DATA SCIENCE

PAGE NO.
DATE: / /

Unit I Introduction to Data science;- Basics of Data science, Data Science, Data Analytics, Machine Learning (Supervised, Unsupervised & reinforcement), Deep Learning (Artificial Neural Networks, CNN), working with data sources- (SQL server, -CSV file, excel file) etc, Real world applications of Machine Learning & Deep Learning, Scope of Data Science.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from noisy, structured and unstructured data & apply knowledge and actionable insight from data to a broad range of application domains. Data science is related to machine learning, data mining and big data.

Data Science is a "concept of unify statistics, data analysis, informatics, & their related ~~queries~~ methods". It uses unique techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, Information science & domain knowledge.

Jim Gray turing award winner - data science as a 4th paradigm of science (empirical, theoretical, computational and now data driven).

- It is interdisciplinary field focused on extracting knowledge from data sets, which are typically large (big data) & applying the knowledge and actionable insight from data to solve problems in a wide range of application domains.

PAGE NO.
DATE: / /

Data Analysis:- Process of inspecting, cleansing, transforming & modelling data with the goal of achieving useful information, informing conclusion and supporting decision making.

Data Analytics process has some components that can help a variety of initiatives. By combining these components, a successful data analytics initiative will provide a clear picture of where you are, where you have been & where you should go.

Types of Data Analytics are:-

1) **Descriptive:-** It helps to answer question like what happened. These techniques summarizes large dataset to describe outcomes to stakeholders. By developing Key performance indicator (KPI). these strategies can help track successes or failures.
(metrics such as return on investment (ROI) are used in many industries)

2) **Predictive:-** It helps to answer questions about what will happen in the future. These techniques uses historical data to identify trends & determine if they are likely to recur. Predictive analytical tools provide valuable insight into what may happen in the future. & this technique uses a wide variety of statistical & machine learning techniques such as KNN, decision trees and regression.
(neural network)

3) **Prescriptive Analytics:-** It helps to answer questions about what should be done. By using insights from predictive analytics, data driven decisions can be made. This allows busines to make informed decisions in the face of uncertainty. It rely on machine learning strategies that can find pattern in large datasets. By analysing past decisions and patterns, the likelihood of different outcomes can be estimated.

It optimise efficiency in many different industries.

4) Diagnostic Analytics :- Diagnostic analytics is form of advanced analytics the examines data or content to answer questions "why did it happens". It is characterised by techniques such as ~~Data~~ Drill down, data recovery, data mining and correlations.

This occurs generally in three steps :-

1) Identify anomalies in the data. They may be unexpected changes in the metric or a particular market.
2) Data that is related to these anomalies is collected.
3) Statistical techniques are used to find the relationship and trends that explain these anomalies.

MACHINE LEARNING

Machine Learning is a field of study that gives computer the ability without being explicitly programmed. It is one of the most exciting technology that one would have ever come across. As it is evident from the name, it gives the computer that make it more similar to humans i.e. the ability to learn or behave like a human.

It is the study of computer algorithm that can improve automatically through experience by the use of data. ML algos build a model based on sample data (training data) in order to make predictions / decisions without being explicitly ~~preformed~~ programmed to do so. ML algos are used in wide variety of applications such as in medicine, email filtering, speech recognition & computer vision.

ML Dept AI

PAGE NO.
DATE: / /

A subset of machine learning is closely related to computational statistics; which focuses on making predictions using computers. The study of mathematical optimization delievers methods, theory and application domains to the field of machine learning.

Data mining is a field of study focusing on 'exploratory data analysis' through unsupervised learning.
In its application across business problems, machine learning is also refered to as predictive analytics.

Types of Machine learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning

Supervised ML is a type of ML in which machines are trained using well-labelled training data and on the basis of that data machines predict the output.

Labelled data means some i/p data is already with the correct o/p.

In this, the training data ~~works~~ provided to the machine works as a supervisor that teaches the machines to predict the o/p correctly.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model.

