U-3 (OS honors).

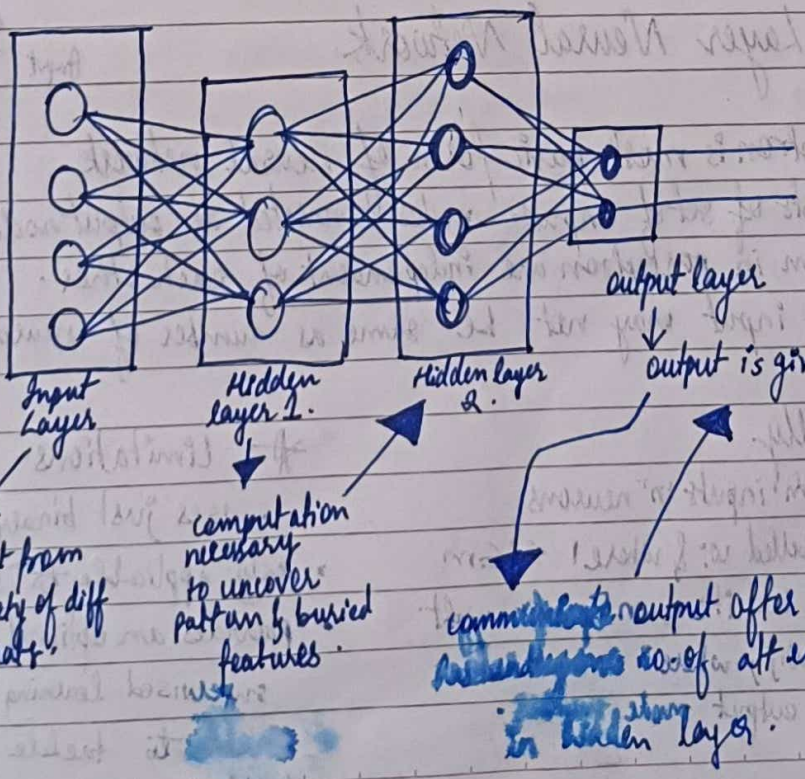
DOMS | Page No.
Date / /

Artificial Neural network -

- inspired by biological neural networks.
- mimic network of neurons like human beings so computers can be make decision in human like manner.

Adv →
- parallel processing capability
- sorting data on entire network.
- capability to work with incompleted knowledge.
- Having memory distribution.
- having fault tolerance.

Disadv →
- assurance of proper network structure
- unrecognizable behaviour of network.
- Hardware dependance
- difficult of showing issues to network
- duration of network is unknown.



DOMS | Page No.
Date / /

Types of ANN

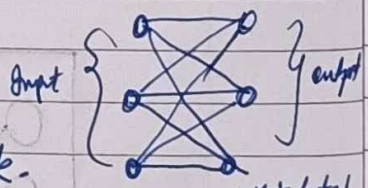
1) Feed Forward ANN
   - → atleast one layer of neurons
   - → network intensity → output & input layer connected neurons
   - → output → network's output in context of it's input
   - → benefit → learns to assess & identify input patterns

2) Feedback ANN
   - ↳ output loops back into network to acheive best internally involved results
   - → addressing optimization results
   - → feed info into themselves
   - → internal system error repairs

Appln of ANN
1) Social Media → Insta → People you may know
2) Sales & Marketing → ecommerce → suggest things acc to history
3) Healthcare → facial analysis
4) Personal Assistants → Siri, alexa, etc

Perceptrons –

1) Single Layer Neural Network



- Perceptron is most basic form of neural network.
- Consists of set of inputs nodes connected to output nodes using connections
- Neuron in perceptron are independent of each other.
- No. of input may not be same as number of neurons (output)

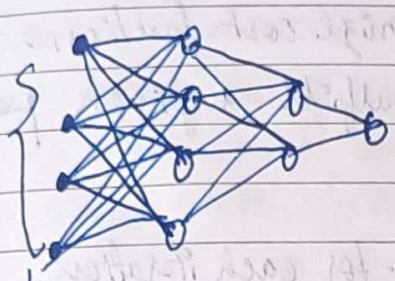
Mathematically
1) Perceptron → 'm' inputs 'n' neurons
2) weights are labelled w_{ij} where $1 \le i \le m$
3) each neuron has it's own weight denoted as W_{ij}, where 'i' is input node no. & 'j' is output node

Limitations-
- uses just binary activation function
- only applicable to linear network
- provides an optimal soln due to supervised learning more training time
- to tackle linear inseparable problems

DOMS | Page No.
Date / /

Multilayer Perceptron

If neural network requires complex decision-making we can create multiple layers of perceptron network.



