Machine learning - Module 1

1.1 Introduction: Need for ML
1.2. Machine learning explained
1.3. Machine learning in relation to other fields
1.4 Types of Machine learning
1.5. challenges of ML.
1.6. Machine learning process
1.7 machine learning applications
1.8. Understanding data 1: Introduction
1.9. Big data Analysis framework.
1.10. Descriptive statistics
1.11. Univariate data analysis & visualization

Need for machine learning:

Business organizations use huge amount of data for their daily activities.

Earlier, the full potential of the huge data was not utilized due to 2 reasons.

1. Data being scattered across different archieve systems & organizations not being able to integrate these sources fully.

2. The lack of awareness about software tools that could helps to unearth the useful information from data.

Machine learning has become popule for 3 reasons

1. High volume of available data to manage:
Big companies such as facebook, twitter & youtube generate huge amount of data that grows at a phenomenal rate.

2. Reduction in cost of storage:
The Hardware cost has also dropped ∴ it is easier now to capture, process, store distribute & transmit the digital information

3. Availability of complex algorithms:
with the advent of deep learning, many algorithms are available for machine learning.

The knowledge pyramid.

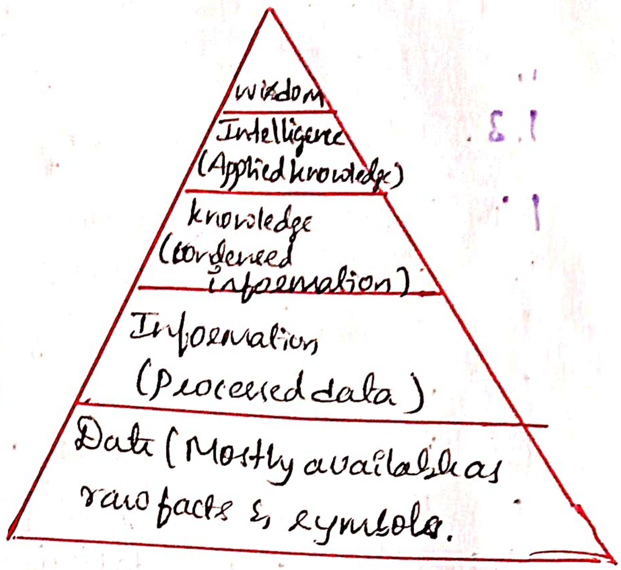

Wisdom
Intelligence (Applied knowledge)
knowledge (condensed information)
Information (Processed data)
Data (Mostly available as raw facts & symbols.

fig: The knowledge pyramid

Data: can be numbers/text that can be processed by computer.

Today, organizations are accumulating vast & growing amounts of data with data sources such as flat files, databases (or) datawarehouses in different storage formats

Information: Processed data which includes patterns, associations (or) relationships among data

E.g: Sales data can be analyzed to extract information. [Fast selling product]

Knowledge: condensed information

E.g: The historical patterns & future trends obtained in the above sales data can be called knowledge.

Unless knowledge is extracted, data is of no use

knowledge is not useful unless it is put into action.

Intelligence: Applied knowledge for actions. An actionable form of knowledge is called intelligence

computer systems have been successful till this stage

The ultimate objective of knowledge pyramid is wisdom that represents the maturity of mind, that is exhibited only by humans

NOTE:-

The objective of machine learning is to process archival data for organizations to take better decisions to design new products, improve the business processes & to develop effective decision support systems

1.2. Machine learning explained

ML is a sub-branch of Artificial Intelligence

ML is the field of study that gives the computers ability to learn without being explicitly programmed. - Arthur Samuel.

The focus of AI is to develop intelligent models

The models can then be used to predict new inputs.

Thus, the aim of ML is to learn a model / set of rules from the given data set automatically so that it can predict the unknown data correctly.