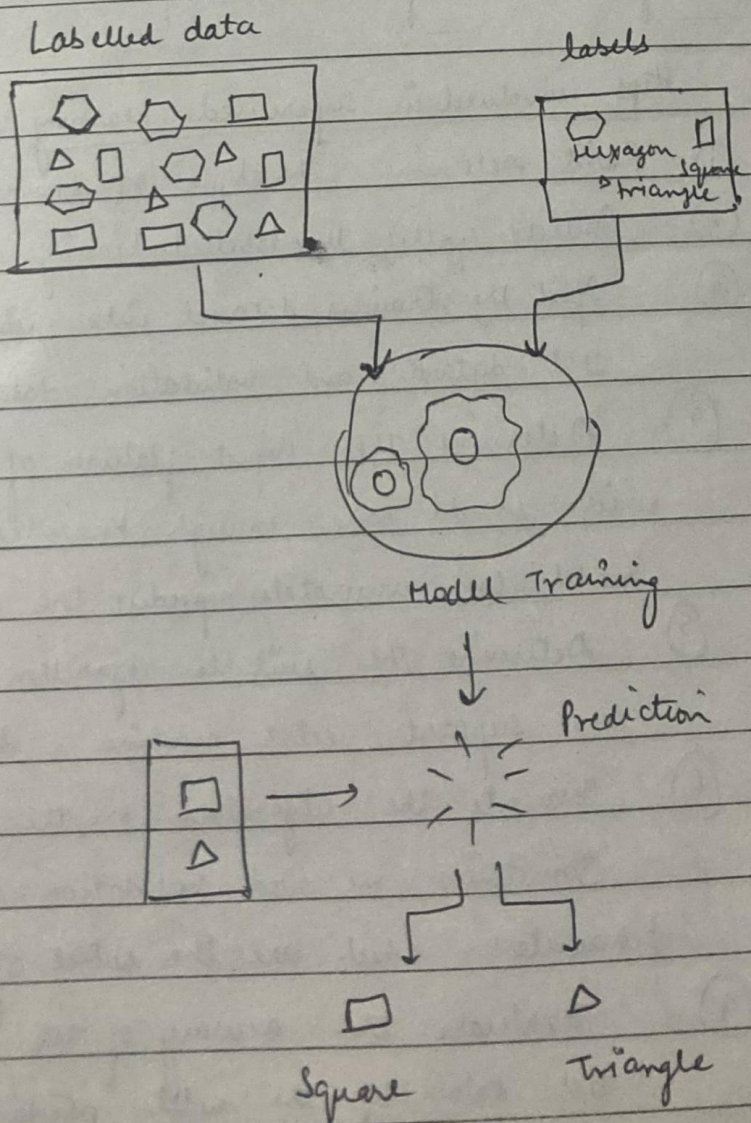
PAGE NO.
DATE: / /

The aim of a supervised learning algorithm is to find a mapping function that maps the input variable(x) to the output variable(y).

In real world, supervised learning can be used for Risk Assessment, Image classification, fraud Detection and spam filtering.

How supervised learning works?

In supervised learning, models are trained using labelled data set where model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of a training set), and then it predicts the output.



PAGE NO.
DATE: / /

Suppose we have a data set of different types of shapes which includes square, rectangle, triangle and polygon. Now the first step is that we need to train the model for each shape.

1) If the given shape has four sides and all the sides are equal, then shape is labelled as Square.
2) If the given shape has three sides then it will be labelled as triangle.
3) If the given shape has six sides, then it will be labelled as Hexagon.

Now, after training, we test our model using the test data and the task of model is to identify the shape. The machine is already trained on all types of shapes and when it finds a new shape, it classifies the shape on the basis of a number of sides and predict the output.