Layer 1.
(process inputs)
its output is input for layer 2
along with output, weights are also given

- atleast one layer b/w input & output layer.
- hidden layers between I/O layer.
- offers b/w soln to every categorizable issue.
- Input → non linear multilayer network with linear discriminants use.

Need for multilayer networks.

- to solve complex problems
- for large input output data
- intricate info is too complex for single layer to handle.
- multilayers go beyond limitation of single layer.

Linear model

- linear rel" b/w features & output.
- straigh lines of hyperplane
- simpler, interpretable, linear rel".
- can't capture non-linear rel"
- eg → predicting house prices
  - loan approval based on income.

Non-linearities.

- non linear rel"
- can be curves, spirals, or complex shapes
- more flexible, complex data.
- more complex & less interpretable.
- eg → image recognition
  - NLP.

DOMS | Page No.
Date / /

* Gradient Descent Algo (finds best fit line for giving training dataset).

- optimization algo → minimize cost function.
- locate least possible value → fulfill a given fee function.

* Types -

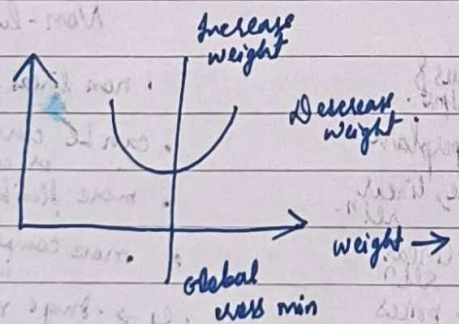
- Batch → processes all training exs. for each iteration
- stochastic → Process 1 training example per iteration. Faster than Batch.
- Mini Batch - Uses examples out of total m training examples are processed per iteration.

- cost function.
- steepest Descent
- Iterative Updates
- convergence
- efficiency
- Variants.
- Local Minima

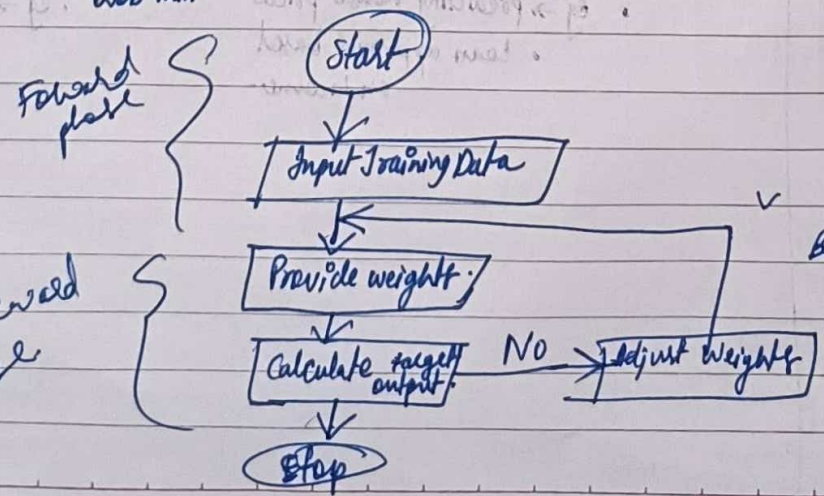
---

* Back propagation -

- algo for supervised learning for ANN.
- keeps adjusting weights of connected neurons with an - to reduce duration of output signal with target output.



- reach global loss mini using backpropagation
- consists of multiple iterations known as epochs.



Features → gradient descent method → case of simple perceptron network with diff unit.

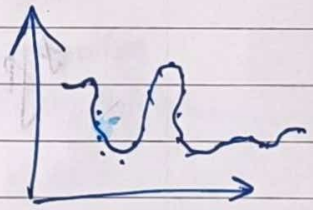
→ weights are calculate in learning period of network:
- feed forward of input training pattern.
- calculation & back propagation of errors.
- updation of weight

Adv → - simple, fast & easy
- only no. of inputs are tuned, not any other parameter.
- flexible & efficient.
- no need for user to learn any special fuction.

