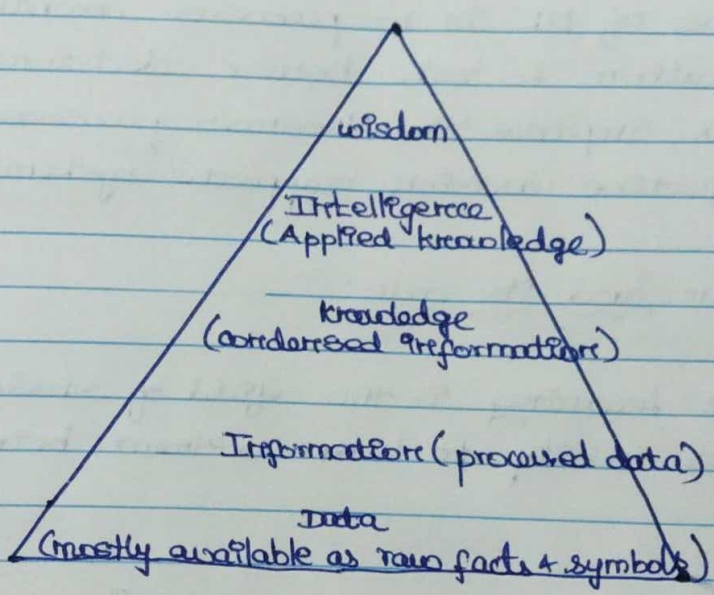
Unit - I

Popularity of ML due to three reasons:

* High Volume of available data to manage
* cost of storage has reduced
* Availability of complex algorithms.

A knowledge pyramid structure:



What is data?

All facts are data.

Data can be numbers or text that can be processed by a computer.

Processed data is called Information.

This includes patterns, associations or relationships among data.

Condensed information is called knowledge.

An actionable form of knowledge is called Intelligence.

The objective of knowledge powered is wisdom that represents the maturity of mind, so far exhibited by only humans.

The objective of ML is to process critical data for organization to take better decisions to design new products, improve the business processes and to develop effective decision support systems.

ML Definition and its role:

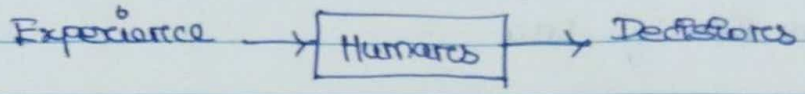
Machine learning is the field of study that gives the computers ability to learn without being explicitly programmed.

The focus of AI is to develop Intelligent systems by using data-driven approaches, where data is used as an input to develop Intelligent models.

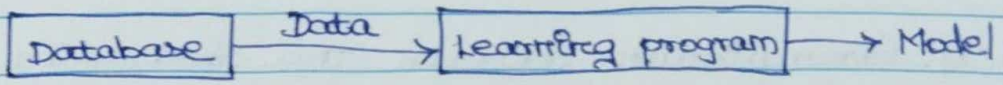
The models can be used to predict new inputs.

The aim of ML is to learn a model or set of rules from the given datasets automatically so that it can predict the unknown data correctly.

(a) A learning system for humans



(b) A learning system for ML



Computer make models based on extracted patterns in the input data and use those data-filled models for prediction and to take decisions.

Quality of data determines the quality of experience and therefore, the quality of the learning system.

A model:

A model is an explicit description of patterns within the data in the form of:

1). Mathematical equations
2). Relative diagram like tree/graphs
3). Logical if/else rules, or
4). Grouping called clusters.

In systems, experience is gathered by following steps:

1). collection of data

2). Once the data is gathered, abstract concepts are formed out of that data. Abstraction is used to generate concepts.

3). Generalization converts the abstraction into actionable form of intelligence.

Generalization involves ranking of concepts, inferencing and formation of heuristics, are actionable aspect of intelligence.

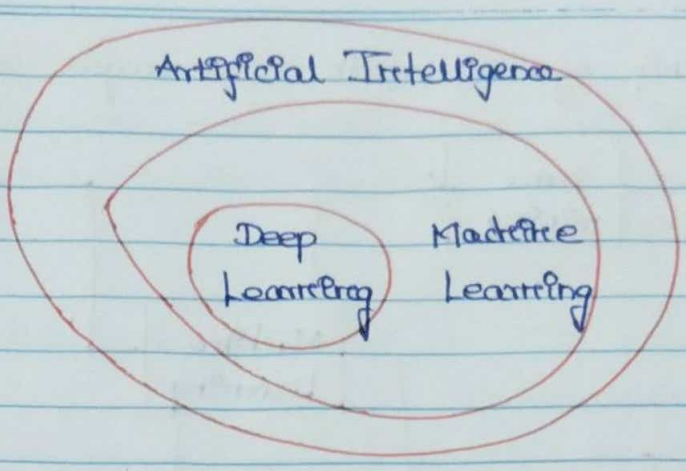
4). Course Corrections are done by taking evaluation measures.

