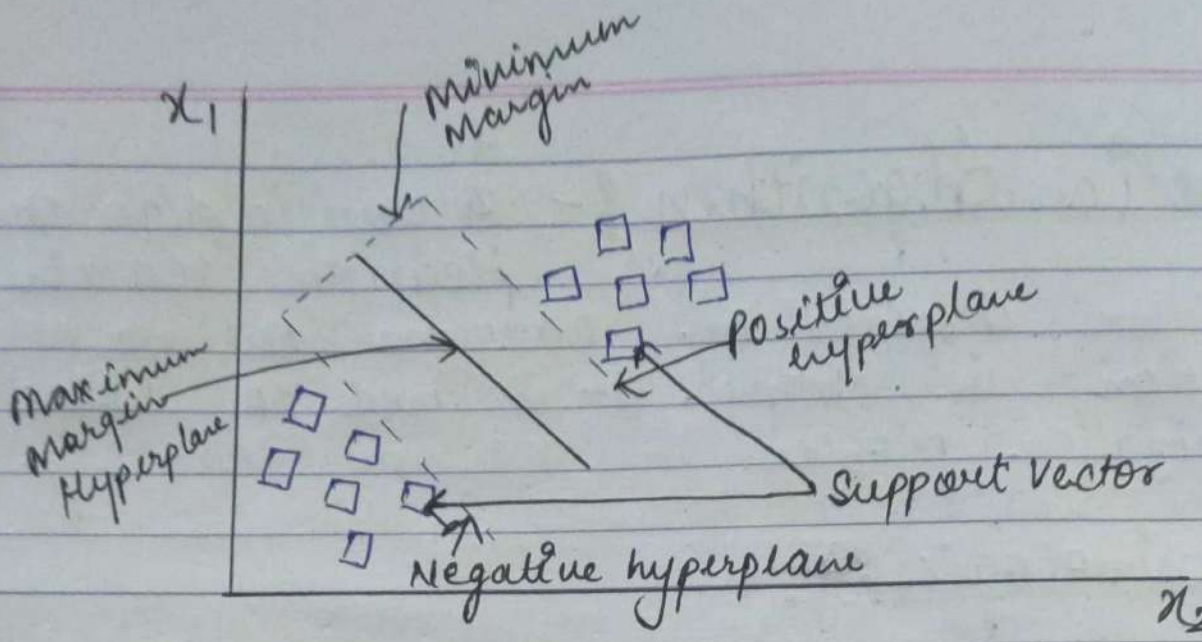
These links represent that one node directly influenced the other node & if there is no directed link that means that node are independent with each other.

The Bayesian Network has mainly two component

(i) Causal component
(ii) Actual component

Support Vector Machine:- Support Vector Machine is one of the most popular learning algorithm which is used for classification and as well as Regression problem. However, primarily it is used for classification problem in Machine learning. The goal of the Support Vector Machine algorithm to create the best line and decision boundary that can seggregrate n dimensional space into classes so that we can easily put the new data point in correct category in the future. This best decision boundary is called hyperplane.

Support vector Machine chooses the extreme points that help in creating the hyperplane. These extreme cases are called as Support Vector and Hence Algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane



SVM algorithm can be used for face detection, image classification, text categorization etc.
SVM is of two types
① Linear SVM
② Non Linear SVM

Advantages of SVM:-

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient.

Disadvantages:-

- SVM algo. is not suitable for large data sets.
- SVM does not perform very well when data set has more noise i.e target classes are overlapping.
- In cases where the no. of features for each data point exceeds the no. of training data samples, the SVM will underperform.

Genetic Algorithm:- A Genetic Algorithm is Heuristic search algorithm inspired by Darwins theory of Evolution in Nature. It is used to solve optimization problem in machine learning.

Terminologies:-

① **Population:-** population is subset of all possible solution.

② **chromosomes:-** One of the solution in the population for the given problem is called chromosomes

③ **Collection of Gene:-** Collection of Gene generate chromosome.

④ **Gene:-** A chromosome is divide into different gene or it is an element of a chromosome

⑤ **Allele:-** It is the value provided to the gene within a particular chromosome.

⑥ **fitness function:-** The fitness function is used to determine the individuals fitness level in the population. It means the ability of an individual to compete with other individuals.

In every iteration, individuals are evaluated based on their fitness function.

6. Genetic operator:- The best individual to generate offspring better than parents. Here genetic operators plays a role in changing the genetic composition of the next generation.

7. Selection:- After calculating the fitness, every existence in the population, a selection process is used to determine which of the individuality in the population will get to reproduce and produce the seed that will form coming generation

Types of Selection Style available.

1. Roulette wheel selection
2. Event selection
3. Rank Grounded selection

Working of Genetic Algorithm:-

13/10

UNIT-2

Regression Analysis

Regression Analysis :-

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables it is mainly used for prediction, forecasting, time series modeling and determining the casual effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints Using this plot, the Machine learning model can make predictions about the data. In simple words, Regression shows a line or curve that passes through all the data points on target predictor graph in such a way that the vertical distance between the data points and the regression line is minimum The distance between data points and line tells whether a model has captured a strong relationship or not.

Some examples of Regression can be as follows

1. Prediction of Rain using Temperature & other factors
2. Determining Market Trends.

3. Prediction of road accidents due to Rash driving

Terminologies Related to the Regression Analysis :-

1. **Dependent Variable :-** The main factor in regression analysis which we gone to predict or understand is called the dependent variable. It is also called Target Variable.

2. **Independent Variable :-** The factors which affect the dependent variables or which are used to predict the values of the dependent variable are called Independent Variable, Also called as a predictor.

3. **Outliers :-** Outliers is an observation which contains either very low values or very high value in comparison other obse-rve values. An Outlier may hamper the result so it should be avoided

4. **Multicollinearity :-** If the independent variables are highly correlated with each other then other variables their such condition is called Multicollinearity. It should

not be present in the data set, Because it creates problem while ranking the most affecting variable.

Underfitting and Overfitting:-

If our algorithm works well with the training data but not well with the test data then such problem is called over fitting. If our algorithm does not perform well even with training data set then such problem is called underfitting.

Regression analysis is a statical method to model the relationship between a target and predictor variables.

Needs of Regression Analysis:-

Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales predictions, marketing trends etc., for such case we need some technology which can make prediction more accurately so for such case we need regression analysis which is a statistical method and used in ML and data science. Below are some other reasons for using Regression Analysis.

1. Regression estimate the relationship between the target & independent variables.
2. It helps to predict real / continuous values.
3. By performing the Regression we can confidently determine the most important factor, the least important factor and how each factor is affecting the other factors.

Types of Regressions:-

There are various types of Regressions which are used in data science and Machine Learning. Each type has its own importance on different scenarios, But at the core, All the regression method analyse the effect of independent variable on the dependent variable. Some important types of Regressions are:-

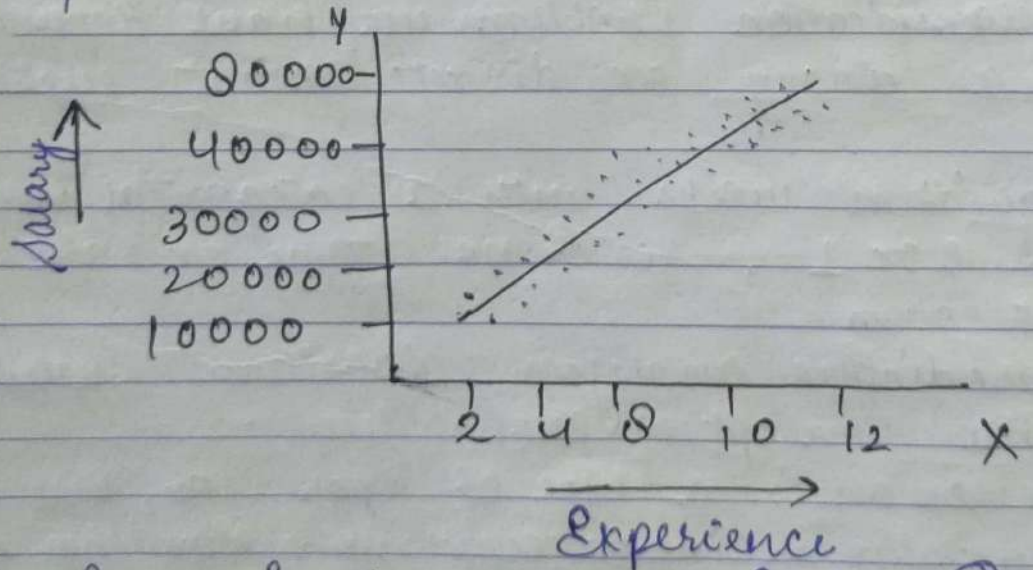
1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Support Vector Regression
5. Decision Tree Regression
6. Random forest Regression
7. Ridge Regression
8. Lasso Regression

Linear Regression:-

- Linear Regression is a statistical regression

method which is used for predictive analysis

- It is one of very simple and easy algorithm which works on regression and shows the relationship between the continuous variable.
- It is used for solving the regression problem in ML.
- Linear Regression shows the linear relationship between the independent variable (x-axis) and dependent variable (y-axis). Hence called linear Regression.
- If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variables, then such linear regression is called Multiple linear regression.
- The relationship between variable in the linear regression model can be explain using the below diagram. Here we are predicting the salary of an employee on the basis of the year of experience.



- Below is the Mathematical eqⁿ of Linear Regression

$$y = \alpha_0 + \alpha_1 x$$

where,

y = target variable / Dependent variable

x = Independent variable / Predictor

α_0 & α_1 = are the Regression coefficient

Note :- $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots \alpha_m x_m$

Above eq(n) is for multiple Linear Regression.

- Some popular application of Linear Regression are:
  - i. Analyzing Trends and sales estimation
  - ii. Salary forecasting
  - iii. Real estate prediction.

Logistic Regression :-

i) Logistic Regression is another supervised learning algorithm which is used to solve the classification problem. In classification problems we have dependent variable in a binary or discrete format such as 0 or 1

ii) Logistic Algorithm works with a categorical variable such as 0 or 1, Yes or No, True or False, spam or Not spam

iii) It is a predictive analysis algorithm which works on the concept of probability

iv) Logistic Regression is a type of Regression, but it is different from Linear regression algorithm in the term how they are used.

vi) Logistic Regression uses sigmoid function or logistic function which is a complex cost function this sigmoid function is used to model the data in logistic Regression. The function can be represented as

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

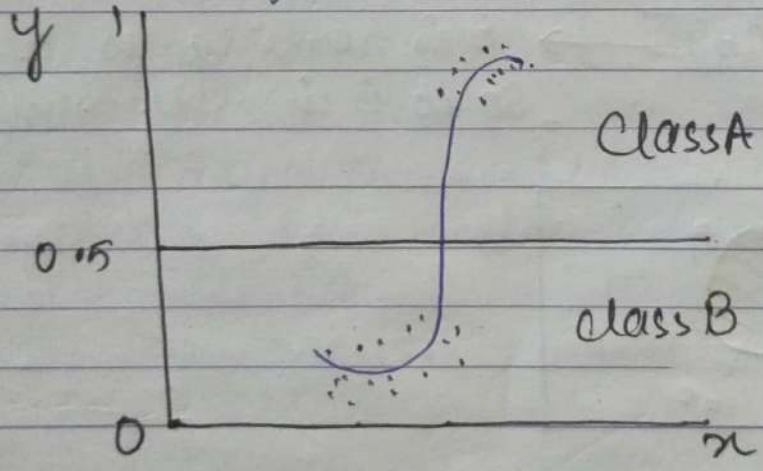
Here,

$f(x)$ = output between the 0 and 1 values

x = input to the function

e = Base of natural logrithm

vii) When we provide the input values to the function, it gives the S-curve as follows.



viii) It uses the concept of threshold level values above the threshold level, values are rounded up to 1 and values below the threshold level are rounded up to 0

ix) There are three types of logistic Regression:-

- Binary (0/1, Pass/fail)
- Multi (cats, dogs, lions)
- Ordinal (low, medium, high)

Bayes Theorem :-

$$\text{Probability} = \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcomes.}}$$

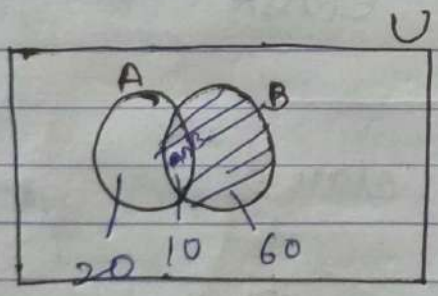
Conditional Probability :- In conditional probability, we find the probability of one event when the other given event has already occur.

for example :- find probability of A, given that event B has already occur.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

→ B is serving as a condition so it is in denominator

ex :-



$$P(A \cap B) = \frac{10}{100} = 0.1$$

$$P(B) = \frac{60}{100} = 0.6$$

$$P(A/B) = \frac{0.1}{0.6} = \frac{1}{6}$$

When there is no condition applied then sample space will be whole universe.

When we apply condition sample space get

reduce and a new sample space created, with the help of it we find (probability) conditional probability of event

Bayes Theorem:- Bayes Theorem is also widely used in Machine Learning where we need to predict classes precisely and accurately. An important concept of Bayes Theorem name Bayesian method is used to calculate conditional probability in ML application that includes classification task. further a simplified version of Bayes theorem is also used to reduce computation time and average cost of the projects.