As humans take decisions based on an experience, computers make models based on extracted patterns in the i/p data & then use these data-filled models for prediction & to take the decisions

For computers, the learnt models is equivalent to human experience

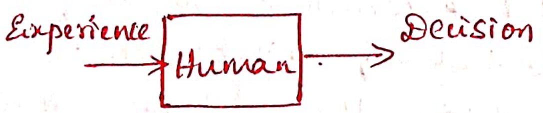


fig (a): A learning system for human

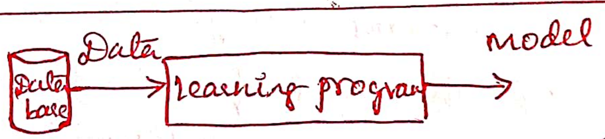


fig (b): A learning system for machine learning (ML)

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

In statistical learning, the relationship b/w the i/p 'x' & o/p 'y' is modeled as a function in the form $y = f(x)$. where f is the learning function that maps the i/p x to o/p y.

Learning of function f is the crucial aspect of forming a model in statistical learning.

The learning program summarizes the raw data in a model.
Model is an explicit description of patterns within the data in the form of

1. Mathematical equation
2. Relational diagrams like trees/graphs
3. Logical if/else rules
4. Grouping called clusters.

NOTE :- Model can be a formula, procedure/representation that can generate data decisions.

A computer program is said to learn from experience 'E', with respect to task 'T' & some performance measure 'P', if its performance on 'T' Measured by 'P' improves with experience E. ~~The~~ Tom Mitchell's definition of Machine learning

The important components of ML are experience 'E', task 'T' & performance 'P'

E.g: The task 'T' could be detecting an object in an image.
The Machine can gain the knowledge of object using training dataset of thousand of images. This is called experience E. So, the focus is to use this experience 'E' for this task of object detection 'T'.
The ability of the system is to detect the object is measured by performance measures like precision & recall
Based on the performance measures, course correction can be be done to improve the performance of the system.

Models of computer systems are equivalent to human experience. & Experience is based on data
Humans gain experience by various means. They gain knowledge by rote learning. They observe others & imitate it.
Humans gain a lot of knowledge from teachers & books.
We learn many things by trial & error.
Once the knowledge is gained, when a new problem is encountered, humans search for similar past situations & then formulate the heuristic & use that for predictions. But, in systems, experience is gathered by the following steps

In systems, experience is gathered by the following steps

1. Collection of data.
2. Once data is gathered, abstract concepts are formed out of that data.
   Abstraction is used to generate concepts. This is equivalent to humans idea of objects
3. Generalization converts the abstraction into an actionable form of intelligence. It can be viewed as ordering of all possible concepts.
4. Heuristics normally works, But occasionally, it may fail too.
   The course correction is done by taking evaluation measures.
   Evaluation checks the thoroughness of the models & to do course. correction, if necessary, to generate better formulations

1.3. Machine learning in relation to other fields.

ML uses the concepts of AI, Data Science & statistics primarily.
It is the resultant of combined ideas of diverse fields.

1.3.1. Machine learning & Artificial Intelligence

ML is an important branch of AI. The aim of AI is to develop intelligent agents.
An agent can be a robot, humans (or) any autonomous systems.
ML is the subbranch of AI, whose Aim is to extract the patterns for prediction

Deep learning is a subbranch of ML.
In DL, models are constructed using Neural N/W technology.
Neural N/ws is based on human neuron models.
Many Neurons form a n/w connected with the activation functions that trigger further neurons to perform tasks

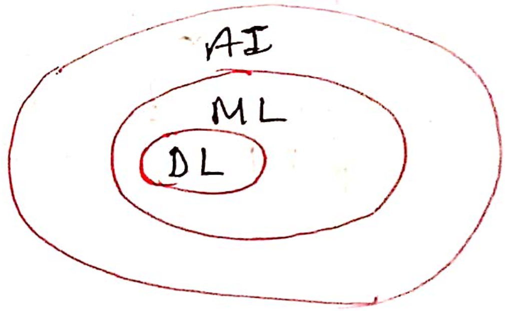


Fig: Relationship of AI with machine learning

1.3.2. Machine learning, Data Science, Data mining & Data Analytics

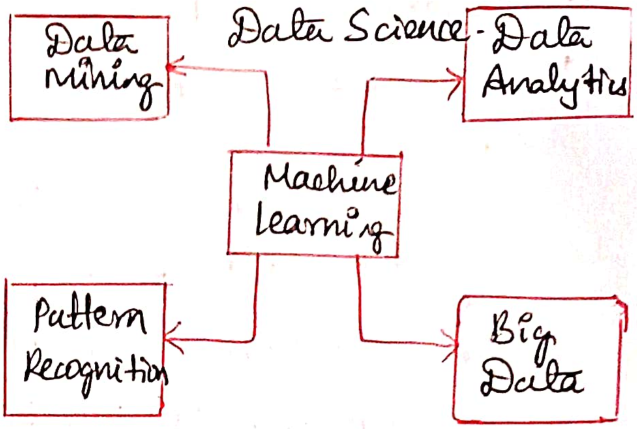


fig: Relationship of ML. with other major fields.