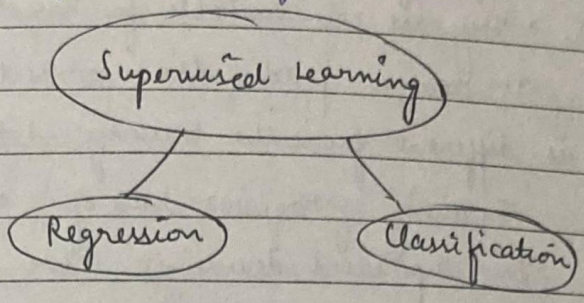
Steps involved in Supervised Learning :-

(1) First determine the type of training dataset.
(2) Collect/ Gather the labelled traing data.
(3) Split the training dataset into training dataset, test dataset and validation dataset.
(4) Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
(5) Determine the suitable algorithm for the model such as support vector machine, decision tree, etc.
(6) Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters which are the subset of training datasets.
(7) Evaluate the accuracy of the model by providing the test data. If the model predicts the correct output, which means our model is accurate.

Types of Supervised Machine Learning Algorithm

PAGE NO.
DATE:- / /

Supervised learning can be further divided into two types of problems:-



① **Regression:-**

Regression algorithms are used if there is a relationship b/w the input variable and the output variable. It is used for the prediction of continuous variable such as Weather forecasting, Market trends, etc. Some popular regression algorithms:-

1. Linear Regression
2. Non-Linear Regression
3. Bayesian Linear Regression
4. Regression Trees
5. Polynomial Regression.

② **Classification:-**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-False, etc.

Spam filtering

- Random Forest
- Decision Trees
- Support vector machines
- Logistic Regression.

Advantages of Supervised Learning model →

1. With the help of the supervised learning, the model can predict the output based on the prior experiences.
2. In this, we can have an exact idea about the classes of objects.
3. It helps us to solve real-world problems such as fraud detection, spam-filtering, etc.

PAGE NO.
DATE: / /

Disadvantages of Supervised Learning models:-

1. They are not suitable for handling complex tasks
2. It cannot predict the correct output if the test data is different from the training data set.
3. Training requires lots of computation times.
4. In supervised learning, we need enough knowledge about the classes of objects.

Unsupervised Machine Learning:-

In supervised machine learning model, models are trained using labelled data under the supervision of training dataset. But there are many cases when we do not have labelled data and need to find the hidden patterns from the given dataset. So, to solve such types of cases in machine learing, we need unsupervised learning techniques.

As name suggests, Unsupervised learning is a ML technique in which models are not supervised using training dataset. Instead models finds the hidden patterns and insights from the given data. It can be compared to the learning that takes place in human brain while learning new things.

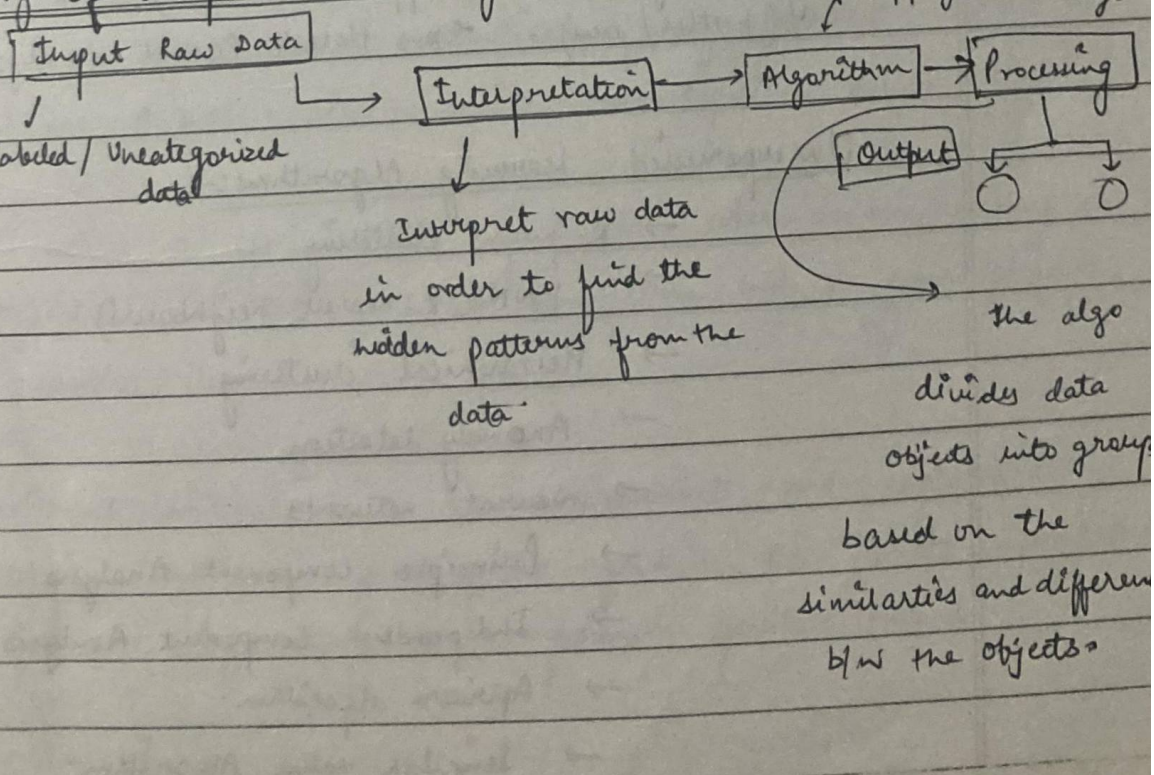
Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to the similarities and represent that dataset in a compressed format.