Bayes Theorem is also known as Bayes rule or Bayes law. Bayes Theorem helps to determine the probability of an event with random knowledge. It is used to calculate the probability of occuring one event by with other one already occur. It is a best method to relate the conditional probability and marginal probability. Bayes Theorem helps to contribute more accurate results. Bayes Theorem is used to estimate the precision of value and provides a method for calculating the condition-al probability. However it is hypothetically a simple calculation but it is used to easily calculate the conditional probability of events where intution often fails.

Bayes Theorem :-

$$\text{Probability} = \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcomes.}}$$

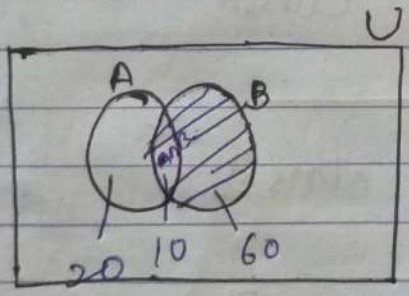
Conditional Probability :- In conditional probability we find the probability of one event when the other given event has already occur.

for example :- find probability of A, given that event B has already occur.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

→ B is serving as a condition so it is in denominator

ex :-



$$P(A \cap B) = \frac{10}{100} = 0.1$$

$$P(B) = \frac{60}{100} = 0.6$$

$$P(A/B) = \frac{0.1}{0.6} = \frac{1}{6}$$

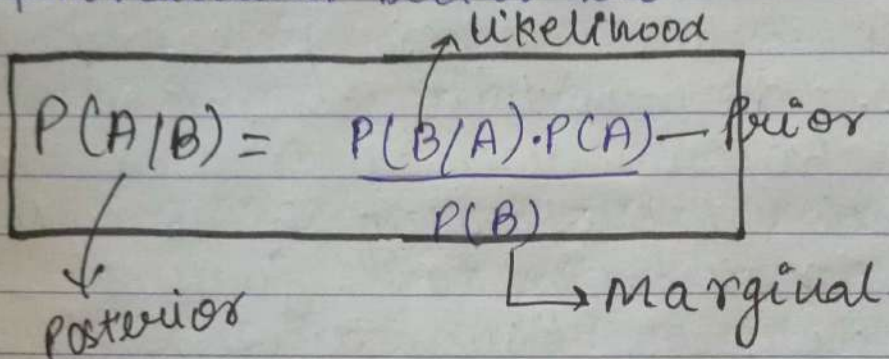
When there is no condition applied then sample space will be whole universe.
When we apply condition sample space get

reduce and a new sample space created, with the help of it we find (probability) conditional probability of event.

Bayes Theorem:- Bayes Theorem is also widely used in machine learning where we need to predict classes precisely and accurately. An important concept of Bayes Theorem name Bayesian method is used to calculate conditional probability in ML application that includes classification task. Further a simplified version of Bayes theorem is also used to reduce computation time and average cost of the projects.

Bayes Theorem is also known as Bayes rule or Bayes law. Bayes Theorem helps to determine the probability of an event with random knowledge. It is used to calculate the probability of occuring one event by while other one already occur. It is a best method to relate the conditional probability and marginal probability. Bayes Theorem helps to contribute more accurate results. Bayes Theorem is used to estimate the precision of value and provides a method for calculating the condition-al probability. However it is hypothetically a simple calculation but it is used to easily calculate the conditional probability of events where intution often fails.

Some of the data scientist assumes that Bayes theorem is used in financial industries but it is not like that. Other than financial industries. Bayes theorem is also extensively applied in health and medical, Research and survey, Aeronautical sector etc.



$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A/B)\,P(B) = P(A \cap B) \quad -(I)$$

$$P(B/A)\,P(A) = P(A \cap B) \quad -(II)$$

$$P(A/B)\,P(B) = P(B/A)\,P(A) \quad \{\text{from eq } (I) \text{ \& } (II)\}$$

$$\boxed{P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}}$$

Note :- Probability of (A/B) is called posterior we need to calculate. It is defined as updated probability after considering the evidence.

- Probability of (B/A) is called a likelihood. It is the probability of evidence when hypothesis is true.

- Probability of (A) is called prior probability, it is the probability of hypothesis before considering the evidence.

Probability of (B) is called marginal probability, it is defined as the probability of evidence under any consideration.

Hence Baye's theorem can be written as -

$$\text{Posterior} = \frac{\text{likelihood} * \text{Prior}}{\text{evidence}}$$

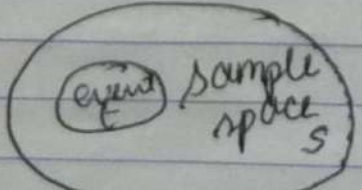
Terminology used in Baye's Theorem :-

Experiment :- planned operation carried out under controlled conditions such as tossing a coin, rolling a dice, drawing a card etc.

Sample space :- During an experiment what we get as a result is called a possible outcomes and the set of all possible outcomes of an event is known as sample space.

for eg :- In rolling a dice, S will be = {1, 2, 3, 4, 5, 6}

Event :- Event is defined a subset of sample space in an experiment further it is also called set of outcomes.



Random Variable:- It is a real value function which helps mapping b/w sample space and a real line of an experiment. A random variable is taken on some random values and each value having some probability. However it is neither random nor a variable but it behaves as a function which can either be discrete, continuous or combination of both.

Exhaustive Event:- As per the name suggests, a set of events where at least one event occurs at a time called exhaustive event of an experiment.

Independent Event:- Two events are said to be independent when occurence of one event does not affect the occurrence of another event. In simple words, we can say that the probability of outcome doesnot depend on another event.

Conditional Probability:-

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Marginal Probability:- It is defined as the probability of an event A occuring independent of any other event B.

Naive Bayes Classifier :-

- Naive Bayes algorithm is a supervised learning Algorithm which is Based on Bayes Theorem and used for solving classification problem
- It is mainly used in text classification that includes a high dimensional training data set.
- Naive Bayes classifier is one of the simple and most effective classification algorithm which helps in building the fast ML models that can may quick prediction.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naive Bayes algorithm are spam filteration, Sentimental Analysis and classify article.
- Naive Bayes algorithm is comprised of two words Naive and Bayes which can be describe as

→ Naive :- It is called Naive because it assumes that the occurrence of the certain feature is independent of the occurrence of other features. Such as if the fruit is identify on the basis of color, shape & taste then red, spherical & sweet fruit is recognized as an Apple. Hence each feature individually contribute to identify that it is an apple without depending on each other

→ Bayes :- It is called Bayes because it depends on the principle of Bayes theorem.

Working of Naive Bayes Classifier:-

(Working of Naive Bayes classifier) It can be understood with the help of example given below.
Suppose we have a data set of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather condition so to solve this problem we need to follow the below steps:-

Step①:- Convert the given dataset into frequency table.

Step②- Generate likelihood table by the finding the probabilities of given features.

Step③- Now use Bayes theorem calculate the posterior probability.

Problem:- If the weather is sunny then the player should play or not.

Dataset:-

	Outlook	Play
0.	Rainy	Yes
1.	Sunny	Yes
2.	Overcast	Yes
3.	Overcast	Yes
4.	Sunny	No
5.	Rainy	Yes
6.	Sunny	Yes
7.	Overcast	Yes

8.	Rainy	NO
9.	Sunny	NO
10.	Sunny	Yes
11.	Rainy	NO
12.	Overcast	Yes
13.	Overcast	Yes

Frequency table

Weather	Yes	NO
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	4

Likelihood table

Weather	NO	Yes
Overcast	0	5=5/14
Rainy	2	2 =4/14
Sunny	2	3 =5/14
	4/14= 0.29	10/14=0.71

$$P(Yes/Sunny) = \frac{P(Sunny/Yes) \cdot P(Yes)}{P(Sunny)}$$

$$P(Sunny/Yes) = \frac{3}{10}$$

$$P(Yes) = \frac{10}{14}$$

$$P(sunny) = 5/14$$

$$\frac{3}{10} \cdot \frac{10}{14} \times \frac{14}{5}$$

$$= \frac{30}{50} = \frac{3}{5} = \boxed{0.6}$$

$$P(NO/Sunny) = \frac{P(Sunny/NO) \cdot P(NO)}{P(sunny)} = \frac{\frac{2}{4} \cdot \frac{4}{14}}{5/14} = \frac{2}{5}$$

$$\boxed{=.4}$$

$$P(Yes/Rainy) = \frac{P(Rainy/Yes) \cdot P(Yes)}{P(Rainy)}$$

$$= \frac{2}{10} \times \frac{10}{14} \times \frac{14}{4}$$

$$= \boxed{0.5}$$

$$P(No/Rainy) = \frac{P(Rainy/No) \cdot P(No)}{P(Rainy)}$$

$$= \frac{3}{4} \times \frac{4}{14} \times \frac{14}{4} = \boxed{0.5}$$

$$P(Yes/Overcast) = \frac{P(Overcast/Yes) \cdot P(Yes)}{P(Overcast)}$$

$$= \frac{5}{10} \times \frac{10}{14} \times \frac{14}{5} = 1$$

$$P(No/Overcast) = \frac{P(Overcast/No) \cdot P(No)}{P(Overcast)}$$

$$= \frac{0 \times \frac{4}{14} \times \cancel{5}}{\cancel{5}}$$

$$= 0$$

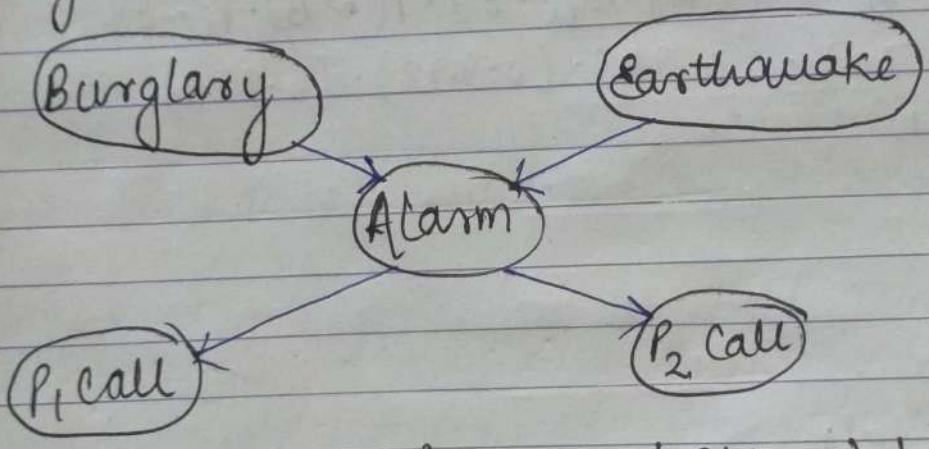
Bayesian Belief Network:-

Note:- Bayesian Belief network is based on joint probability distribution and conditional probability

It is convenient for representing probablistic relation between multiple events.

If we have variables x_1, x_2, x_3 and so on upto x_n then the probabilities of a different combination of $x_1, x_2, x_3 \ldots$ upto x_n are known as joint probablity distribution

Ques. Calculate the probability that alarm has sounded, but there is neither a Burglary nor an earthquake occured an Neighbour1, Neighbour 2 both call the landlord



Q. $P(P_1, P_2, A, \sim B, \sim E)$

$P(B=T) = 0.001$

$P(B=F) = 0.999$

$P(E=T) = 0.002$

$P(E=F) = 0.998$

A	$P(P_2=T)$	$P(P_2=F)$
T	0.70	0.30
F	0.10	0.90

B	E	$P(A=T)$	$P(A=F)$
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	$P(P_1=T)$	$P(P_1=F)$
T	0.90	0.10
F	0.05	0.95

Sol^{n} $P(P_1|A) \cdot P(P_2|A) \cdot P(A|\sim B, \sim E) \cdot (P \sim B) \cdot P(\sim E)$

$$P(P_x|A) = \frac{P(A|P_i) \cdot P(P_i)}{P(A)}$$

$= (0.90) \cdot (0.70) \cdot (0.001) \cdot (0.999) \cdot (0.998)$
$= 0.000628$

$P(P_1|A) \cdot P(P_2|A) \cdot P(A|B, \sim E) \cdot P(\sim B) \cdot P(\sim E)$
$= (0.90) \cdot (0.70) \cdot (0.94) \cdot (0.999) \cdot (0.998)$
$= 0.5904$

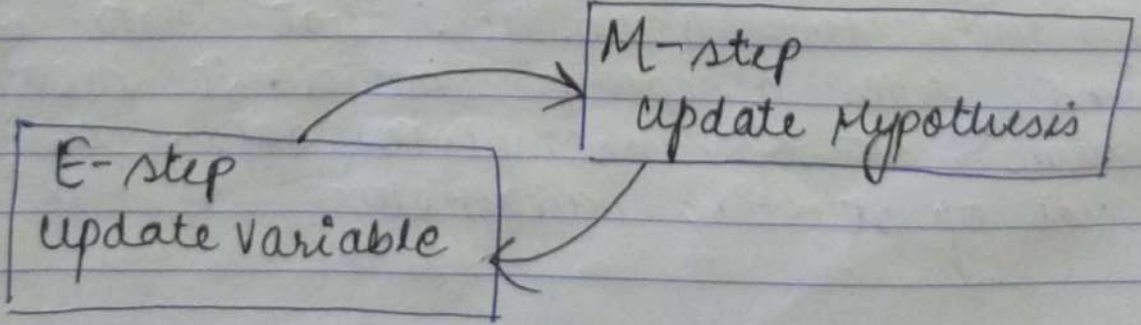
EM Algorithm:-

(Expectation maximization Algorithm)