Disadv → - sensitive to noisy data & irregularities.
- performance → dependent → data.
- too much time training.
- matrix based approach preffered over mini batch.

---

* Overfitting -

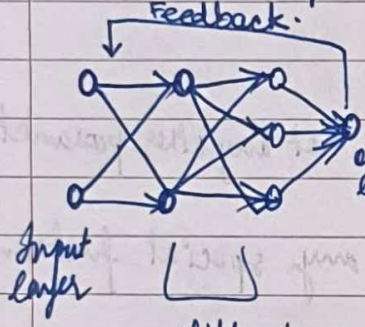


- statistical model fits against its matching data.
- Algo/model can't perform well on unseen data
- Low bias & high variance.
- match input data to target data non existed.
- model → complex enough → match all datapoint & performs well.
- Reasons → • noisy data • training data is too small • large no. of features.

* To avoid Overfitting -
- Cross Validation.
- Train with more data.
- Removing features
- Early stopping the traing
- Regularization
- Ensembling

Recurrent Neural Network (RNN)

- Feed-Forward Neural Network
- Perceptors are arranged to layers, hidden layers are not connected
- All nodes fully connected some layer are not connected with outside world
- need to access previous info in current iteration.
- commonly used in language translation, text to speech, etc.



- Adv → • remember each info through time
  - useful in time series prediction → Feature to remember
  - extend effective pixel neighborhood input cnn.

Disadv → • gradient vanishing & exploding problems.
- training is an difficult task.
- can't process very sequence if
- using tanh or relu as activation function.

Types of RNN → one to one
one to many.
many to one
many to many.

U-4 (DS Honors).

DOMS | Page No.
Date / /

* Hadoop Ecosystem -

- collection of open source s/w projects that facilitate storing, processing & managing big data.
- essentially framework provides tools for various tasks in big data analytics.

Core components -
- HDFS
- YARN
- MapReduce

Benefits, advantages -
- Scalability
- Cost-effectiveness.
- Flexibility
- Open source.
- Fault tollerance.

Disadv -
- complexity
- Learning curve.
- Performance overhead
- Security concerns.
- Vendor Lock-in

Appl^n →
- Log Analysis
- Social Media Analytics.
- Fraud Detection.
- Scientific computing
- Recommendation systems

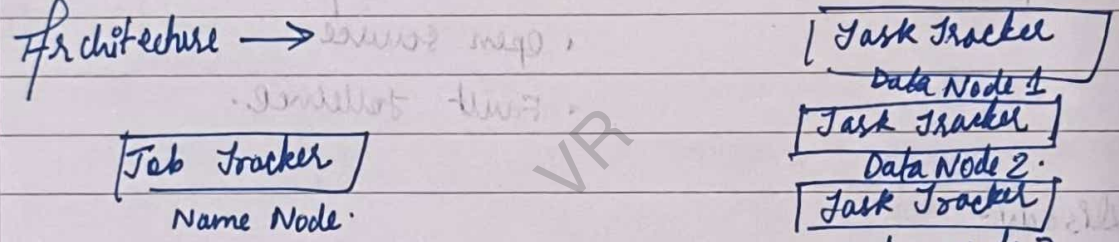
DOMS | Page No.
Date / /

HDFS (Hadoop Distributed File System).

↓
distributed file system designed to run on clusted & low cost hardware.

Characteristics & features of HDFC-
- Fault tollerance
- streaming data access.
- Large Data Sets.
- write once Read many models.
- Highly portable.

Architecture →



HDFS → master slave architecture.
- Name Node → single Name Node → master slaves.
- Data Nodes → multiple Data Nodes → storage attached to it

Adv →
- Scalability
- cost-efficient.
- Reliability
- Fault Tolerance
- Integration.

Disadv →
- Complexity
- Learning Curve
- Overhead
- security Concerns
- Not ideal for small data

DOMS | Page No.
Date / /

* Map Reduce -

- programming Model & software framework.
- developed by Google.

Map() Reduce()

Characteristics →
- Very Large Scale data.
- write once & read many data.
- Map & reduce main operations → simple code
- all maps → reduced operation starts.
- map & reduce → physical processor
- No. of map tasks & reduced tasks.