PAGE NO.
DATE: / /

Example:-
Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset which means it does not have any idea about the features of the dataset. The task of unsupervised learning is to identify the image features on its own.
~~Then~~ This will perform this task by clustering the image dataset (ULA) into the groups according to similarities between images.

Why use Unsupervised learning? (Importance)
(1) It is helpful for finding useful insights from the data.
(2) It is much similar to a human learning from their own experiences, which makes it closer to the real AI.
(3) It works on unlabeled and uncategorized data which makes unsupervised learning more important.
(4) In real world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

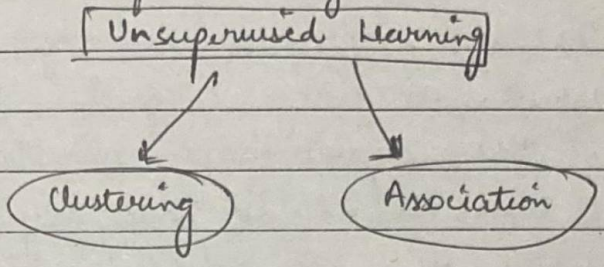
Working of Unsupervised learning



PAGE NO.
DATE: / /

Types of Unsupervised Learning Algorithm

ULA can be further categorized into two types of problems



① **Clustering :-** Clustering is a method of grouping the objects into clusters such that the objects with most similarities remain into a group and has less no similarities with the objects of another group. Cluster analysis finds the commonalities b/w data objects and categorizes them as per the presence and absense of those commonalities.

② **Association :-**

An association rule is an unsupervised learning method which is used for finding the relationships b/w variables in the large database. It determines the set of items that occur together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to buy Y (butter/jam). → Ex → Market Basket Analysis.

Unsupervised Learning Algorithms :-

- → K-means clustering
- → KNN (K-nearest neighbours)
- → Hierarchical clustering
- → Anomaly detection
- → Neural networks
- → Principle Component Analysis
- → Independent Component Analysis
- → Apriori algorithm
- → Singular value Algorithm.

PAGE NO.
DATE: / /

Advantages of Unsupervised learning :-

1. USL is used for more complex tasks as compared to SL because in USL, we don't have labeled input data.
2. It is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised learning :-

1. It is intrinsically more difficult than supervised learning as it does not have corresponding output.
2. The result USL algo might be less accurate as input data is not labelled and algorithms do not know the exact output in advance.

What is Reinforcement learning :-

⇒ Reinforcement learning is a machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of action. For each good action, agent gets the positive feedback and for each bad action agent gets the negative feedback or penalty.

⇒ In this, agent learn automatically using feedback without any labelled data., learns by its own experience only.

⇒ RL solves a specific type of problem where decision making is sequential and the goal is long-term such as game-playing, robotics, etc.

⇒ The agent interacts with the environment and explores it by itself. The main goal of the agent in RL is to improve the performance by getting the maximum positive rewards.