The EM algorithm is one of the most commonly used term in Machine learning to obtain maximum likelihood estimates of variables that are sometimes observable and sometimes not.
It has various real world application in statistics including the mode of the posterior Marginal distribution of parameter in Machine learning and Data Mining application.
The expectation - maximization (EM) algorithm is defined as the combination of various

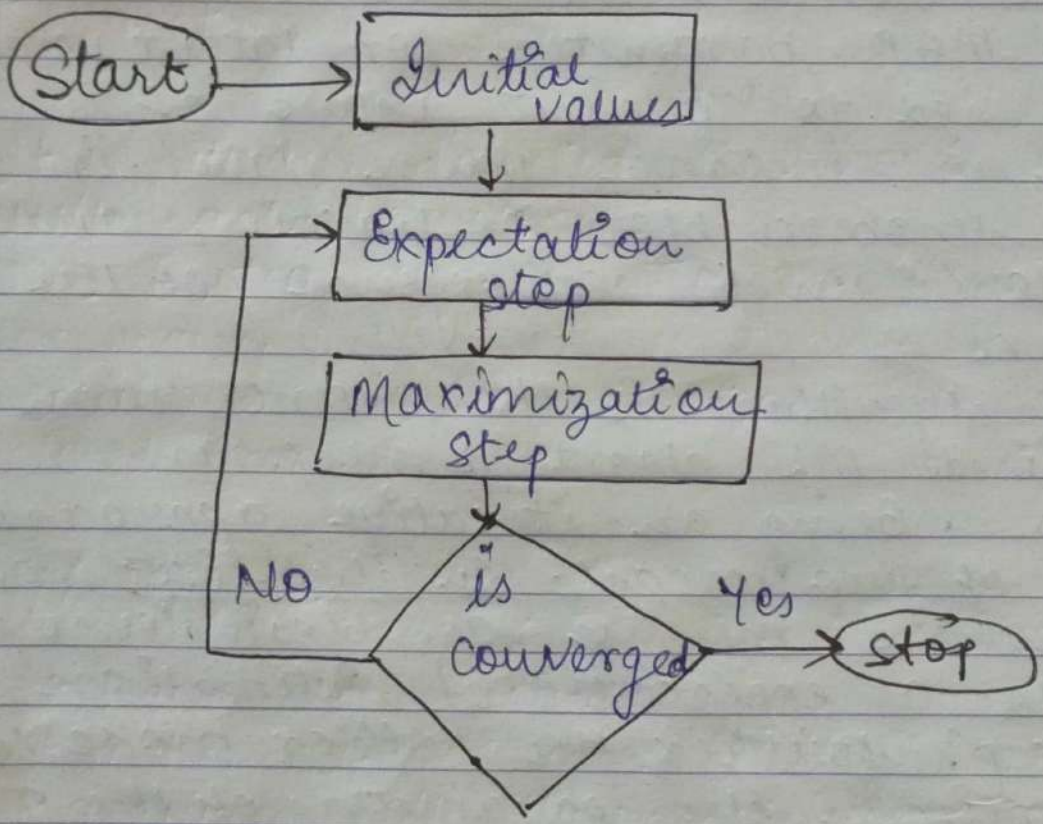
Unsupervised ML algorithm which is used to determine the local maximum likelihood estimates (MLE) or maximum posterior estimate for unobservable variables in statistical model. further it is a technique to find maximum likelihood estimation when the latent variables are present it is known as the latent variable model determine MLE and MAP parameters for latent variable. It is used to predict values of parameters in instance where data is missing or unobservable for learning and this is done until convergence of the values occurs

The EM algorithm is the combination of Unsupervised ML algorithm such as K-means. Being an iterative approach it consist of two model. In the first mode we estimate the missing variables hence it is referred to as expectation/ Estimation step (E step) further the other mode is used to optimize the parameters of the model so that it can explain the data more clearly The second mode is known as the maximization step or (M step)



Steps in EM algorithm:-

The EM algorithm is completed mainly in four step which include initialization step, expectation step, Maximization step and convergence step. There steps are explained as follows:-



Step 1:- The very first step is to initiate the parameter values. Further the system is provided with incomplete observed data with the assumption that data is obtained from a specific model. It means it gives set of incomplete data, consider a set of starting parameter.

Step2:- This step is known as Expectation or E step which is used to estimate or guess the values of the missing or incomplete data using the observed data of the dataset

Step3:- This step is known as Maximization or M step where we use complete data obtained from second step to update the parameter values. further M step primarily updates the hypothesis.

Step4:- The last step is to check if the values of latent variables (variables that are not directly observable and are actually infer from the values of the other observed variables) are converging or not. If it get yes then stop the process else repeat the process from step 2 until the convergence occurs.

Uses of EM algorithm:-

1. It is used to fill the missing data in the sample. 2. It use the bases of VSL cluster.
3. It can be used for the purpose of estimating the parameter of hidden Marco Model (HMM)
4. It is used for discovering the values of latent varibles.

Note:- About the convergence in EM algorithm:- convergence is defined as the specific situation in probability based on intuition. for ex:- If there

are two random variables that have very less difference in their probability, then they are known as converged. In the other words, whenever the values of given variables are matched with each other, it is called Convergence.

Advantages of EM algorithm:-

- It is always directed that likelihood will increase with each iteration.
- The E step and M step are often pretty easy for many problems in terms of implementation.
- solutions to the M steps often exist in the closed form.

Disadvantages of EM algorithm:-

- It has slow convergence.
- It makes convergence to the local optima only.
- It requires both the probabilities forward & backward.

Support Vector Machine:-

Support Vector Machine is relatively simple supervised ML Algorithm used for classification and Regression. It is more prefer for classification but it is sometime very useful for regression as well.

Basically SVM finds hyperplane that creates a boundary between the types of data. In 2 dimensional surface this hyperplane is nothing but a line. In SVM, we plot each data item in the data set in an n-dimension -l space, where n is the no. of feature in the data.

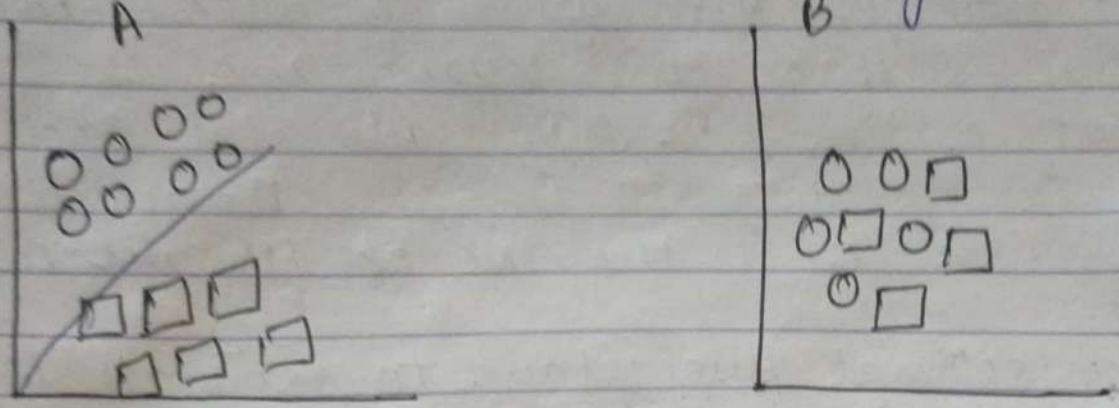
Find the optimal Hyperplane to separate the data. By this you must have understood that inherently SVM can only perform binary classification (i.e. choose b/w two classes). However, there are various techniq -ue to use for multiclass problem. SVM on multiclass problem, we can create a binary classifier for each class of data. The two result of each classifier will be
- The datapoint belongs to that class.
- The datpoint does not belongs to that class.

for example, in a class of fruits to perform for multiclass classification we can create a binary classifier for each fruit. So for example we can say the 'mango' class will be a binary classifier to predict if it is a 'Mango' or it is not a 'Mango'.

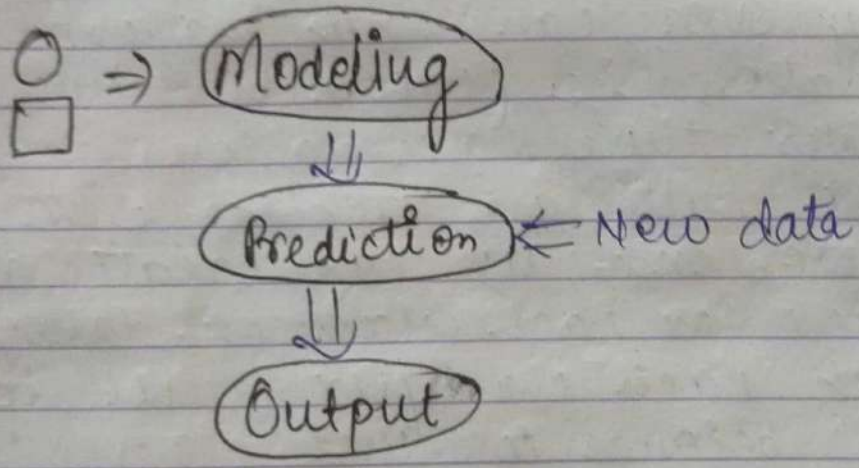
The classifier with the highest score is choosen at the output of SVM.

SVM for complex (Non linearly separable) works very well without any modification for linearly separable data. Linearly separable data is any data that can be plotted in a graph and can be

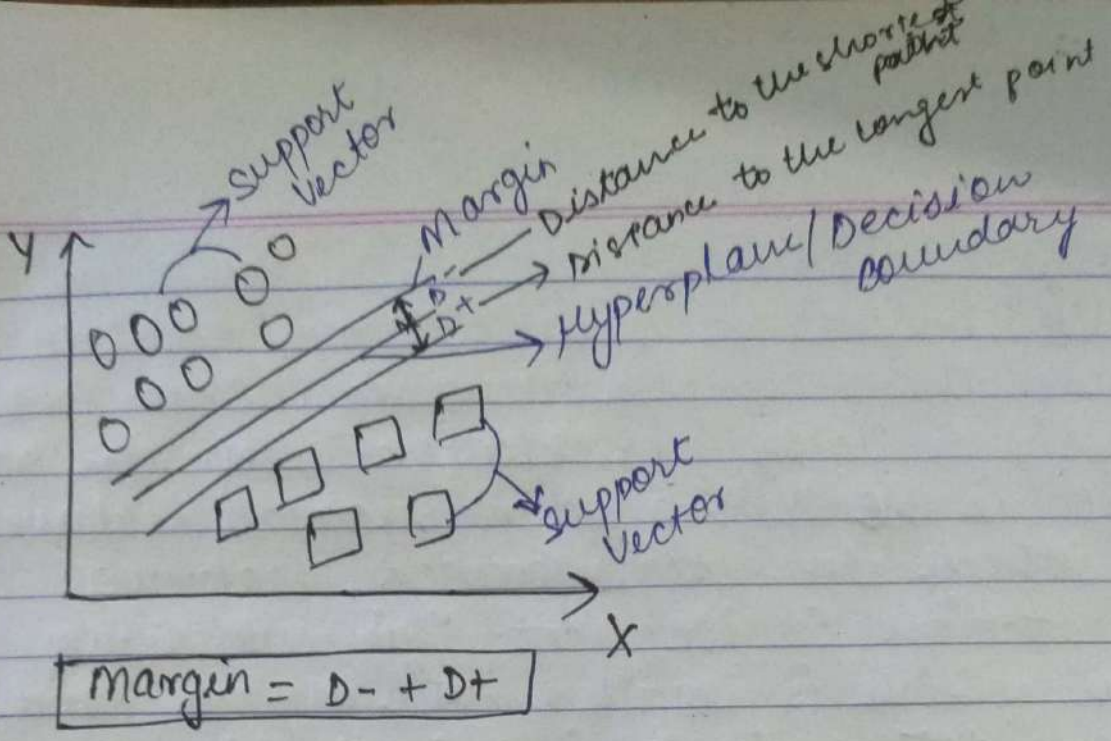
separated in two classes using a straight line.



Ex:-



We use Kernelised SVM for Non linearly separable data. We have some non linearly separable data in 1-Dimension. We can transform this data into 2-Dimension and the data will become linearly separable in 2-Dimensions. This is done by mapping each one dimension data point to corresponding 2-Dimension order pair. So for any non linearly separable data in any dimension we can just map the data to higher dimension and then make it linearly separable.



- Margin will play important role in deciding which hyperplane exist or which will not exist.

SVM Kernel:- A Kernel is nothing but a measure of similarity between datapoints. The Kernel function in Kernalised SVM. Tells that given two data points in the original feature space, what the similarity is between the points in the newly transform feature space. There are various kernel function available but two are very popular.

1. **Radial Basis function Kernel**
2. Polynomial Kernel

A very interesting fact is that SVM does not actually have to perform this actual transformation on the datapoints to the new dimensional feature space. This is called the kernel trick

Kernel → 1D ⇒ Kernel ⇒ 2D
lower ⇒ Kernel ⇒ higher

Kernel trick:- Internally the kernelised SVM can compute these complex transformations just in terms of similarly calculations between pairs of points in the higher dimensional feature space. where the transformed feature representation is implicit. This similarity function which is mathematically a kind of complex dot product is actually the Kernel of Kernelised SVM.