Data Science is an 'umbrella'. term that encompasses many fields.

Machine learning starts with data therefore, DS & ML are interlinked

ML is a branch of Data Science.

DS deals with gathering of data for analysis.

It is a broad field that includes.

* Bigdata
* Data mining
* Data Analytics
* Pattern Recognition.

Big data:

Data science concerns about collection of data.

characteristics of Big data are as follows

1. volume: Huge amount of data is generated by big companies like facebook, Twitter & youtube
2. Variety: Data is available in variety of forms like images, Videos & in different formats
3. velocity: It refers to the speed at which the data is generated & processed.

Bigdata is used by many machine learning algorithms for applications such as ~~so~~ language translation & image recognition.

Bigdata influences the growth of Deeplearning which is a branch of ML. deals with constructing models using neural networks.

Data Mining

It aims to extract the hidden patterns that are present in the data.

Machine learning aims to use it for prediction.

Data Analytics:

It Aims to extract useful knowledge from crude data.

There are different types of analytics.

Predictive data analytics is used for making predictions.

ML is closely related to this branch of analytics & shares almost all algorithms.

Pattern recognition:

It is an engineering field. It uses ML algorithms to extract the features for pattern analysis & pattern classification.

1.3.3. Machine learning & statistics

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

* Statistical methods look for regularity in data called patterns.
* Initially, statistics sets a hypothesis & performs experiments to verify & validate the hypothesis inorder to find relationship among data.
* Statistics requires knowledge of the statistical procedures & the guidance of a good statistician.
* Statistical methods are developed in relation to the data being analysed

Machine learning has less assumptions & requires less statistical knowledge. But, it often requires interaction with various tools to automate the process of learning.

1.4. Types of Machine learning

There are 4 types of ML.

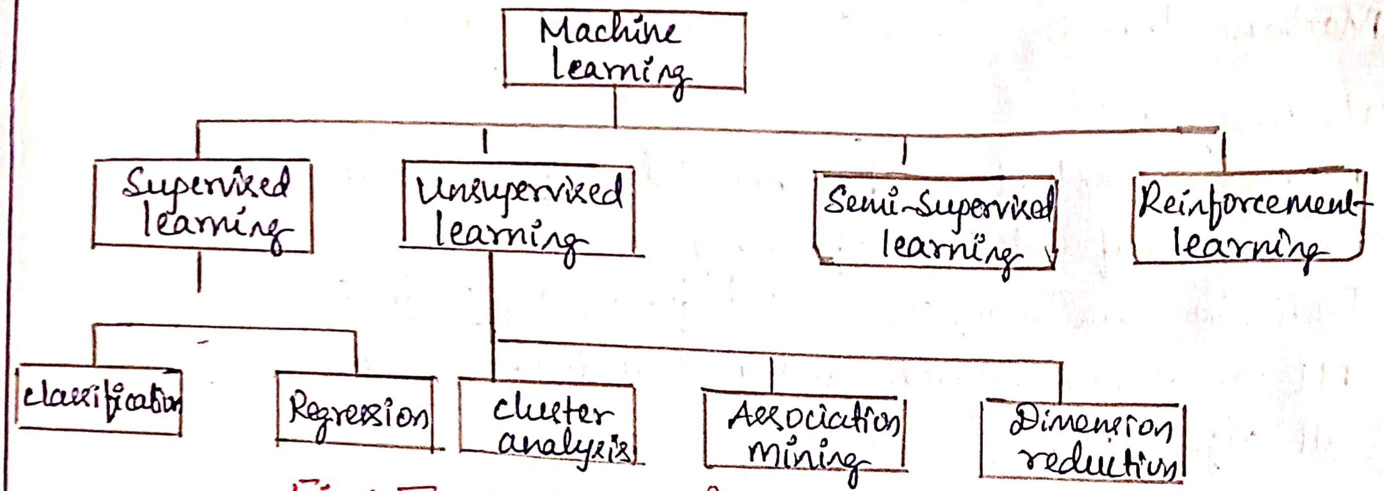


Fig: Types of Machine learning

Labelled & Unlabelled data.