PAGE NO.
DATE: / /

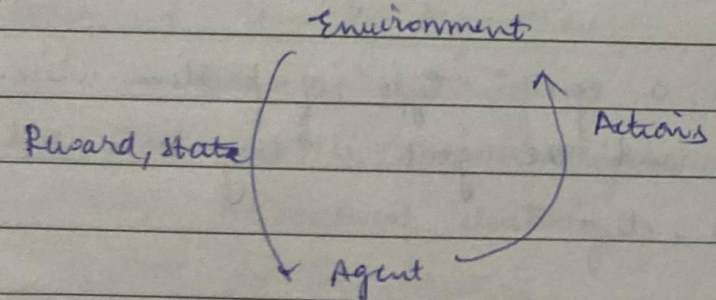
⇒ The agent learns with the process of hit and trial, and based on the experience, it learns to perform the task in a better way. Hence, Reinforcement learning is a type of machine learning method in which an intelligent agent (computer program) interacts with the environment and learns to act within that.

Ex: Robotic dog learns the movement of its arms.

Example:- Suppose there is an AI agent present within a maze environment and his goal is to find the diamond. The agent interacts with the environment by performing some actions and based on these actions the state of agent will get change and it receives a reward or penalty as feedback

The agents continues doing these three things →
(take action, change state/ remain in the same state, and get feedback) and by doing these actions he learns and explores the environment.

The agent learns that what actions lead to positive feedback and what actions lead to negative feedback. As a positive feedback the agent gets a positive reward and as a penalty, it gets negative ~~reward~~ point.



Deep Learning

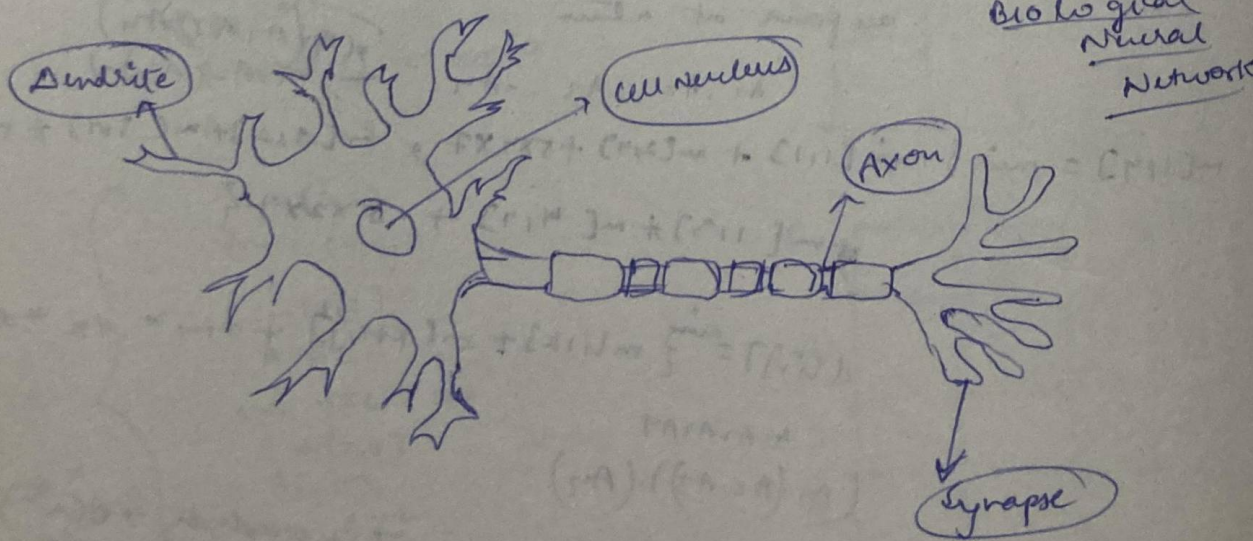
Deep learning is a class of machine-learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input. For ex→ in image processing lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.

The adjective 'deep' in deep learning refers to the use of multiple layers in the network.