Adv →
- Scalability
- Cost-efficient
- Fault Tolerance
- Ease of programming
- security & authentication

Disadv →
- Limited Functionality
- Verbosity
- Shuffle Bottleneck.

* Python with Hadoop Streaming -

- allows to write MapReduce Programs.
- leverage power of distributed processing in Hadoop.

Benefits/Adv
→ readability & Maintainability
→ Rich Libraries
→ Familiarity for Python Developers.

Disadv -
- Limited Functionality
- Performance Overhead
- Debugging challenges.

DOMS | Page No.
Date / /

Spark -

- open source framework for large scale data preprocessing.
- addresses some of limitations of traditional MapReduce & significant improvements in terms of speed, performance & flexibility.

Core Concepts -
→ In Memory Processing
→ Unified Platform

Spark Components -
→ Spark core
→ Spark SQL
→ Spark Streaming
→ MLlib
→ GraphX

Adv →
- Speed
- Flexibility
- Ease of Use
- Rich Ecosystem

Disadv →
- Hardware Requirements
- Learning curve
- complexity

Pyspark -

- Is Python API for Apache Spark.
- spark's functionalities accessible & usable from Python Programming Language

Benefits →
- Readability & Maintability
- Familar for Python Development.
- Rich Ecosystem of Python library
- Apache Spark Functionalites.

Disadv →
- Performance Overhead
- Hardware Requirement
- Learning Curve.

U-5 (DWDM Honors).

DOMS | Page No.
Date / /

* Data Warehousing -

- process of constructing & using a data warehouse.
- Integrating data from multiple heterogenous sources → analytical reporting.

Data warehouse → store historical info from mutiple sources to allow you to analyze & report on related data.

Goals → • reporting / analysis • maintain → organization's historical info
• foundation → decision making.

organizations → info → datawarehouse.
- Increasing customer focus
- repositioning products & managing products.
- Analysing operations & looking for profit
- managing customer relationships, making environmental corrections & managing cost of corporate assets.

* Characteristics of DW

- Subject Oriented
- Integrated
- Non-Volatile
- Time Varient
- access & highspeed query.
- end user → time sensative.
- large amt → historical data
- Queries → large amt of data.

DOMS | Page No.
Date / /

* Data Mining :-

process of extracting knowledge & insights from large datasets.

Techniques → • Statistic Analysis • Machine Learning Algo
• Data Visualization Tool.

Data Mining Process -
1) Data collection 2) Data preparation 3) Data Selection.
4) Model Selection 5) Model Building 6) Evaluation 7) Deployment

Benefits	Challenges
→ Uncover hidden meaning	→ Data Quality.
→ Improved Decision making	→ Complexity of Techniques
→ Enchaced customer Understanding	→ Privacy Concern
→ Fraud Detection.	

Characteristics -
- Focus on Pattern & Relationship.
- Predictive Capabilities
- Statistical Foundation
- Machine Learning Integration
- Data Driven Insights.
- Iterative Process
- Focus on Specific Q's
- Subject to Data Quality

DOMS | Page No.
Date / /

Data Analysis Using Hive -

Hive - powerful tool for data warehousing & data analysis on large dataset stored in HDFS.

Key Functionalities
→ SQL like interface
→ Data Warehousing Concepts
→ Scalability
→ Integration with Hadoop Ecosystem.

Process
→ Data loading
→ Schema Definition.
→ Data Querying
→ Data Analysis
→ Data Exploration

Adv -
- Ease of Use
- Scalability
- Flexibility
- cost effectiveness
- Integration with Hadoop Systems

Disadv -
- Limited Functionality
- Performance Overhead.
- Learning Curve.

Characteristics of Hive for Data Analysis
- SQL like Interface
- Data Warehousing Capabilities
- Batch Processing
- Schema on Read
- Distributed Processing

DOMS | Page No.
Date / /

* Data Ingestion -

process of acquiring, importing & preparing data for storage & analysis in data warehouse, data lake or other target systems

Process
→ Data Extraction
→ Data Transformation
→ Data validation
→ Data Loading

Benefits →
→ Improved Data Quality
→ Enhanced data accessibility
→ Faster time to insights
→ Streamlined Analytics workflow

Challenges -
→ Data Volume & Variety
→ Data Quality Issues
→ Real time Data Integration.
→ Security & Compliance.