Evaluation checks the thoroughness of the models and to-do course correction, if necessary, to generate better formulations.

Artificial Intelligence and Machine Learning:

ML uses the concepts of Artificial Intelligence, Data sciences and statistics.

ML is the subbranch of AI, whose aim is to extract the patterns for predictions.



Deep learning:

DL is the subbranch of machine learning.

In DL, the models are constructed using neural network technology.

Neural networks are based on the human neuron models.

Many neurons form a network connected with the activation functions that trigger further neurons to perform tasks.

Data science:

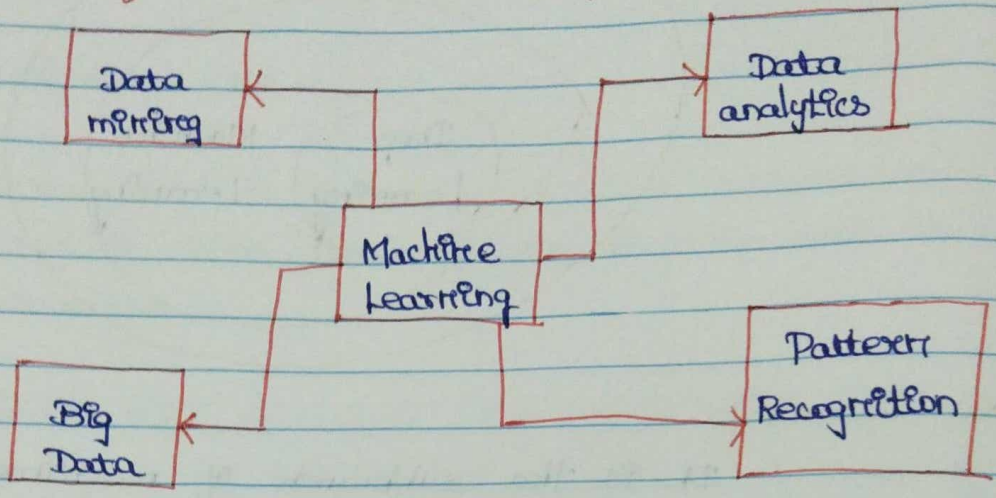
DL is a branch of data science.

Data science deals with gathering of data for analysis.

It includes:

* Big data
* Data mining
* Data analytics
* Pattern recognition

Relationship of ML with other major fields:



Data mining aims to extract the hidden patterns that are present in the data
whereas ML aims to use it for prediction.

Data Analytics aims to extract useful information knowledge from raw data.

ML algorithms to extract the features for pattern analysis and pattern classification.

Statistics:

Statistics is a branch of mathematics that has a solid theoretical foundation.

Initially, statistics sets a hypothesis and perform experiments to verify and validate the hypothesis in order to find relationship among data.

Statistics	ML
* Mathematics intensive and models are often complicated equations involving many assumptions.	* ML has less assumptions and requires less statistical knowledge.
* Statistical methods are coherent and rigorous.	* It often requires interaction with various tools to automate the process of learning.

Pattern and model:

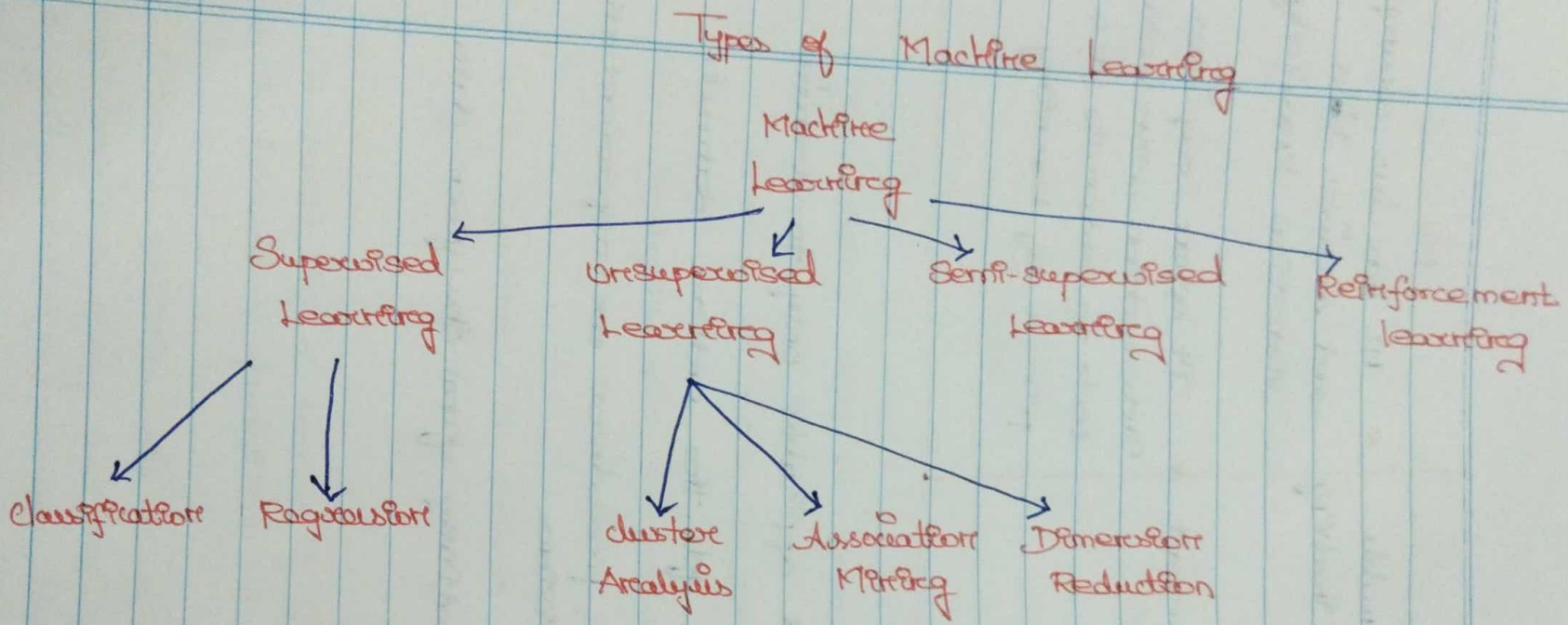
pattern is local and applicable only to certain attributes.

Model is global and fits the entire datasets.

Model can be a formula, procedure or representation that can generate data decisions.

Model is generated automatically from the given data.

Types of Machine Learning



Data represented in the form of Table.

Row of the table represents a data points.
Column of the table as attributes

Features are attributes or characteristics of an objects

Out of all attributes, one attribute is important and called label.

Label is the feature that aim to predict.

Supervised Learning	Unsupervised learning
* There is a supervisor component	* No supervisor component
* Uses labelled data	* Uses unlabelled data.
* Assigns categories or labels	* Performs grouping process such that similar objects will be in one cluster.

Supervised Learning uses labelled datasets.

Unsupervised learning is by self-instruction.

The process of self-instruction is based on the concept of trial and error.

Semi-supervised learning:

It is applicable where the datasets has a large collection of unlabelled data and some labelled data.

This algorithms use unlabelled data by assigning a pseudo-label.

Then the labelled and pseudo-labelled dataset can be combined.

Reinforcement learning:

This technique is reward-based, goal-oriented algorithms

RL allows the agent to interact with the environment to get rewards.

The agent can be human, animal, robot or any predoperdort program.

The rewards enable the agent to gain experience

The agent aims to maximize the reward.