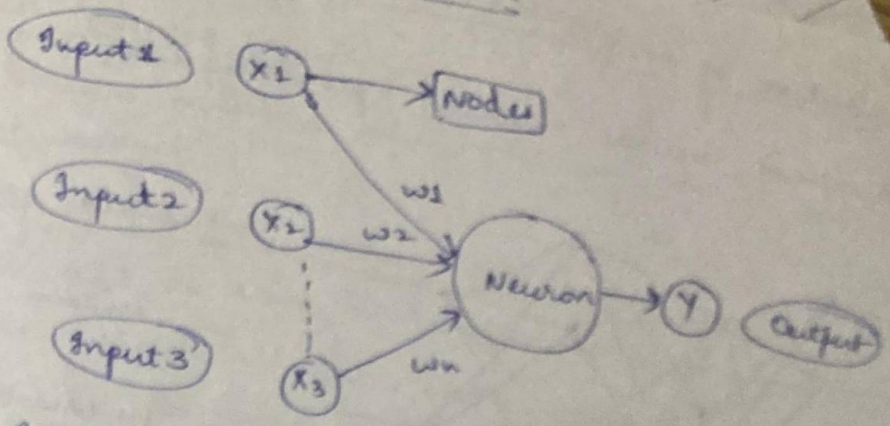
Artificial Neural Network :-

The term "Artificial Neural Networks" is derived from Biological Neural Networks that develop the structure of human brain. Similar to the human brain that has neurons interconnected with one another, artificial neural networks also have neurons that are connected to one another in various layers of the networks. These neurons are known as "nodes".

Biological Neural Network



Artificial Neural Network



Relationship b/w BNN and ANN

BNN	ANN
Dendrites	Inputs
Cell Nucleus	Nodes
Synapse	Weights
Axon	Output

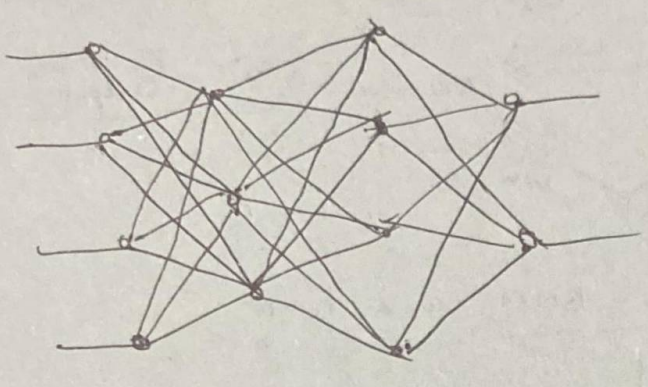
An Artificial Neural Network in the field of Artificial intelligence where it attempts to mimic the network of neurons makes up a human ~~brain~~ brain so that computers will have an option to understand things and make decisions in a human-like manner.

In BNN, each neuron has an association point somewhere in range 1000 and 10000. In the human brain, data is stored in a such a manner as to be distributed, and we can extract more than one piece of this data when necessary from our memory (Parallely).

⇒ (OR gate example)

Architecture of ANN :-

Artificial Neural Networks primarily consists of three layers →



- Input layer
- Hidden layer 1
- Hidden layer 2
- Output layer

Input Layer :-

As the name suggests, input layer accepts input in several different formats provided by the programmer.

Hidden Layer :-

The hidden layer presents in btw input and output layers.
It performs all the calculations to find features and patterns.

Output Layers :-

~~The input layer presents in btw input and output layers~~
~~It pe~~
the input goes through a series of transformations using the hidden layer, which finally results in output which is conveyed using this layer.

The ~~on~~ artificial neural network takes input and computes the weighted sum of the inputs.

the computation is represented in the form of a transfer function.

$$\sum_{i=1}^{n} Wi * Xi + b.$$

⇒ Advantages of Artificial Neural Network:-

1) Parallel Processing Capability:-
ANN have a numerical value that can perform more than a task simultaneously.

2) Storing data on the Entire Network:-
Data that is used in traditional programming is stored on the whole network, not on the database.

3) Capability to work with incomplete Knowledge:-
After ANN training, the information may produce output even with inadequate data. The loss of performance here relies upon significance of missing data.