Data Ingestion Tools
→ ETL (Extract, Transform, Load)
→ Data Integration Platforms
→ Cloud Based Data Ingestion Services

Scalable ML with Spark

Trad" ML struggles
↳ slow training time
↳ large datasets
↳ data silos hinder.

• Adv
↳ Distributed Processing
↳ In Memory Computing
↳ Unified Platforms.

Feature
↳ spark MLlib
↳ spark Dataframes
↳ spark SQL
↳ MLflow Integration

Benefits
→ Faster Training
→ efficient Resource Utilization
→ Improved Model Accuracy
→ Simplified Workflow

Use case -
• Recommendation System
• Fraud Detection
• Customer Segmentation
• Image & Text Recognition.
• Social Network Analysis.

DOMS | Page No.
Date / /

U-6 (DS Honors)

* NLP - (unlocking meaning from Text)

empowers Machines to understand & process human language.

Steps

1) Text Preprocessing
   - Tokenization
   - Normalization
   - Stop Word Removal.

2) Feature Extraction
   - Bag of Words (BOW) (text as a dictionary of word counts)
   - TF-IDF. (weights to words based on their frequency)
   - Word Embeddings. (words as numerical vectors)

3) Applying NLP Techniques
   - Sentiment Analysis
   - Named Entity Recognition (NER)
   - Text Summarization
   - Machine Translation.

Features -
- Automative Text Analysis
- Improves Machine Understanding
- Enables Powerful Appln
- Uncovers Hidden Insights.
- Improves Human Computer Interaction

Disadv
- Data Dependency
- Limited Context Understanding
- Computational Requirements
- Explanability challenges
- Potential for Bias.

Appln of NLP
- Comm" & Interaction — Machine Translation, chatbots, smart replies, sentiment analysis
- Content Creation & Summarization — Text Summarization, machine generated content
- Information Extraction & Retrieval — Question answering systems, name entity recognition & text classification
- Addn appln — voice search, Speech Recognition, paraphrasing text, Legal/financial text processing

DOMS | Page No.
Date / /

* Sentiment Analysis (NLP)-

1) Preprocessing Text (Clean & prepare text data)
2) Feature Extraction (BoW, TF-IDF) Or word embeddings)
3) Model Training → Text samples with predefined sentiment labels.
4) Sentiment Prediction: → Predict sentiment of new, unseen text data.

* Computer Vision

empowers machine to interpret & understand content of digital img & videos.

Strengths
- Automate Visual Analysis
- Improve Machine Perception
- Unlock Valuable Insights
- Enhanced Human Compute Interaction
- Wide range of appln.

Disadv-
- Reliance on Quality Data
- Limited Generalizability
- Computational Demands
- Privacy Concerns
- Vulnerability to Adversial Attacks

Appln-

- Transportation
  - self driving cars
  - advanced driver assistance sys
  - Traffic monitoring & management.
- Security & Surveillance
  - Facial recognition
  - Video Analytics
  - object Tracking
- Retail & Ecomm
  - self checkout sys
  - product recognition
  - Visual search
- Manufacturing & Quality control-
  - Automated Visual Inspection
  - Inventory Management
  - Robot Vision.
- Healthcare & Medical Imaging
  - Medical Image Analysis
  - Surgical Robotics
  - Patient Monitoring

DOMS | Page No.
Date / /

Steps in Computer Vision-

1) Image Preprocessing-

- Resizing / Cropping - adjusting img size for uniformity or focusing on specific regions or interests.
- Noise Reduction - Removing unwanted artifacts or noise from image to improve clarity
- Color Normalization - Adjusting color variations for consistency.

2) Feature Extraction-

- Edge Detection - Identifying boundaries & edge within images.
- Color Histograms - Analyzing distribution of colors in image.
- Local Features - Extracting keypoints or regions of interest with img.

3) Applying Machine Learning Algorithms-

- Image classification - Categorize image content
- Object detection - Identifying & locating objects within image.
- Image Segmentation - Grouping pixels into meaningful segments corresponding to objects or regions.

Object Detection - Spotting objects in image.
- crucial CV appln aims to identify & localize objects within an img or video.
  1) Object proposal - system generates regions in img that contain objects.
  2) Feature Extraction - Features are extracted from regions.
  3) classification & Bounding Box Prediction - ML classifies each region, predicting whether it contains an object & its bounding box.