Kernel function:- Kernel function is a method used to take data as input and transform it into the required form of processing data. Kernel function generally transform training set of data so that a non linear decision surface is able to transform in a linear equation in a higher number of dimension spaces.

Standard Kernel function:-
(Types of Kernel)

① **Gaussian Kernel:-** It is used to perform transformation when there is no prior knowledge about data.

② **Gaussian Kernel Radial Basis function (RBF):-** RBF is a default Kernel used in support vector machine It is used to transform

feature space in an exponentially decaying function of distance between the vectors. and the original input space.

③ Sigmoid Kernel:- This function is equivalent to a two layer perception model of the neural network which is used as an activation function for artificial neurons.

④ Polynomial Kernel:- The polynomial kernel takes an additional parameter 'degree' that controls the model complexity and computational cost of the transformation

⑤ Linear Kernel:- Linear Kernel is used when the data is separated using a single line. It is one of the most common Kernel to be used. It is mainly used when there are large number of features in a particular data set.

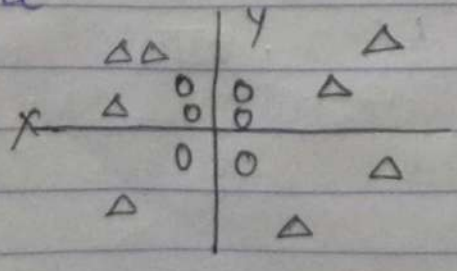
Hyperplane (Decision Boundary/Surface):-
(Best line decision boundary)
There can be multiple lines or decision boundary to seggregrate the classes in n dimension space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of hyperplane depends on the feature present in the data set which means if there are two features then hyperplane will be straight line. If there are three features, then hyperplane will be 2 dimensional plane.
We always create hyperplane that has maximum margin, which means maximum distance between the data points.

Margin:- Distance between the vectors and hyperplane is called as Margin. And the goal of SVM is to maximize this margin. The hyperplane with the maximum margin is called optimum margin.

Support Vector:- The data points or vector that are closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane Hence called as Support Vector.

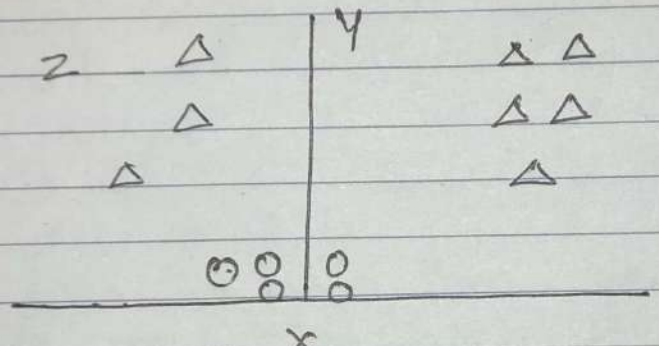
Non Linear SVM:- If the data is linearly arranged then when we can separate it by using the straight line but for non linear data we cannot draw a single line or straight line



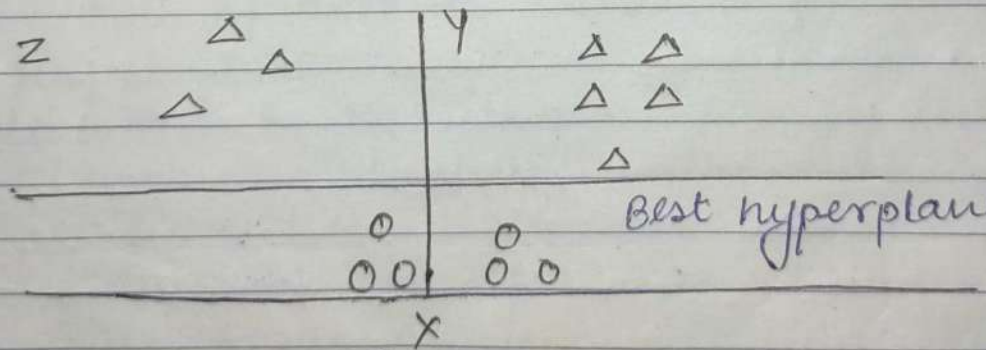
Consider the above image, to separate these points we need to add one more dimension. for linear data we have use 2 dimension x and y so for non linear data we will add a 3rd dimension z. it can be calculated as

$$z = x^2 + y^2$$

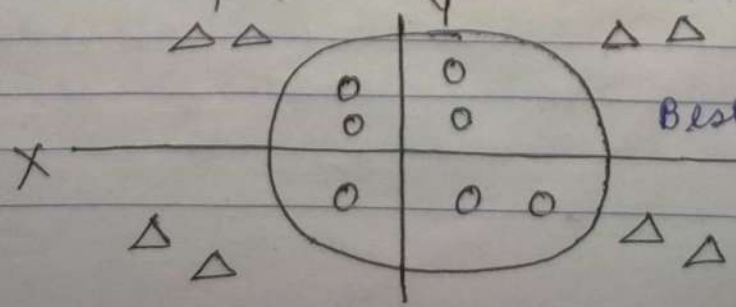
By adding the 3rd dimension, the sample space will become as the below image



So now SVM will divide the datasets into classes in the following way



Since we are in 3-D space, hence it is looking like a plane parallel to the x-axis. If we convert it into 2D space with z=1, then it will become



Hence, we get a circumference of radius 1 in case of Non linear data.

UNIT-3

Decision Tree

Decision Tree :- A tree has many analogies in real life and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making.

Classification is a two step process, learning step and prediction step in machine learning. In learning step, the model is developed based on the given data set. In the prediction step the model is used to predict the response for given data.

Decision tree algorithm belongs to the family of SL algorithm. Unlike other SL algorithms, the decision tree algorithm can be used for Regression and classification problem too.

The goal of using the decision tree is to create a training model that can predict the class or value of the target variable by learning simple decision rule inferred from prior data (trainning data). The decision tree is one of the easiest and popular classification algorithm to understand and interrupt.

→ root = decision node
→ interior node = terminal node
leaf node

In decision trees for predicting a class level for a record we start from the root of the tree. We compare the values of the root attributes on the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Types of Decision Trees:-

Types of decision tree are based on the type of target variable we have. It can be of two types

① Categorical variable decision tree:- Decision tree which has a categorical target variable is called a categorical variable decision tree.

② Continuous variable decision tree:- Decision tree which has a continuous target variable is called continuous variable decision tree.

for ex:- We have a problem to predict whether a customer will pay his renewable premium with an insurance company (Yes/No)

Here we know that the income of customer is a significant variable but insurance company does not have income details for all customer. Now as we know this is an important variable, then we

can build a decision tree to predict customer income based on occupation, product and various other variable.

Terminologies related to decision tree:-

1. Root node
2. Splitting
3. Decision node
4. Leaf / Terminal node
5. Pruning
6. Branch / Sub tree
7. Parent and child node

- **Root node:-** Root node is a from where the decision tree start. It represents the entire data set which further get divide into two or more homogeneous (data)sets
- **Leaf node:-** Leaf nodes are the final node and the tree can not be seggregated further after getting the leaf node.
- **Splitting:-** Splitting is the process of dividing the decision node / Root node into sub nodes according to the given conditions.
- **Branch / subtree:-** A tree formed by the spilliting the tree.
- **Pruning:-** Pruning is a process of removing the unwanted branches from the tree.

• Parent/child node:- The root node of the tree is called the parent node and the other nodes are called child nodes.

Assumption while creating decision tree

(1) In the beginning, the whole training set is considered as the root.
(2) Feature values are preforced to be categorical. If the values are continuous then they are discretized prior to the building of the model.
(3) Records are distributed recursively on the basis of attribute values.
(4) Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

Decision trees follow sum of product representation (SOP). The sum of product is also known as disjunctive normal form.
For a class every branch from the root of the tree to a leaf node having the same class as in conjunction (product of values). Different branches ending in that class from a disjunction
The primary challenge in decision tree implementation is to identify which attribute do we need to consider as the root node and each level handling this is

to known as attribute selection we have different attrib-ute selections measures to identify the attribute which can be considered as the root node at each level.

Steps Used for Making Decision tree :-

step ① :- Begin the tree with the root node, which contains the complete dataset

step ② :- find the best attribute in the dataset using attribute selection measure.

step ③ :- Divide the root node into subsets that contain the possible values for the best attribute

step ④ :- Generate the decision tree node which contains the best attribute.

step ⑤ :- Recursively make new decision tree using the subsets of the dataset created in step 3. Continue this process until the stage reached where you cannot further classify the nodes and called final node as leaf node.

Attribute Selection Measure :- While implementing a decision tree. The main issue arises that how to select a best attribute for root node and for sub nodes. So to solve such problem there is a technique which is called attribute selection measure or ASM. By this measurement we can easily select best attribute for the node of the tree. There are two popular techniques for ASM which are :- 1. I.G 2. Gini Index

① **Information Gain :-** • Information gain is a measurement of changes in entropy after the segmentation of the dataset

based on an attribute.

- It calculates how much information of a feature provides us about a class.
- According to the value of IG we split the node and build the decision tree. A decision tree algorithm always tries to maximise the value of IG and a node / attribute having the highest IG is split first. It can be calculated using the below formula.

$$IG = Entropy(S) - [Weighted\ Avg * Entropy(each\ feature)]$$

Entropy :- Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data.
It can be calculated as.

$$Entropy(S) = -P(yes)\log_2 P(yes) - P(no)\log_2 P(no)$$

Ex:-

Age	Competition	Type	Profit = TV.
old	Yes	S/W	Down
old	NO	S/W	Down
old	NO	H/W	Down
mid	Yes	S/W	Down
mid	Yes	H/W	Down
mid	NO	H/W	UP
mid	NO	S/W	UP
new	Yes	S/W	UP
new	NO	H/W	UP
new	NO	S/W	UP

(for IG(TV)
P = No. of down
N = No. of up

① $IG = -\frac{P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \left(\frac{N}{P+N}\right)\log_2\left(\frac{N}{P+N}\right)$

② Entropy = $\sum_{i=1} \frac{P_i+N_i}{P+N} I(P_i N_i)$

③ Gain = IG(TV) − Entropy (undertaking attribute)

$IG(TV) = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$

$= -\left[\frac{5}{10}\log_2\left(\frac{5}{10}\right) + \frac{5}{10}\log_2\left(\frac{5}{10}\right)\right]$

$= -\left[0.5\log_2 2^{-1} + 0.5\log_2 2^{-1}\right]$

$= -\left[-0.5 \times 1 + (-0.5 \times 1)\right]$

$= -[-1]$

$= 1$

(Undertaking Attribute) Age

	Down	UP
old	3	0
Mid	2	2
New	0	3

Entropy = IG × Prob. (Age)

$IG(old) = -\left[\frac{3}{3}\log_2\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] = 0$

Entropy(old) = IG(old) * Probability of old

$= 0 * \frac{3}{10}$

$= 0$

$IG(mid) = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] = 1$

Entropy = IG(mid) * Probability of old

$= 1 * \frac{4}{10} = 0.4$

$IG(new) = -\left[\frac{0}{3}\log\left(\frac{0}{3}\right) + \frac{3}{3}\log_2\left(\frac{0}{3}\right)\right] = 0$

Entropy $= 0 * \frac{3}{10}$

$= 0$

Entropy (Age) $= 0 + 0.4 + 0$

$= 0.4$

Gain $= IG(TV) - \text{Entropy (Age)}$

$= 1 - 0.4$

$= 0.6$

competition	down	up
yes	3	1
NO	2	4

$IG(yes) = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

$= -[0.75 \times (-0.415) + 0.25(-2)]$

$= -[-0.31125 - 0.5]$

$= 0.81125$

Entropy $= 0.81125 \times \frac{4}{10}$

$= 3.245/10$

$= 0.3245$

$IG(NO) = -\left[\frac{2}{6}\log_2\left(\frac{2}{6}\right) + \frac{4}{6}\log_2\left(\frac{4}{6}\right)\right]$

$-[0.333\log_2\left(\frac{2}{6}\right) + 0.667\log_2\left(\frac{4}{6}\right)]$

$= -[-0.1590 - 0.1173]$

$= -[-0.2763] = 0.2763$

Entropy $= 0.2763 \times \frac{6}{10} = 0.16578$

Entropy (Competition) = 0.3245 + 0.16578 = .49028

Gain = 1 − 0.49028 = 1 − 0.4 = 0.6 (0.124)

≠ 0.6

Type	Down	UP
H/W	2	2
S/W	3	3

$$IG(H/W) = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$$

$= 1 \Rightarrow$ Entropy $= 1 \times \frac{4}{10} = 0.4$

$$IG(S/W) = -\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$$

$= 1$

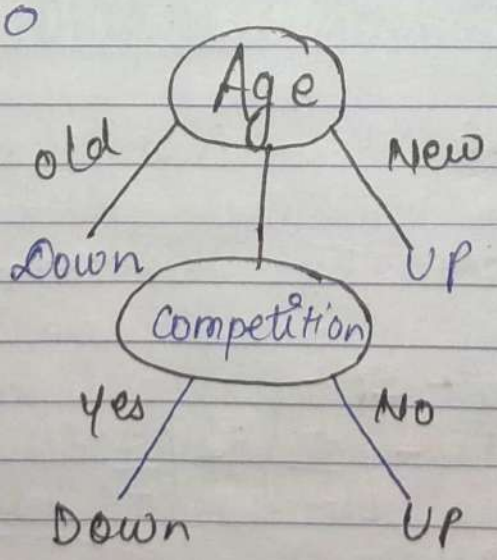
Entropy $= 1 \times \frac{6}{10} = 0.6$

Entropy (Type) $= \frac{4}{10}(0.4 + 0.6)$

$= 1$

Gain $= 1 - 1$

$= 0$



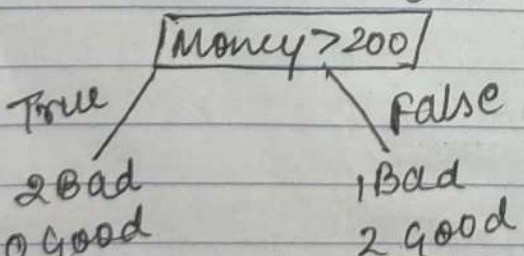
Gini Index:- (Tree splitting criteria).