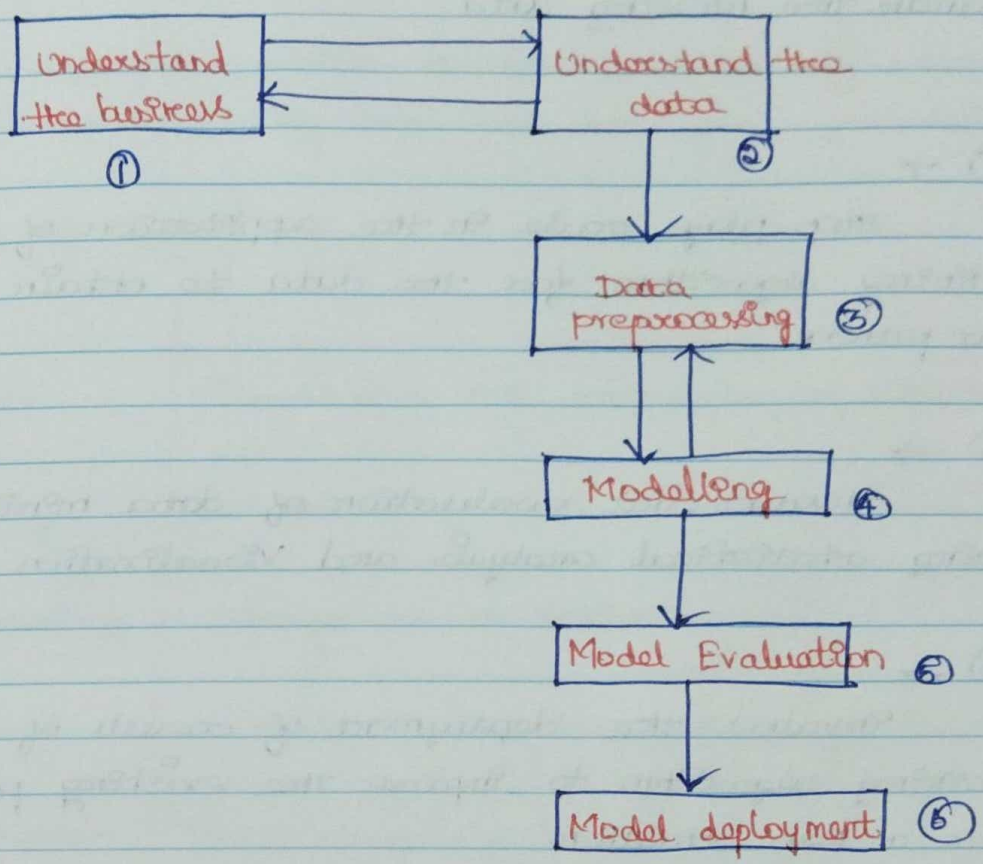
The reward can be positive or negative (punishment)

when the rewards are more, the behaviour gets reinforced and learning become possible.

Machine learning process:

Emerging process model for the data mining solutions for business organization is CRISP-DM.

Cross Industry Standard Process - Data Mining



① → Involves the formulation of the problem statement for the data mining process

② → steps like data collection, study of characteristics of data, formulation of hypothesis and matching of patterns to the selected hypothesis

③ →

producing the final datasets by cleaning the raw data and preparation of data for the data mining process.

Suitable strategies should be adopted to handle the missing data.

④ →

Step plays crucial role in the application of data mining algorithm for the data to obtain a model or pattern.

⑤ →

Involves the evaluation of data mining results using statistical analysis and visualization methods.

⑥ →

Involves the deployment of results of data mining algorithm to improve the existing process or for a raw situation.

Challenges of Machine Learning:

* Huge data
* High Computation power
* Complexity of the algorithms
* Bias/Variance
* Overfitting and Underfitting

Variance is the error of the model. This leads to a problem called bias / variance tradeoff.

A model that fits the training data correctly but fails for test data, i.e. general lacks generalisation is called overfitting.

The reverse problem is called underfitting where the model fails for training data but has good generalisation.

ML Applications Survey Table:

Domains	Applications
Business	Predicting the bankruptcy of a business firm
Banking	Prediction of bank loan defaulters
Image processing	Image search engines
Audio/Voice	Chatbots like Alexa, Microsoft Cortana
Telecommunication	Trend analysis & identification of bogus calls
Marketing	Retail sale analysis
Natural language Translation	Google translate, sentiment analysis
Web analysis & services	Viruses, detection of e-mail spams, search engine like google

Supervised learning has two methods:

Classification

Regression

Classification:

The input attributes are independent variables

the target attributes called label are dependent variable

The relationship between the input and target variables represented in the form of structure called classification model.

The focus of classification is to predict the 'label' that is in discrete form (a set of finite values)

In classification, learning takes place in two stages:

First stage: Training stage

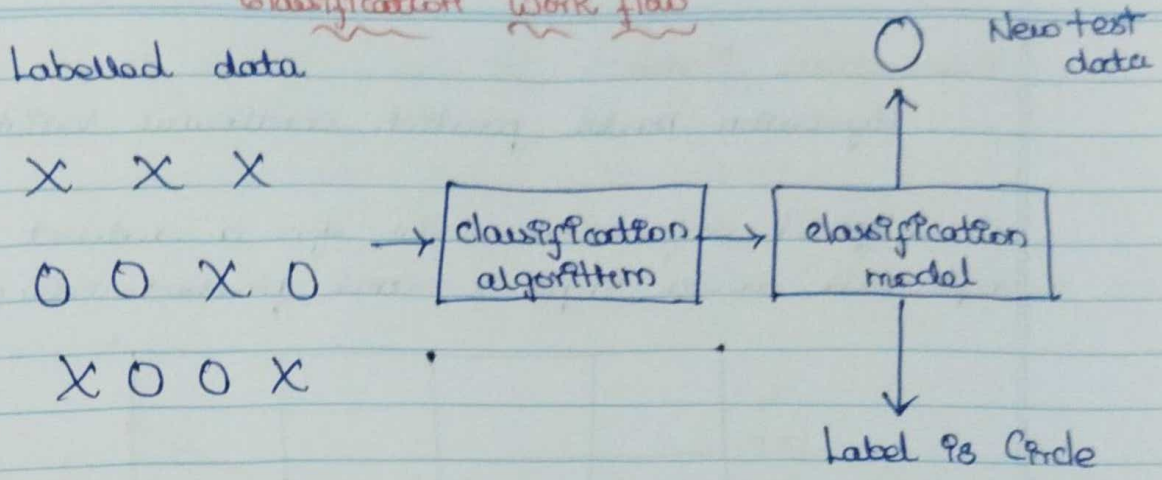
Takes a labelled dataset and starts learning.

After the training set, samples are processed and the model is generated

Second stage: Testing Stage

The constructed model is tested with test or unknown samples and assigned a label.

Classification Work flow



Classification model can be classified as generative models and discriminative models.

Generative models deals with the process of data generation and its distribution.

Ex: Probabilistic models

Discriminative models do not care about the generation of data. Instead, concentrate on classifying the given data.

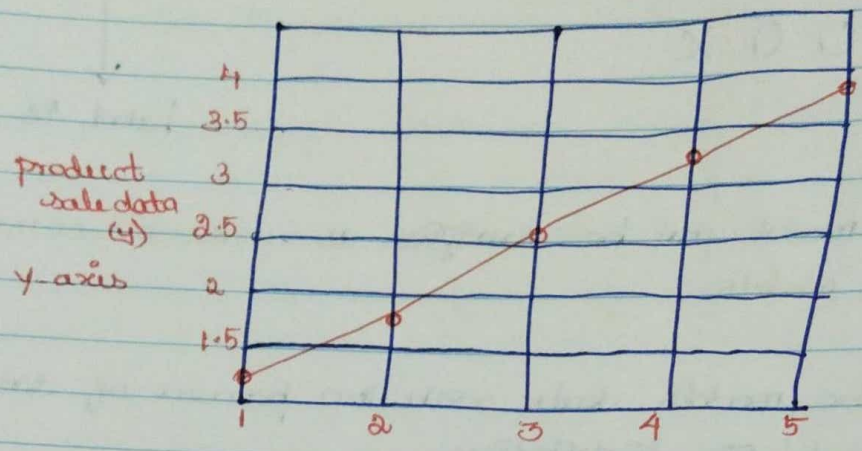
Key algorithms of classifications:

- Decision Tree
- Random Forest
- Support Vector Machines
- Naive Bayes
- Artificial neural networks

Regression Models:

Regression model predict continuous variables like price

A fitted regression model for a dataset that represents weeks input x and product sale y.



A regression model of the form $y = ax + b$

x is the independent variable may be one or more attributes

y is the dependent variable

Regression model takes input x and generates a model in the form of fitted line of the form $y = f(x)$

The advantage of this model is that predictions for product sales (y) can be made for unknown week data (x).

Linear regression takes the training set and tries to fit it with a line

product sales = 0.66 x week + 0.54

0.66 and 0.54 are all regression coefficients that are learnt from data.

Difference between classification and Regression:

Both have a supervisor and concept of training and testing.

Regression model predicts continuous variables

Classification concentrates on assigning labels

Unsupervised Learning two methods:

Cluster Analysis

Dimensional reduction

Cluster Analysis:

aims to group objects into disjoint clusters or groups

It clusters objects based on its attributes.

key clustering algorithms:

K-means algorithm,

Hierarchical algorithm

Dimensional reduction: takes a higher dimension data as input and outputs the data in lower dimension by taking advantages of variance of the data.

It is a task of reducing the datasets with few features without losing the generality.

—x—x—

Bias → error due to overly simplistic assumptions in learning algorithm

→ model has poor performance on both training and test data → high bias → underfitting

Variance → error due to model sensitivity to fluctuations in the training data.

high variance → model learns the training data's noise and random fluctuations rather than the underlying pattern.

model perform well on the training data, poorly on testing data - overfitting

Underfitting - high bias and low variance

Overfitting - high variance and low bias

Good fitting - low bias, low variance