Data is represented in the form of a table.

Data can also be referred to as a data point, sample (or) an example.

Each row of the table represents a data point.

Features are attributes / characteristics of an object.

columns of the table are attributes.

Each attribute is known as label: which is to predict.

There are 2 types of data

labelled data

Unlabelled data .

Table: Iris flower dataset

Sl No	length of Petal	width of Petal	length of Sepal	width of Sepal	class
1	5.5	4.2	1.4	0.2	Setosa
2	7	3.2	4.7	1.4	Versicolor
3	7.3	2.9	6.3	1.8	Virginica

labelled data :

let us consider the Iris flower data set

It has 4 attributes — length & width of sepals & petals.

The target variable is ~~cal~~ called class.

There are 3 classes - Iris setosa, Iris Virginica, Iris versicolor.

A data need not be always numbers.

It can be images/video frames.

DNN can handle images with labels.

DNN takes images of dogs & cats with labels for classification

In unlabelled data, there are no labels in the data set

input	label
	Dog
	Cat

(a) labelleld dataset

(b) unlabelled data set

1.4.1 Supervized learning:

Supervized algorithms use. labelled dataset.

In supervized learning algorithms, learning takes place in 2 stages.

* First stage, the teacher communicates the information to its student. The student receives the information & understands it. During this stage, the teacher has no knowledge of whether the information is grasped by the student

* Second stage, the teacher asks the student a set of questions to find out how much information has been grasped by the student. Based on these questions, the student is tested. & the teacher informs the student about his assessment.

This kind of learning is called Supervized learning.

2 Methods of Supervised learning.

①. classification

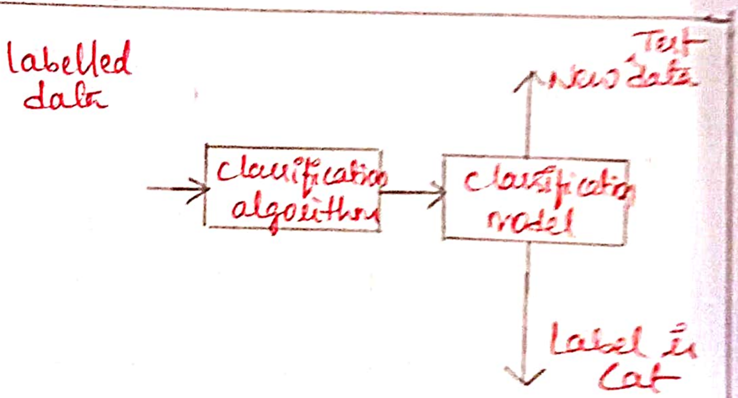
②. Regression.

Classification;

It is a supervised learning method.

The input attributes are independent variables

The target attributes (Label) ~~are~~. are dependent variable.



fig(a) An example of classification system

The relationship b/n the i/p & target variable is represented in the form of a structure called as classification model.

classification Model is to predict the label in a discrete form. (Finite set of values).

- Fig(a) shows the classification system, which takes a set of labelled data images such as dogs & cats to construct a model that can later be used to classify the unknown test image data.

In classification learning takes place in 2 stages.

1. Training stage.
2. Testing stage

1. Training stage - The learning algorithm takes a labelled data set & starts learning. After the training set, samples are processed & the model is generated

2. Testing stage - The constructed model is tested with test or unknown sample & assigned a label. This is the classification process.

* In fig(a), Initially the classification learning algorithm learns with the collection of labelled data & constructs the model. Then, a test case is selected & the model assigns a label.