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) Algorithm in the CART algorithm.
- An attribute with low Gini index should be preferred as compare to the high gini index
- Gini index can be calculated using below formula.

$$\text{Gini index} = 1 - \sum_i p_i^2$$

Eg:-

Day	Invested Money	Profit status
1	100	Bad
2	200	Good
3	250	Bad
4	150	Good
5	300	Bad



Sol: Gini index $= 1 - \Sigma p^2$

(i) $= 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right]$

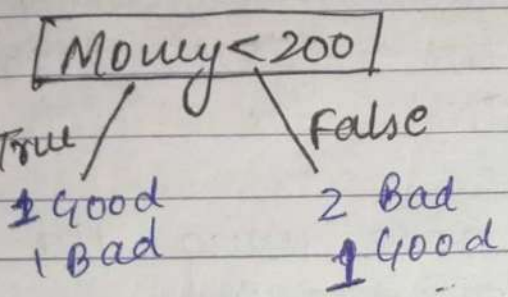
$= 0$

(ii) $= 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{4}{3}\right)^2\right] = 1 - \left[\frac{1}{9} + \frac{2}{9}\right] = 1 - \frac{5}{9}$

$= \frac{4}{9} =$

Final Gini index $= \frac{2}{5} \times 0 + \frac{3}{5} \times \frac{4}{9}$

$= \frac{12}{45} = \frac{4}{15} = 0.2667$



- Gini index will be check for all possible splits at a particular node. Minimum gini index will be selected as tree spliting criteria at a particular node
- In the case of gini index, the best case for gini index will alwdys lie in between 0 to 1. So gini index is a default splitting criteria for making decision tree.

ID3 Algorithm :- (Iterative Dichotomiser)

ID3 stands for iterative dichotomiser and is named such because the algorithm iteratively (repeteadly) divides features into two or more groups at each level.

ID3 uses a top down greedy approach to build a decision tree. Top down approach means that we start building the tree from the top and the greedy approach means that at each iteration we select the best feature at the present moment to create a node.

Steps for ID3 Algorithm:-

Step ①:- Calculation the I.G of each feature

Step ②:- Considering that all rows don't belong to the same class, split the dataset into subsets using the feature for which the I.G is maximum

Step ③:- Make a decision tree node using the feature with the maximum gain.

Step ④:- If all the rows belong to the same class, make the current node as a leaf node with the class as its label.

Step ⑤:- Repeat for the remaining attribute until we run out of all the features (attribute) or the decision tree has all leaf node.

Ex:-

Draw decision tree by using ID3 algorithm for the given dataset

T.V

Day	Outlook	temp	wind	humidity	Play Tennis
D1	Sunny	Hot	weak	High	No
D2	Sunny	Hot	strong	High	No
D3	Overcast	Hot	weak	High	Yes
D4	Rain	mild	weak	High	Yes
D5	Rain	cool	weak	Normal	Yes
D6	Rain	cool	Strong	Normal	No
D7	Overcast	cool	strong	Normal	Yes.
D8	Sunny	mild	Weak	High	No
D9	Sunny	cool	Weak	Normal	Yes
D10	Rain	mild	Weak	Normal	Yes
D11	Sunny	mild	Strong	Normal	Yes.
D12	Overcast	mild	strong	High	Yes
D13	Overcast	Hot	weak	Normal	Yes
D14	Rain	mild	strong	High	No

Attribute : Outlook

Value (Outlook) = Sunny, Overcast, Rain

$S[9+, 5-]$

$$\text{Entropy}(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$S_{Sunny}[2+, 3-]$

$$\text{Entropy}(S_{Sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$S_{Overcast}[4+, 0-]$

$$\text{Entropy}(S_{Overcast}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$S_{Rain}[3+, 2-]$

$$\text{Entropy}(S_{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

$$\text{Gain}(S, \text{Outlook}) = \text{Entropy}(S) - \sum_{V \in (\text{Sunny, Overcast, Rain})} \frac{|S_V|}{S}\,\text{Entropy}(S_V)$$

$$= \text{Entropy}(S) - \frac{5}{14}\text{Entropy}(S_{Sunny}) - \frac{4}{14}\text{Entropy}(S_{Overcast}) - \frac{5}{14}\text{Entropy}(S_{Rain})$$

$$= 0.94 - \frac{5}{14}\times 0.971 - \frac{4}{14}\times 0 - \frac{5}{14}\times 0.971 = 0.2464$$

Attribute : temp

Value (temp) = Hot, Cold, mild

$S[9+, 5-] = 0.94$

$S_{Hot}[2+, 2-]$

$$\text{Entropy}(S_{hot}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$S_{mild}[4+, 2-]$

$$\text{Entropy}(S_{mild}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = .3866 + .526 = .9132$$

0.1516 (H)
0.0478 (W)

S_{cool} [3+, 1−]

Entropy(S_{cool}) $= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = .315 + .5 = .815$

Gain $= 0.94 - \frac{4}{14} \times 1 - \frac{6}{14} \times .9132 - \frac{4}{14} \times .815$

$= 0.0289$

Attribute (wind) :-

value = (weak, strong)
(wind)

S[9+, 5−] = 0.94

$S_{(strong)}$ [3+, 3−]

Entropy(S_{strong}) $= -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} =$

$=$

$=$

&

S(weak) [6+, 2−]

Entropy(S_{weak}) $= -\frac{6}{8}\log_2\frac{6}{8} - \frac{6}{8}\log_2\frac{6}{8}$

$=$

$=$

Gain $= 0.94 - \frac{6}{14} \times \quad - \frac{8}{14} \times$

$= 0.0478$

Attribute (humidity) :-

value (High, Normal)
humidity

$S[9+, 5-] = 0.94$

$S_{(High)}[3+, 4-] = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$

=

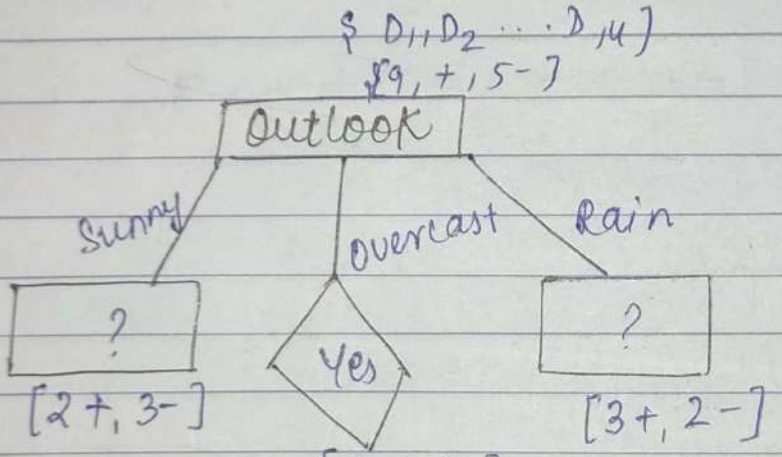
=

$S_{normal}[6+, 1-] = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$

=

=

$Gain = 0.94 - \frac{7}{14} \times \quad - \frac{7}{14} \times \quad = 0.1516$



Day	Temp	Humidity	Wind	play T
D_1	Hot	high	weak	NO
D2	Hot	high	strong	NO
D8	mild	high	weak	NO
D9	cool	Normal	weak	Yes
D11	mild	Normal	strong	Yes

Attribute (Temp) :-

Value (Temp) = Hot, mild, cool

$S_{sunny}[2+, 3-]$

$Entropy(S_{sunny}) = \frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$S_{hot}[0+, 2-]$

Entropy $(S_{hot}) = 0.0$

$S_{mild}[1+, 1-]$

Entropy $(S_{mild}) = 1$

$S_{cool}[1+, 0-]$

Entropy $(S_{cool}) = 0$

Gain = Entropy (S) $- \frac{2}{5}$ Entropy $(S_{hot}) - \frac{2}{5}$ Entropy (S_{mild}) $- \frac{1}{5}$ Entropy (S_{cool})

$= 0.97 - \frac{2}{5} \times 0 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0$

$= 0.570$

Attribute (Humidity)

$S_{sunny}[2+, 3-]$

Entropy $(S_{sunny}) = 0.97$

$S_{high}[0+, 3-]$

Entropy $= 0$

$S_{normal}[2+, 0-]$

Entropy $= 0$

Gain $= 0.97 - \frac{3}{5} \times 0 - \frac{3}{5} \times 0$

$= 0.97$

Attribute (wind)

Values (wind) = weak, strong

$S_{sunny}[2+, 3-]$

Entropy $(S_{sunny}) = 0.97$

S_{weak} [1+, 2-]

$= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = -\left(\frac{1}{3}\times(-1.584) + \frac{2}{3}\times(-0.584)\right)$

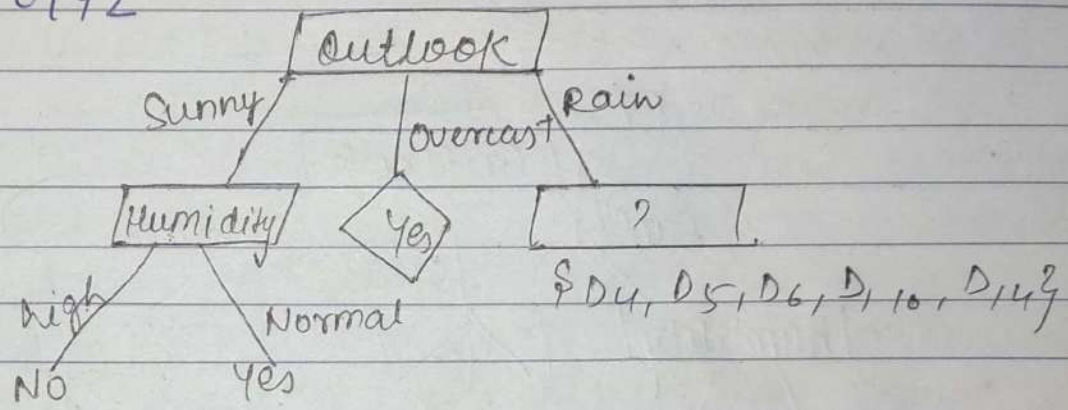
$= 0.9183$

S_{strong} [1+, 1-]

$= 1$

$Gain = 0.97 - \frac{3}{5}\times 0.9183 - \frac{2}{5}\times 1$

$= 0.0192$



Day	Temp	Wind	Play T
D4	mild	weak	yes
D5	cool	weak	yes
D6	cool	strong	NO
D10	mild	weak	yes
D14	mild	strong	NO

Attribute (Temp)

S_{Rain} (3+, 2-] = 0.971

[S_{Hot} (0+, 0-] = Entropy = 0] X

S_{mild} (2+ 1-] = Entropy = 0.9183

S_{cool} (1+, 1-] = Entropy = 1

Gain =

$= 0.97 - \frac{0}{5}\times 0 - \frac{3}{5}\times 0.918 - \frac{2}{5}\times 1$

$= 0.0192$

Attribute (wind)

Values (wind) = strong weak

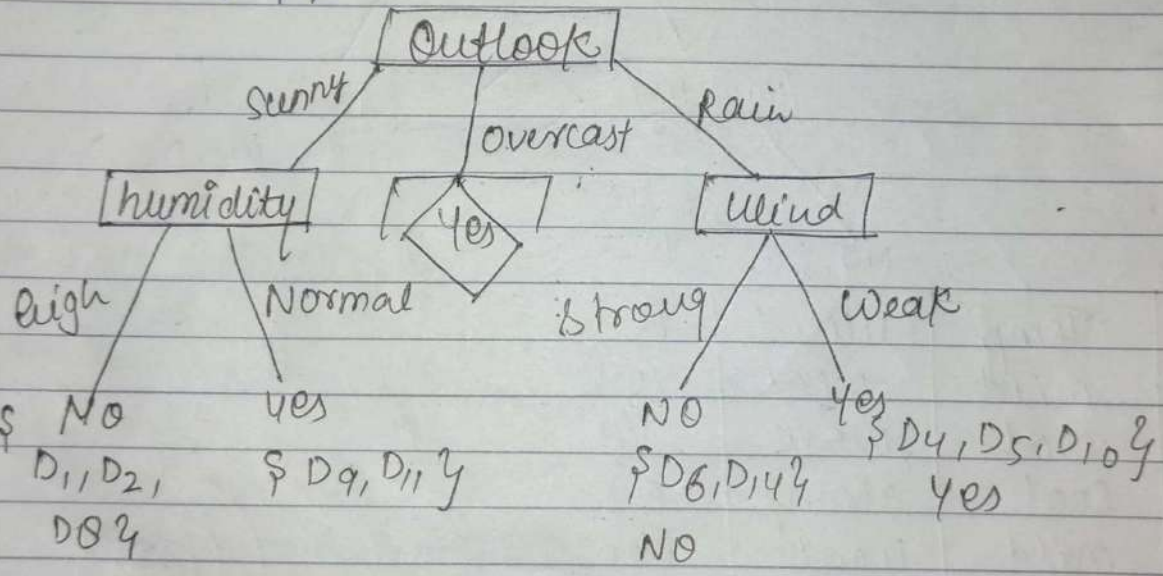
S_{rain} (3+, 2−) = 0.97

S_{strong} (0+, 2−) Entropy = 0

S_{weak} (3+, 0−) = Entropy = 0

$Gain = 0.97 - \frac{2}{5} \times 0 - \frac{3}{5} \times 0$

$= 0.97$



Pruning:- Pruning is the process of deleting the unnecessary node from a tree in order to get the optimal decision tree. A too large tree increases the risk of overfitting and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning data without reducing accuracy is known as pruning. There are mainly two types of tree pruning