4) Having a memory distribution:-
The succession of the network is directly proportional to the chosen instances and if the event can't appear to the network in all aspects, it can produce false output

5) Having fault tolerance:-
Exortion of one or more cells of ANN does not prohibit it from generating output, and this features makes the network fault-tolerance.

⇒ Disadvantages of ANN

1) Assurance of proper Network structure:- There is no particular guideline for determining the structure of artificial neural networks.

2) Unrecognized Behaviour of the Network:- When ANN produces a testing solution, it does not provide insight concerning why and how. It decreases trust in the network

3) Hardware dependence:- ANN needs processors with parallel processing power. Therefore, realization of equipment is dependent.

4) Difficulty of showing issue to the network:- The presentation needs in to resolve the problems into numerical values before being introduced to ANN will directly impact the performance of the network.

5) Duration of Network is unknown:- The network is reduced to a specific value of error, & this value

How do Artificial Networks work?

1) ANN can be represented as weighted directed graphs, where
2) artificial neurons forms nodes.
The association b/w neuron outputs and inputs can be viewed as the directed edges with weights.
3) ANN receives input signals from external source in the form of patterns and image in the form of vector.
4) These inputs are then mathematically assigned by the notation $x(n)$ for every n number of inputs.
5) Now, each input is multiplied by its corresponding weight.
6) All weighted inputs are summarized inside the computing unit.
7) If weighted sum is zero, then bias is added to make the output non-zero. Bias has same input and weight equals to 1. A maximum value is benchmarked to keep response in limit. and total of weighted inputs is passed through the activation function.
8) The activation functions refers to the set of transfer functions to achieve the desired output. Ex→ Binary, Linear, Tan hyperbolic sigmoidal activation functions.

Binary

output → 0 or 1
If net weighted input of neurons ~~is~~ >1, then O/p → 1 otherwise 0.

Sigmoidal hyperbolic → S shaped curve.

Tan hyperbolic f^n is used to approximate o/p from actual net i/p.

Types of ANN

Feedback ANN

The feedback networks feed information back into itself and are well suited for optimization problems.
Ex→ Internal system error corrections

Feed-forward ANN

→ Input layer, output layer & at least one layer of neuron. It figure out how to evaluate and recognize i/p patterns.

CNN (Convolutional Neural Network) (ConvNet)

It is a Deep learning algorithm which can takes in an input image, assign importance to various aspects / objects in the image and be able to differentiate one from another. The pre processing required in ConvNet is much lower as compared to other classification algorithms.

ARCHITECTURE

↳ Input layer
↳ Hidden layers → middle layers called hidden layers bcoz their inputs and outputs are masked by the activation function.

Convolutional Layers :- In CNN, the input is a tensor with a shape : (no of inputs) × (input height) × (input width) × (input channels). After passing through a convolutional layer, the image becomes abstracted to a feature map, also called an activation map, with shape : (no of inputs) × (feature map height) × (feature map width) × (feature map channels).

→ These layers convolve the i/p and pass its result to next layers.
→ Convolution reduces the number of free parameters allowing the n/w to be deeper.
→ Convolutional NN are ideal for grid-like topology.

Pooling layers :- Pooling layers reduce the dimensions of data by combining the o/p's of neuron clusters at one layer into a single neuron in the next layer.

Fully connected layers :-
They connect every neuron in one layer to every neuron in another layer.

Receptive field → Receptive field is the area in the layer from which the neurons in the next layers receives the i/p. In case of ~~CNN~~ convolution layer, it is typically (5 by 5 neurons) area of square and in fully connected layer, receptive field is the entire previous layer.

Weights → The function that is applied to the input values is determined by a vector of weights and **bias**.

Advantages of Data Science

- Data science can be fun.
- Multiple Job Designations
- Ease of Job Hunting
- Customize the Products
- A Highly Paid career
- Cost optimization
- AI is the future.

Drawbacks

- Data Security
- Complexity
- Term is Misleading
- Does not Allows to expertise.

Application of Data Science

- Fraud and Risk Detection
- Health care
- Augmented Reality
- Gaming
- speech Recognition
- Image Recognition

future scopes

- Health care sector
- Transport sector
- E-commerce sector