* Similarly, in the case of Iris flowerdataset, if the test is given as (6.3, 2.9, 5.6, 1.8, - -?-), the classification will generate the label for this. This is called classification

* Eg of Classification is Image recognition which includes classification of diseases like cancer, classification of plants etc

* The classification models can be categorized based on the implementatio technology such as decision trees, probabilistic methods, distance measure Soft computing methods

* Classification models can also be ~~model~~ classified as generative & discriminative models.

* Generative models deal with the process of data generation & its distribution.

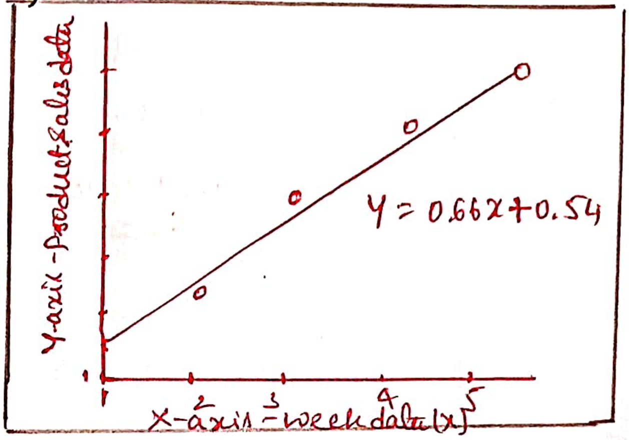
* Probabilistic models are examples of generative models. Discriminative models do not care about the generation of data, instead they simply concentrate on classifying the givendata.

* Some of the key algorithms of classification are

1. Decision tree
2. Random forest
3. Support vector Machines
4. Naive Bayes
5. ANN & DLN like CNN)

Regression models

It predict continuous variables like price. [It is a number]



The regression model takes i/p x & generates a model in the form of a fitted line of the form $\boxed{y = f(x).}$

where x is the independent variable.
y is the dependent variable.

In fig(6), linear regression takes the training set & tries to fit it with a line

Product sales = 0.66 x week + 0.54.

Here 0.66 & 0.54 are all regression co-efficients that are learnt from data.

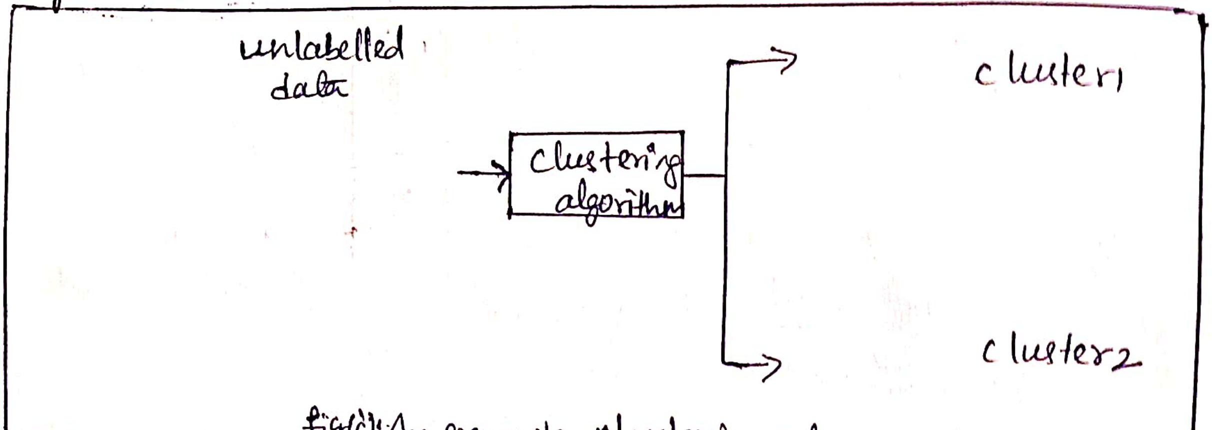
1.4.2. Unsupervised learning:

* There is no supervisor (Or) teacher components.
* Self instruction is the most common kind of learning process.
* This process of self instruction is based on the concept of trial & error.
* Here, program is supplied with objects, but labels are defined.
* The algorithm itself observes the examples & recognizes the patterns based on the principles of grouping.
* Grouping is done in ways that similar objects form the same group.
* Cluster Analysis & dimensional reduction algorithms are examples of unsupervised algorithms.