① Cost complexity pruning
② Reduced Error pruning

Inductive Bias :-

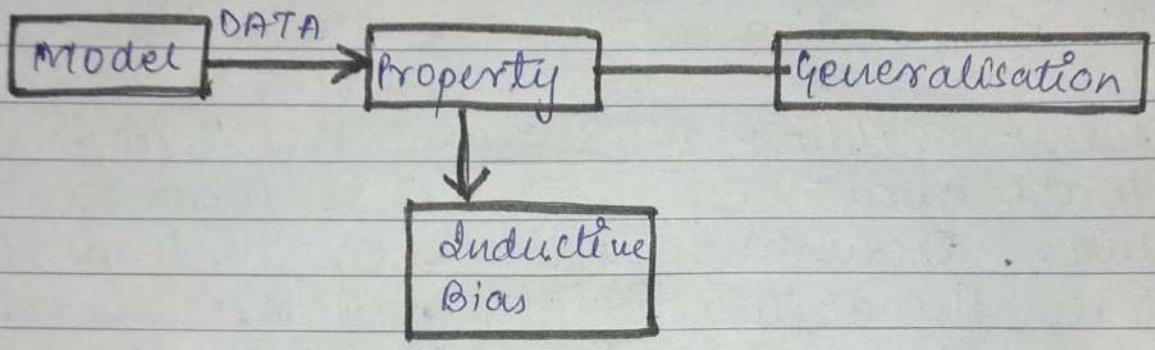
Every machine learning model require some type of architecture design and possibly some initial assumptions about the data we want to analysis. Generally every building block and every Belief that we make about the data is a form of inductive bias. Inductive Biases play an important role in the ability of ML models to generalize to the unseen data. A strong inductive bias can lead our model to coverage the global optimum. On the other hand, a weak inductive bias can cause the model to find only the local optima and be greatly affected by random changes in the initial stages. Inductive reasoning starts with an observation and leads to a possible generalisation hypothesis.

Valid observation may lead to different hypothesis and some of them can be false Before learning a model for given data and a learning algorithm, there are a few assumptions These assumptions are actually known as inductive bias. It is like the property of the algorithm.

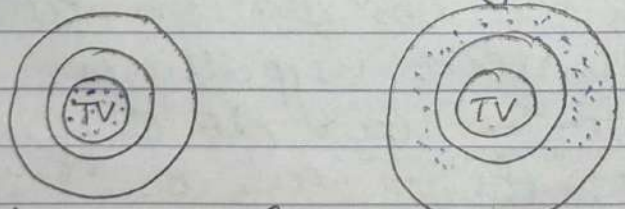
for example in the case of decision trees, the depth of the trees is the inductive bias. If the depth of the tree is too large, there will be too less generalization and wide

testing the model on a new example we might recall a particular example used to train the model.

If the depth of the tree is too low then there is too much generalization in the model



Decision tree is a method that uses inductive biases for assumption that an objective can be achieved by asking a series of binary questions. As a result, the decision boundary of the tree classifier becomes orthogonal.



Low inductive bias

Higher inductive bias

Inductive inference with Decision Tree:-

(Two types of inference are there)

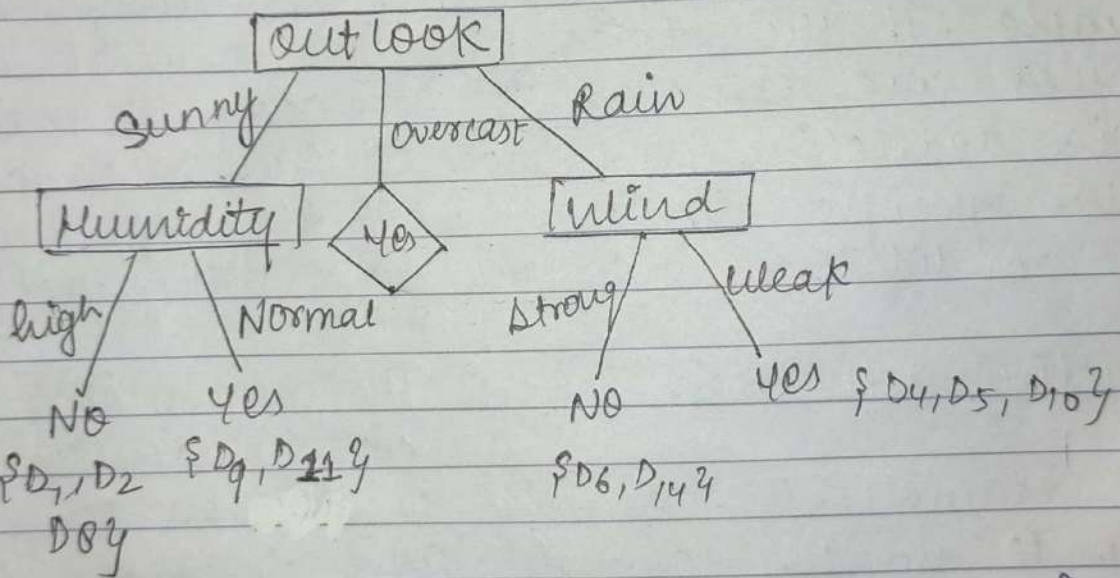
Decision tree is a method that uses inductive inference to approximate a target function, which will produce discrete values. It is widely used, robust too noisy data and

considered a practical method for learning disjunctive expressions. Inference are made when a person or machine goes beyond the available evidence to form the conclusion.
There are two types of inference:-

① Deductive Inference:- It moves from general to more specific conditions. A conclusion always follows the stated premises. In other words, if the premise are true then the conclusion is always valid. for example, if the stated premise is that all employees are expected to be good at maths, Then via deductive inference it can be conclude that the specific employee roll at XYZ corporation requires candidates who are good at maths.

② Inductive Inference:- It is the process of reaching a general conclusion from specific examples. The general conclusion should apply to unseen examples and is based on evidence from the past, which may or may not hold (evidence) true at this point or for a specific instance.
for example, if a certain pattern of medical parameters were found in the set of diabetic patient, then an inductive inference method could generalize that the same pattern of medical parameters will be found in all diabetic patients.

Decision tree are one of the most widely used classification technique that use inductive reference to approximate the target function. They classify instances by sorting them down the trees from root to leaf node. The normal nodes correspond to feature, the branch of the trees corresponds to the test condition on a feature and the leaf node corresponds to a class label.



Decision tree are easy to interrupt and visualize and they can easily capture non linear patterns. They require less data preprocessing from the user. for example - there is no need to normalize column they can also be used in feature engineering, for example - for predicting missing values. Algorithm commonly used for building decision trees are ID_3, ASSISTANT, and C 4.5.

Issues in Decision Tree Learning:-

- Determine how deeply to grow the decision tree.
- Handling continuous attributes.
- Choosing an appropriate attribute selection measure.
- Handling training data with missing attribute values.
- Handling attributes with differing costs.
- Improving computational efficiency.

Advantages of Decision Learning:-

- Highly intuitive and easy to understand.
- Less number of data preparation steps unlike other machine learning algorithms.
- It is a non-parametric algorithm i.e it does not require lot of assumptions.
- Highly versatile algorithm and can perform multiple roles apart from the standard predictions.

Disadvantages of Decision Tree

- Overfitting
- Feature Reduction & Data Resampling
- Optimization
- Unstable
- Limited performance in Regression.

Merits:-
- Interpretability
- Less Data Preparation
- Non-Parametric
- Versatility
- Non-Linearity

Instance Based Learning :-

Instance based learning are the systems that learn the training examples and then generalizes to new instances based on some similarity measures. It is called instance based because it builds the hypothesis from the training instances. It is also known as memory based learning or lazy learning. The time complexity for this algorithm depends upon the size of training data. The worst case time complexity of this algorithm is $O(n)$. where n is the number of training instances.

for example if we were to create a spam filter with an instance based learning algorithm instead of just flagging emails that are already marked as spam emails, our spam filter would be programmed to also flag emails that are very similar to them this requires a measure of resemblence between two emails could be the same sender or the repetitive use of the same keyword or something else.

Advantages of Instance Based Learning :-

- Instead of estimating for the entire instance set, local approximations can be made to the target function
- This algorithm can adapt to new data easily, One which is collected as we go further.

Disadvantages:-

- classification cost are higher
- Large amount of memory require to store the data and each query involves starting the identification of a local model from scratch.

Some of the instance based learning algorithms are —

1. K nearest neighbour (KNN)
2. Self organizing Map (SOM)
3. Learning vector quantization (LVQ)
4. Locally weighted learning (LWL)

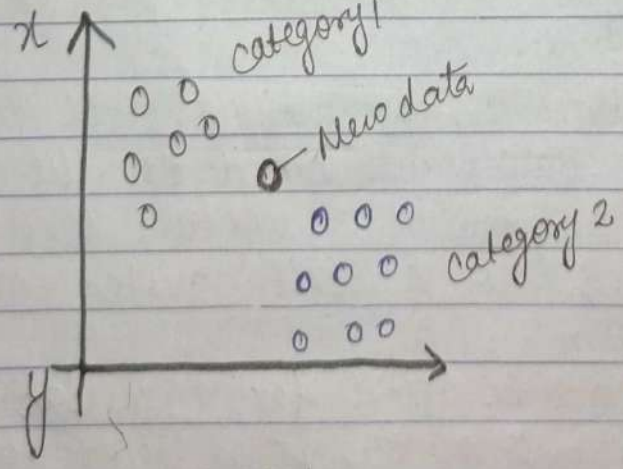
K Nearest Neighbour (KNN) Algorithm

- KNN is one of the simplest ML algorithm based on SL technique
- KNN algorithm assumes the similarity between the new data and available data and put the new data into the category that is most similar to the available categories.
- KNN algorithm stores all the available data and classifies a new data points based on the similarity This means when new data appears then it can be easily classified into a well suited category by using KNN algorithm.
- KNN algorithm can be used for regression as well as for classification but mostly it is used for the classification problems.
- KNN is the non-parametric algorithm, which means it does not make any assumption on underlying data.

- It is also called a lazy learner algorithm because it does not learn from the training set immediately, it stores the data set and at the time of classification it performs an action on the data set.
- KNN algorithm at the training phase just stores the data set and when it gets new data, then it classifies that data into a category which is much similar to the new data.

Need of the KNN algorithm:-

Suppose there are two category i.e category A & category B and we have a new data point x_1, so this data point will lie in which of these categories? To solve this type of problem we need a KNN algorithm. With the help of KNN we can easily identify the category or class for a particular dataset.



Working of KNN algorithm:-

The KNN working can be explain on the basis

of below steps.

Step 1:- Select the number K of the neighbour

Step 2:- Calculate the Euclidean distance of K number of neighbours.

Step 3:- Take the K nearest neighbour as per the calculated Euclidean distance.

Step 4:- Among these K neighbours count the number of data points in each category.

Step 5:- Assign the new data points to that category for which the number of the neighbour is maximum.

KNN classification problem:-

Qus°	English	Math	Result	Distance
1.	4	3	fail	5.38
2.	6	7	Pass	1
3.	7	8	Pass	1
4.	5	5	fail	3.16
5.	8	8	Pass	2

Qus. Student (English = 6, math = 8)

K = 3

Euclidean distance = $\sqrt{(x_{01}-x_{A1})^2+(x_{02}-x_{A2})^2}$

for 1 = $\sqrt{(6-4)^2+(8-3)^2} = \sqrt{4+25} = \sqrt{29} = 5.385$

for 2 = $\sqrt{(6-6)^2+(8-7)^2} = 1$

for 3 = $\sqrt{(6-7)^2+(8-8)^2} = 1$

for 4 = $\sqrt{(6-5)^2+(8-5)^2} = \sqrt{1+9} = \sqrt{10} = 3.162$

for 5 = $\sqrt{(6-8)^2+(8-8)^2} = 2$

After following the steps we take the 3 nearest distance as per the calculated Euclidean distance.

1, 1, 2

There are 3 pass and 0 fails

So,

3 > 0

Then this is belong to pass category.

Ques:2 We have a document classification task, A document $d_5[3,7]$ needs to be assigned a class. Below are 4 training document belonging to classes C_1 & C_2

	x_1	x_2	class	
d_1	7	7	C_1	4
d_2	7	4	C_2	5
d_3	3	4	C_1	3
d_4	1	4	C_1	3.605

K = 3

for 1 = $\sqrt{(3-7)^2+(7-7)^2} = \sqrt{16} = 4$

for 2 = $\sqrt{(3-7)^2+(7-4)^2} = \sqrt{25} = 5$

for 3 = $\sqrt{(3-3)^2+(7-4)^2} = \sqrt{9} = 3$

for 4 = $\sqrt{(3-1)^2+(7-4)^2} = \sqrt{4+9} = \sqrt{13} = 3.605$

3, 3.60, 4

There are 3 C_1 & 0 C_2

So

3 > 0

Then this belong to C_1 category

KNN Regression Problem:-

Name	Height	Age	Weight	
P_1	6	40	60	3.011
P_2	6.11	26	55	11.004
P_3	5.9	30	56	6.783
P_4	5.8	32	58	5
P_5	5.3	33	75	4.031
P_6	5.6	35	78	2.009
P_7	5.5	37	80	0.3

P_8 (H = 5.8, Age = 37, W = ?

for P_1 = $\sqrt{(5.8-6)^2+(37-40)^2} = \sqrt{0.04+9} = 3.011$

for P_2 = $\sqrt{(5.8-6.11)^2+(37-26)^2} = \sqrt{0.0961+121} = 11.004$

for P_3 = $\sqrt{(5.8-5.9)^2+(37-30)^2} = \sqrt{0.01+49} = 6.783$

for P_4 = $\sqrt{(5.8-5.8)^2+(37-32)^2} = 5$

for P_5 = $\sqrt{(5.8-5.3)^2+(37-33)^2} = \sqrt{0.25+16} = 4.031$

for P_6 = $\sqrt{(5.8-5.6)^2+(37-35)^2} = \sqrt{0.04+4} = 2.009$

for P_7 = $\sqrt{(5.8-5.5)^2+(37-37)^2} = \sqrt{0.09+0} = 0.3$

when K = 3

= 0.3, 2.00, 3.0

(P_1, P_6, P_7)

weight (P_8) = $\frac{60+78+80}{3}$ = 72.66

Selection of the K Value in KNN Algorithm

- There is Particular way to determine the best value for K, so we need to try some values to find the best out of them. The most preffered value for is 5
- A very low value for K, such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- large values for K are good but may find some difficulties.

Advantages of KNN algorithm :-

- It is simple to implement
- It is robust to the noisy training data.
- It can be more effective if the training data is large.

Disadvantages of KNN algorithm :-

- Always need to determine the value of K which may be complex sometimes.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Locally Weighted Regression :-

- LWR is instance based learning regression.
- The phrase "Locally Weighted Regression" is called
  i- Local because the function is approximated

based on data near the query point

ii- Weighted because the contribution of each training example is weighted by its distance from the query point.

iii- Regression because this is the term used widely in the statistical learning community for the problem of approximating real valued functions.

- Given a new instance x_q, the general approach in locally weighted regression is to construct an approximation $\hat{f}$ that fits the training examples in the neighbourhood surrounding x_q.
- This approximation is then used to calculate the value of $\hat{f}(x_q)$, which is output as the estimated target value for the query instance
- Consider locally weighted regression in which the target function f is approximated near x_q using a linear function of the form

$$\hat{f}(x) = w_0 + w_1 a_1(x) + \cdots + w_n a_n(x)$$

where $a_i(x)$ denotes the value of the ith attribute of the instance x.

- Gradient decent of ANN,

$$E(W) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)$$

t_d - targeted output

o_d - calculated output

30/11/22

UNIT-4

Artificial Neural Networks

Artificial Neural Networks- Artificial Neural Networks contains neurons which are called units. These units are arranged in a series of layers that together constitute the whole artificial neural network in a system. A layer can have only a dozen unit or millions of units as this depends on the complexity of the system.

Commonly artificial neural network has an input layer, output layer as well as hidden layers. The input layers receive the data from the outside world which the neural network needs to analyze or learn about. Then this data passes through one or multiple hidden layer that transform the input into data that is valuable for the output layer. Finally the output layer provides an output in the form of a response of the ANN to input data provided. In the majority of neural network units are interconnected from one layer to another. Each of these connections has weights that determine the influence of one unit on another unit. As the data

transfers from one unit to another, the Neural networks learns more and more about the data which eventually results in a output from the output layer.

Examples of ANN learning :-

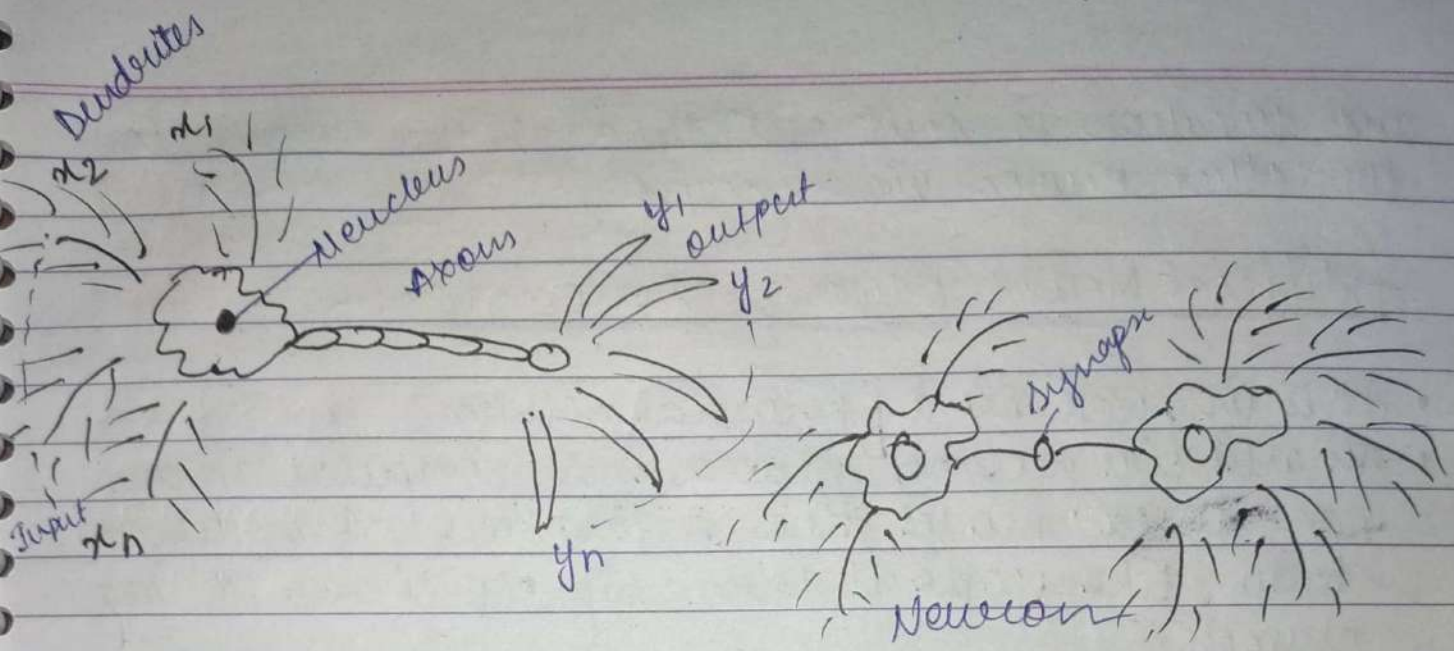
How machine will learn from the given data :-

ANN are trained using a training set for ex :-
Suppose you want teach an ANN to recognise a cat. Then it is shown thousands of different images of cats so that the network can learn to identify a cat once the neural network has been trained enough using images of cats, then you need to check if it can identify cat images correctly. This is done by making the ANN classify the images it is provided by deciding whether they are cat images or not. The output obtained by ANN is a human provided description of whether the image is a cat image or not. If the ANN identifies incorrectly then back propagation is used to adjust whatever it has learnt during training. Back propogation is done by fine tuning the weights of the connection in ANN units based on the error rate obtained this process continues until the ANN can correctly recognise a cat in an image with minimal possible error rates

Types of ANNs-

Working of Biological Neurons-

- Neurons are the fundamental unit of the brain
- ANN is made up of number of processing elements called as Neuron (A Neural Network)
- Human Brain consist of a huge number of neurons approx 10^{11} with numerous interconnections
- A neural network consist of many interconnected neurons.
- The brain consist of billions of cells called neurons.
- Neurons are connected together by synapse which are the connection between the neurons by which a neuron can send impulse to the another neuron.
- There are approx 10^4 synapses per neuron in the human brain.
- The synapses have processing value or weight
- Each neuron accept input from the external world via dendrites or from output of other neurons, process it and gives the output through axons
- A neuron is composed of nucleus, cell body, dendrites (or tree like networks) made up of nerve fibre connected to the cell body.
- Axon is a single, long interconnection, extending from the cell body and carrying signal from the neurons

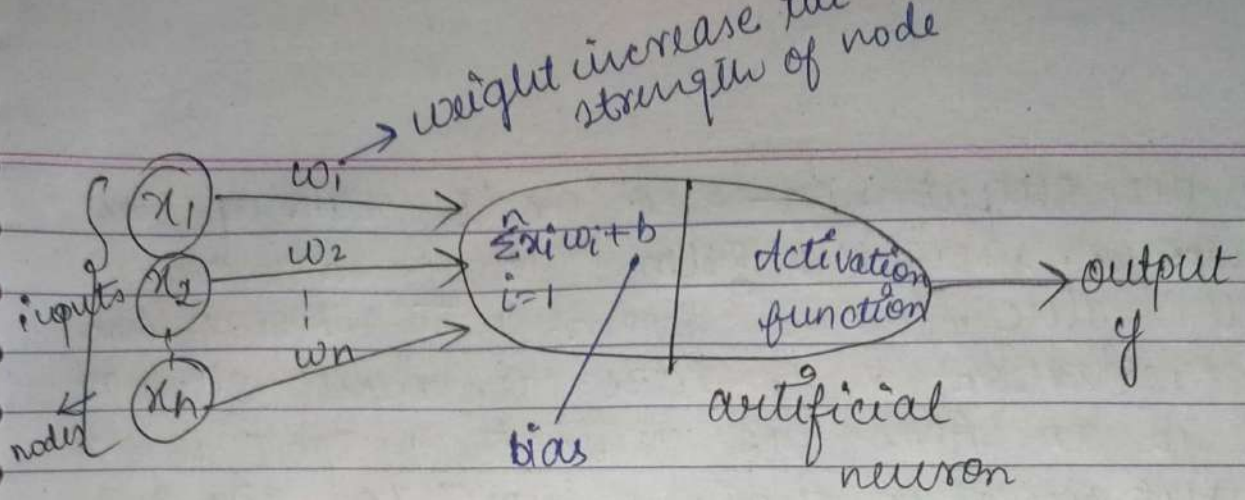


- The end of the axon splits into fine strands and each strands terminates into a small bulb like organ called synapse.
- It is through synapse that neuron introduce its signal to the other nearby neuron.
- In the brain neuron collect signals from other neurons through a host of fine structures called dendrites. The neuron sends out spikes of electrical activity through the axon which can split into thousands of branches. At the end of each branch a synapse converse the activity from the axon into electrical effect that inhibit activity on the contacted (target) neurons.
- When a neuron sends an excitatory signal to other neuron then this signal will be added to all of the other input of that neuron
- If the commulative input received by the cell body raises the internal electric potential of a cell to threshold then the receiving cell fire a pulse or action potential of fixed strength

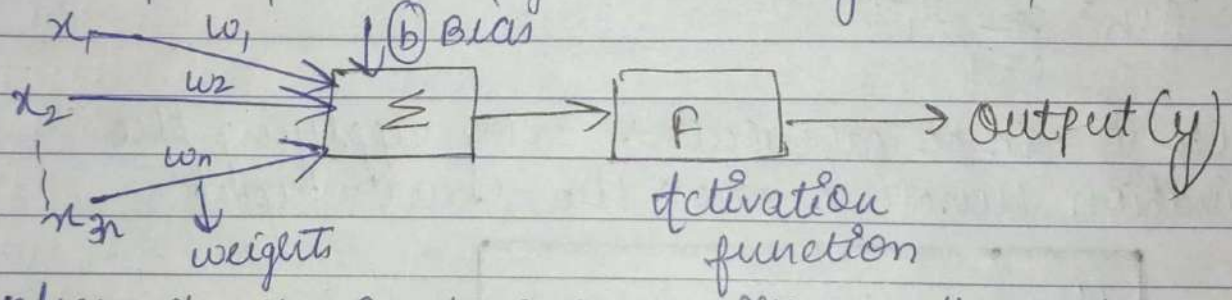
and duration is sent out through the axons to the other neuron via synapse.

Artificial Neural Network phenomenon:-

- It is inspired by a biological neuron.
- An artificial neuron network is computational non-linear model that is inspired from the brain like people it learn by experience or by examples.
- ANN can perform task like classification, predication decision making, visualization and others just by considering examples.
- ANN consist of large collection of artificial neurons or processing elements which operates in parallel.
- Every neuron is connected with other neuron through a connection link.
- Each connection link is associated with weight that has information about the input signal.
- Weight is the most useful information for neurons to solve a particular problem because it is usually inhibits the signal that is being communicated.
- Every neuron has weighted input (synapse), an activation function (that defines the output) and one output.
- The neurons/nodes can take input data and perform simple operations on the data and the result of these operations is passed to other neurons.
- ANN are capable of learning which takes place by altering weight values.