Cluster Analysis:-

* It is an example of unsupervised learning.
* It aims to group objects into disjoint clusters/groups.
* Cluster analysis clusters objects based on its attributes.
* All the data objects of the partitions are similar in some aspect & vary from the data objects in the other partitions significantly.

Some of the examples of clustering processes are.
Segmentation of aregion of interest in an image, detection of abnormal growth in a medical image, determining clusters of signatures in a gene database.



fig(c): An example clustering scheme.

fig(c) shows the clustering scheme, where clustering algorithm takes a set of dogs & cats images & groups it as 2 clusters dogs & cats.

It can be observed that the samples belonging to a cluster are similar & samples are different radically across clusters.

Some of the key clustering algorithms are.

* k-means algorithm
* Hierarchical algorithms.

Dimensionally Reduction.

It is an unsupervised algorithms. It takes a higher dimension data as i/p & o/ps the data in lower dimension by taking advantage of the Variance of the data.
It is a task of reducing the dataset with few features without losing the generality

Differences b/w. Supervised & Unsupervised learning

Sl No	Supervised learning	Unsupervised learning
1	There is a supervisor component	NO supervisor component
2	Uses labelled data	Uses unlabelled data
3	Assigns categories/labels	Performs grouping process such that similar objects will be in one cluster

1.4.3 Semi-Supervised learning.

There are circumstances where the dataset has a huge collection of unlabelled data & some labelled data labelling is a costly process. & difficult to perform by the humans

Semi-supervised algorithms use unlabelled data by assigning a pseudo label. Then, the labelled & pseudo labelled dataset can be combined.

1.4.4 Reinforcement learning:

It mimics human beings. Like human beings use ears & eyes to perceives the world & take actions, reinforcement learning allows the agent to interact with the environment to get rewards.

The agent can be human, animal, robot (or) any independent program.

The rewards enable the agent to gain experience.

The agent aims to maximize the reward.

The reward can be positive (or) negative.

When the rewards are more, the behavior gets reinforced & learning becomes possible

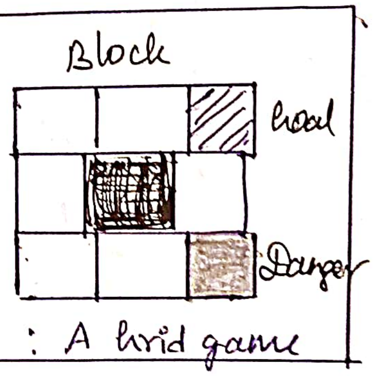


fig : A Grid game

1.5. Challenges of Machine learning.

Some of the challenges of ML are

① Problems:- ML can deal with the "well-posed" problems where specifications are complete & available.
Computers can't solve "ill-posed" problems

② Huge data: Availability of a quality data which should be large & should not have data problems such as missing data / Incorrect data

③. High computation power:-
With the availability of big data, the computational resource requirement has also increased.
Systems with Graphics processing Unit (GPU)/Tensor processing unit (TPU) are required to execute ML algorithm.
ML tasks have become complex & hence time complexity has increased & that can be solved only with high computing power.

④ Complexity of the algorithms:
The selection of algorithms, describing the algorithms, appl^n of algorithms to solve ML task & comparision of algorithms have become necessary for ML/ data scientist now.

⑤ Bias/Variance :- It is the error of the model. This leads to a problem called variance/bias tradeoff.
A model that fits the training data correctly but fails for test data in general lacks generalization, is called overfitting.
The reverse problem is called underfitting where the model fails for training data but has good generalization
Overfitting & underfitting are great challenges for ML algorithms

1.6. Machine learning process:-

ML process involves 6 steps.

1. Understand the business:

This step involves understanding the objectives & requirements of the business organization.

Generally, a single data mining algorithm is enough for giving the solution.