ANN can be viewed as directed graphs in which artificial neurons are nodes and directed edges with weights are connections between neuron outputs and neuron inputs. A single layer neural network is called perceptron. It gives a single output.



where x_1, x_2 and so on x_n represents various input (independent variable to the network). Each of these inputs is multiplied by a connected weight $w_1, w_2, \ldots w_n$ show the strength of a particular node. b is bias value, it allows to shift the activation function up or down. Product of input weights are sumed and fed to an activation function to generate a result and this result is sent as output.

ANN receives input from the external world in the form of pattern and images. Each input is multiplied by its corresponding weight and all weighted input are sumed up inside computing unit (Artificial Neuron).

In case the weighted sum is 0 bias is added

to make the output non-zero or to scale up the system response. Weighted sum correspond to any numerical value ranging from 0 to 1 the main purpose of activation is to take an input signal of a node in an ANN to generate output signal. this output signal is used as input to the next layer in the network.

from the general model of ANN the net input can be calculated as

$$y_{in} = x_1 \times w_1 + x_2 \times w_2 + \ldots + x_n \times w_n$$

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

The output can be calculated as by applying the activation function over the net input.

$$y = f\left(\sum_{i=1}^{n} x_i w_i + bias\right)$$

In order to limit the response at desire value, the threshold value is set up. The activation function is set of the transfer function use to get desired output.

Activation function :-

- Activation function is used to get the output of a node. It convert an input signals of a node in new output signal.
- Activation function is used to maps the resulting values in between 0 to 1 or -1 to 1 depending upon the type of function used.

There are mainly two types of activation function
(I) Linear activation function
(II) Non-linear activation function.

- Activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias to it.
- Activation function introduce non-linearity into the output of a neuron. If we do not apply activation function then the output signal would be simply linear function (one degree polynomial.)
- Activation function is of two types

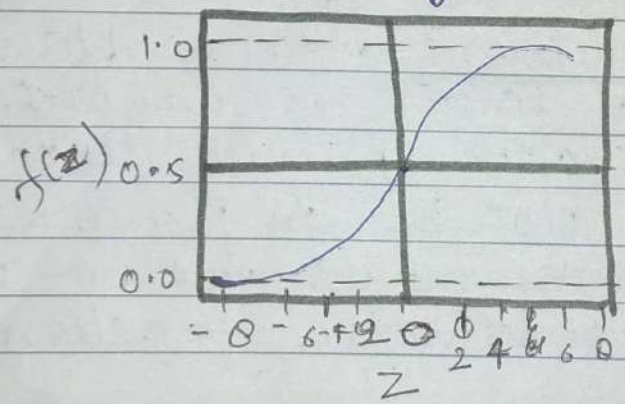
(1) Linear Activation function:- Linear Activation function is used as linear or identify activation function. The output after applying linear activation function will not be confined between any range. Neural Network would not be able to learn under linear activation function.

(2) Non Linear Activation function:- NL Activation function makes easy for the model to generalize or adapt with variety of data and to differentiate between the output. It makes the network more powerful and add ability to learn something complex and complicated for data. Some non linear activation functions are logistic, sigmoid etc.

(II)
The Non Linear Activation function are mainly divided on the basis of their range or curves.

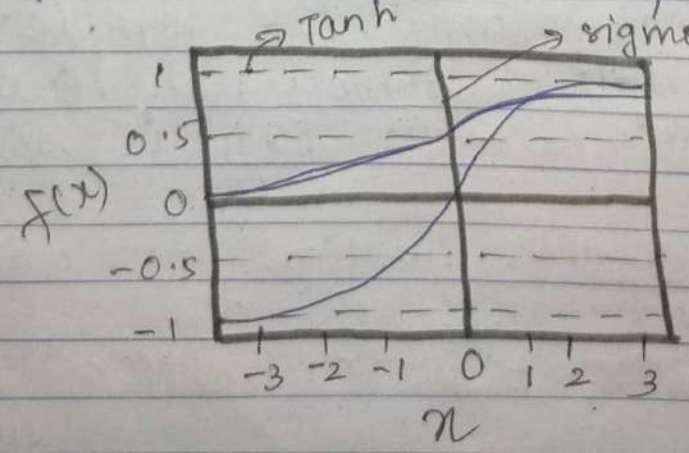
(1) Sigmoid or Logistic Activation function The sigmoid function curve

looks like S shape, the main reason why we use sigmoid function is because it exist between 0 to 1. Therefore it is specially used for models where we have to predict the probability as an output. The function is differentiable that means we can find the slope of sigmoid curve at any two points. The function is monotonic but function derivative is not. The logistic sigmoid function can cause a neural network to get stuck at the training time.



② Tanh or Hyperbolic Tangent Activation function:-

Tanh is also like logistic sigmoid but it is better than that. The range of the Tanh function is from -1 to 1. Tanh is also S shaped

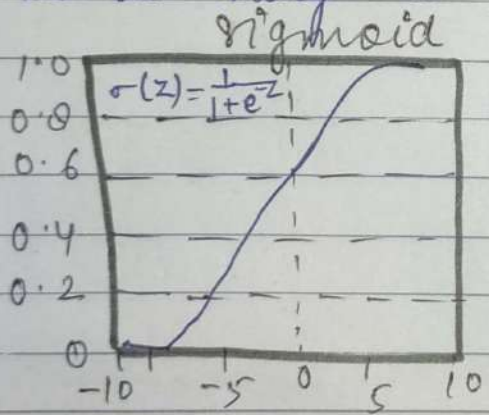


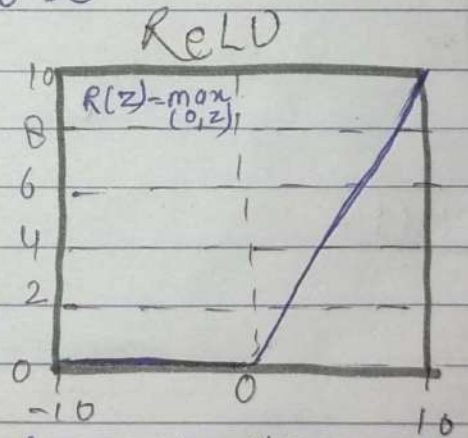
The function is montonic while its derivation is not.

The advantage is that the negative input will be mapped strongly negative and the zero input will be mapped near zero in the tanh graph. The function is differentiable. The tanh function is mainly used classification between two classes. Both Tanh and logistic sigmoid activation function are used in feed forward network.

③ Rectified Linear Unit (ReLU) Activation function:-

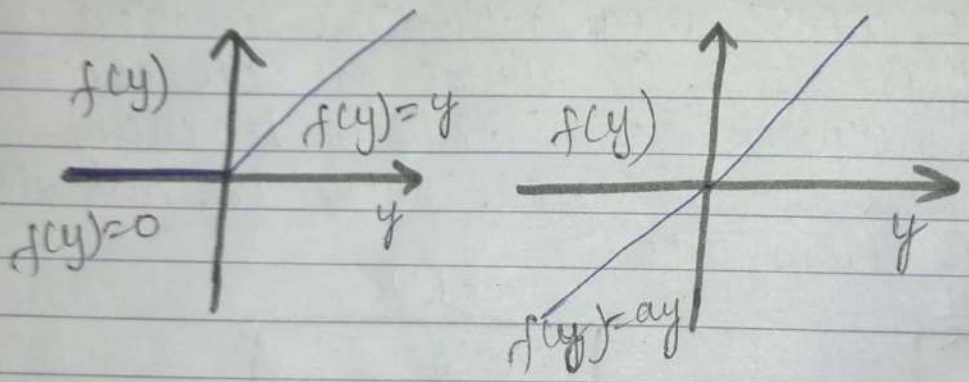
The ReLU is the most used activation function in the world. It is used in almost all the convolu-tional neural network (CNN) or deep learning. The function and its derivative both are monoton-ic its range is from 0 to ∞





The ReLU is rectified from bottom. f(z) is zero when z is less than zero and f(z) = z when z is above or equal to zero. The issue is that all the negative values become zero immediately which decreases the ability of the model to fit or train from the data properly that means any negative input given to the ReLU activation function turns the value into zero immediately in the graph which in turn affect the resulti-ng graph by not mapping the -ve values appropriately.

④ Leaky ReLU Activation function:-
In this function the leak help to increase the range of the ReLU function. Usually the value is 0.01. When it is not 0.01 then it is called Randomised ReLU. Therefore the range of the Leaky ReLU is $-\infty$ to ∞. Both Leaky and Randomised ReLU functions are monotonic in nature and also their derivatives are monotonic in nature



Perceptrons:-

Machine Learning Technique

Unit - 5 [One Shot]

Most important topics:

1. Reinforcement learning, tasks & examples
2. Markov Decision Process
3. Q-Learning (functions, algorithm)
4. Deep Q-learning (introduction)
5. Genetic Algorithm

Correct learning

① Reinforcement Learning:

- It is a type of learning in ML where an agent learns to make decision by taking actions in an environment to receive feedbacks as reward or penalty.

* Key components:

- Agent: Learner or decision-maker.
- Environment: External system through which the agent interacts.
- State (S): A representation of current situation of agent.
- Action (A): Choices available to the agent.
- Reward (R): Immediate positive feedback from environment after an action.
- Penalty (P): Immediate negative feedback.
- Policy (π): Strategy that agent follows to determine its actions.

* Tasks in Reinforcement Learning:

- Exploration and exploitation: The agent must explore the environment to find the best actions and exploit known actions that provide high rewards.
- Learning from interaction: The agent learns by continuously interacting with the environment and updating its strategy based on

* Examples:

- Game playing: like chess is played by agents that learn strategies using Reinforcement Learning.
- Robotics: RL is used by robots to learn tasks like walking, navigating, or manipulating objects.

(ii) Markov Decision Process (MDP):

- It is a mathematical framework for modelling decision-making situations where outcomes are partly random and partly under the control of the agent.

* Key components:

- Set State (S): Set of all possible situations in which the agent can be.
- Actions (a): Set of all possible actions an agent can take
- Transition model (T): The probability of transitioning from one state to another due to action a.
- Reward function (R): The immediate reward received after transition.
- Policy (π): Defines agent's behaviour.

* MDP framework:

1. The agent starts in a state 's'.
2. It chooses an action 'a' based on it's policy 'π'.
3. It receives a reward R(s,a) and transitions to a new state 's'' according to a transition model T(s,a,s').
4. The process repeats, and the agent aims to maximize the cummulative reward over time.

Q-Learning:

- It is a popular, model-free Reinforcement learning algorithm that learns the quality of actions, denoted as Q-values, which represents the expected utility of taking an action in a given state and following the optimal policy.

- Q-Function (Q):

- The Q-value $Q(s,a)$ represents the expected cumulative reward of taking action 'a' in state 's' following the optimal policy.

- The goal is to learn the optimal Q-function $Q^*(s,a)$ which gives the highest expected reward for any state-action pair.

* Algorithm:

- Initialize the Q-table $Q(s,a)$.

- For each episode:
  - Initialize the starting state 's'.
  - Choose an action 'a' using an exploration strategy.
  - Take action 'a' and observe the reward 'r' and the next state 's'.
  - Update the Q-value using the Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

where,

α: learning rate, control the amount of updation

γ: discount factor, determines importance of future rewards

  - Update the state s to the new state s'.

Deep Q-Learning :

- It is an extension of Q-Learning that uses a neural network to approximate the Q-values, making it feasible to handle large or continuous state spaces.

Introduction:

- Neural network: Instead of a Q-table, DQL uses a neural network to approximate the Q-values.
- Experience replay: To stabilize training, DQL stores the agents experiences in a replay buffer.
- Target Network: DQL uses two neural networks: the primary network Q and the target network Q'. It helps in stablizing the training by providing consistent targets during updates.

@brainlearning

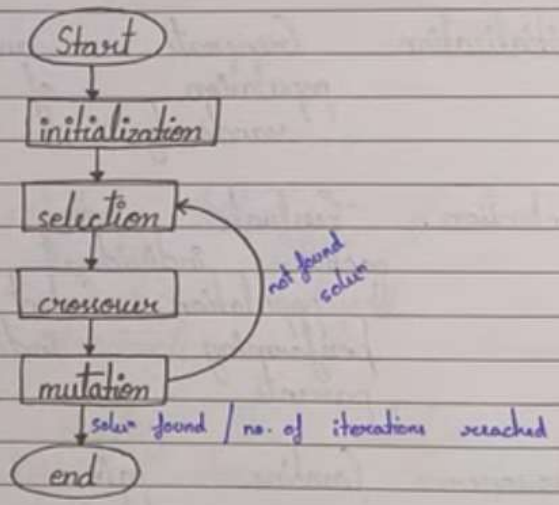
Genetic Algorithm.

- It is an optimization technique based on the principles of natural selection and genetics.

Steps:

1. Initialization: Generate an initial population of candidate randomly.
2. Selection: Evaluate the fitness of each individual in the population. Select best performing individuals to be parents.
3. Crossover: Combine pairs of parents to create new offspring.
4. Mutation: Introduce random changes to some offspring to maintain genetic diversity and explore new parts of the solution space.
5. Replacement: Replace the old population with the new generation of idividuals.

6. Iteration: Repeat the selection, crossover, mutation, and replacement steps for a no of generations or until a solution is found.



Thanks for watching !!

@bruvilearning