This step also involves the formulation of the problem statement for the data mining process

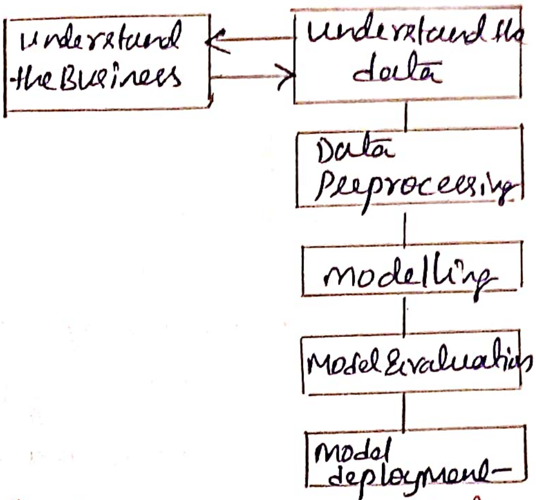


fig (d): A Machine learning/ Data mining Process

2. Understanding the data:

It involves data collection, study of the characteristics of the data, formulation of hypothesis, matching of patterns to the selected hypothesis

3. Preparation of data:

It involves producing the final dataset by cleaning the raw data & preparation of data for the data mining ~~data~~ process.

The missing values may cause problems during both training & testing phases

Missing data forces classifiers to produce inaccurate results.

Hence suitable strategies should be adopted to handle missing data.

4. Modelling:

This step plays a role in the application of data mining algorithm for the data to obtain a model/pattern.

5. Evaluate:

It involves the evaluation of the data mining results using statistical analysis & visualization methods

The performance of the classifier is determined by evaluating the accuracy of the classifier.

6. Deployment-

It involves the deployment of results of the data mining algorithm to improve the existing process/for a new situation

1.7 Machine leaening ~~process~~ Applications :

ML technologies are used widely in different domains

1 Sentimental analysis : It is an application of NLP where the words of documents are converted to sentiments like happy, sad & angry which are captured by emotions effectively.

e.g : For movie reviews/ product reviews, 5 stars/1 star are automatically attached using sentiment analysis programs.

2. Recommandation systems :

Systems which make personalized purchases possible.

e.g : Amazon recommends users to find related books/books bought by people, who have the same task like you & Netflix suggests shows/ related movies of your taste.

The recommendation systems are based on ML.

3. Voice assistants :

Products like amazon Alexa, Microsoft cortana, Apple siri & google assistant are all examples of voice assistants.

They take speech commands & perform tasks. These chatbots are the result of ML technologies.

4. Technologies like google maps & those used by Uber are all examples of ML which offer to locate & navigate shortest paths to reduce time.

1.8 Understanding data - I

Data is available in different data sources like flat files, databases (or) data warehouses.

It can either be an operational data / a non-operational data.

Operational data is used in normal business procedures & processes.

E.g: Daily sales data is operational.

Non-operational data is a kind of data used for decision making.

Elements of Big data :-

Data whose volume is less & can be stored & processed by a small scale computer is called "small data". These data are collected from several sources, integrated & processed by a small-scale computer.

Data whose volume is much larger than small data - Big data

Characteristics of Big data are.

1. Volume : Small Data is measured in terms of GB, TB.
   Big data is measured in terms of petabyte (PB) & Exabyte (EB)
   1 Exabyte - 1 million terabyte.

2. Velocity : The fast arrival speed of data & its increase in data volume is noted as velocity.
   It helps to understand the relative growth of big data & its accessibility by users, systems & applications

3. Variety : The variety of big data includes.
   1. Form :- text, graph, audio, video, maps.
   2. Function :- data from various sources like human conversation, transaction records & old archieve data
   3. Source of data : open / public data, Social media data & multimodal data.

4. Veracity of data :- Conformity to the facts, Truthfulness, believability & confidence in data. There may be many sources of error such as technical errors, typographical errors & human errors.

5. Validity : Accuracy of the data for taking decisions / any other goal that are needed by the given problem

6. Value : It is the characteristic of big data & is extracted from the data & its influence on the decisions that are taken based on it.

1.8.1. Types of data :

There are 3 kinds of data.

① Structured data

2. Unstructured data

3. Semi-structured data.

Structured data :-

Data is stored in an organized manner such as database & it is available in the form of Table.

The data can also be retrieved in an organized manner using tools like SQL.

The structured data frequently encountered in ML are listed below

1. Record data.
2. DataMatrix
3. Graph data
4 Ordered. data
   - Temporal data-
   - Sequential data
   - Spatial data

Records data :-

It is a ~~set~~ collection of measurements taken from a process collection of objects in a dataset & each object has a set of measurements

The measurements can be arranged in the form of a matrix.

Rows in the matrix represent an object & can be called as entities, cases/ records.

The columns of the dataset are called attributes, features/fields.

The table is filled with observed data

Label is the term used to describe the individual observations.

Data Matrix :-

It is a variation of the record data type because it consists of numeric attributes.

The standard matrix operations can be applied on these data.

Graph Data : ,

It involves the relationships among objects.

E.g : A web page can refer to another web page. This can be modeled as a graph.

The nodes are webpages & the hyperlink is an edge that connects the nodes.

Ordered data :-

Ordered data objects involves attributes that have an implicit order among them.

1. Temporal data :

It is the data whose attributes are associated with time.

E.g: the customer purchasing patterns during festival time is sequential data.

Time series data is a special type of sequence data where the data is a series of measurements over time

2. Sequence data :- It is like sequencial data but not have time stamps. This data involves the sequence of words/letters.

E.g: DNA data is a sequence of 4 characters - A T G C.

3. Spatial data :

It has attributes such as positions/areas.

E.g: maps are spatial data where the points are related by location

Unstructured data :-

It includes video, image & audio. It also includes textual document, programme & blog data.

It is estmated that 80% of the data are unstructured data.

Semi Structured data :

Partially structured & partially unstructured.

E.g: XML (Extensible Markup language] -

JSON (Java Script object Notation]

RSS (Really simple syndication)

Hierarchical data

5.1.2 Data storage & representation

The goal of data storage Management is to make data available for analysis.

There are different approaches to organize & manage data in storage files & systems from flat file to data warehouse.

Some of them are listed below.

Flat files: These are the simplest & most commonly available data source

Some of the popular spreadsheet formats are listed below.

CSV file: Comma separated value. file & are used by spreadsheet & database applications.

The first row may have attributes & the rest of the rows represent the data

TSV file :- Tab separated files where values are separated by Tab.

Both CSV & TSV files are generic in nature & can be shared.

Database system:-

It consists of database files & a database management system.

Database file contain original data & metadata.

DBMS aims to manage data & improve operator performance by including various tools like database administrator, query processing & transaction manager.

A relational database consists of set of tables.

Tables have rows & columns.

The columns represent the attributes.

The rows represent tuples.- corresponds to either an object/a relationship b/n objects.

A user can access & manipulate the data in the database using SQ

different types of databases are as follows

1. Transactional database:- collection of transactional records. Each record is a transaction.
   A transaction may have a timestamp., identifier & a set of items, which may have links to other table.

2. Time series database:- Stores time related information like log files
   This data represents the sequences of data, which represent values/ events obtained over a period (or) repeated time span.

3. Spatial database: contain spatial information in a raster/ vectorformat.
Raster format are either bitmaps / pixel maps. [Image data].
vector format are used to store maps / maps use basic geometric primitives like points, lines, polygons & so forth.

WWW :- It provides a diverse, worldwide online information source
The objective of data mining algorithms is to mine interesting patterns of information present in WWW.

XLM: Extensible markup language.
It is both human & machine interpretable data format, that can be used to represent data that needs to be shared across the platforms

Data Stream :-
It is dynamic data, which flows in & out of the observing environment
Typical characteristics of data stream are
Huge volume of data.
Dynamic
Fixed order movement
Real time constraints

RSS (Really Simple Syndication): For sharing instant feeds across services.

JSON (Java Script Object Notation): Used in ML algorithms.

1.9. Big data Analysis framework.

Many frameworks are proposed for performing data analytics.
All proposed analytics frameworks have some common factors
Big data framework is a layered architecture.

1. Data connection layer
2. Data management layer
3. Data analytics layer
4. Presentation layer.

① Data connection layer:-

It has data ingestion mechanisms & data connectors.
Data ingestion means taking raw data. & importing it into appropriate data structures.
It performs the tasks of ETL process.- Extract, Transform & load operation.

2. Data Management layer:

It performs preprocessing of data. This layer allows parallel execution of queries, read, write & data management tasks.

3 Data Analytic layer:

It has many functionalities such as statistical test, ML algorithms to understand & construction of ML models.
This layer implements many model validation mechanisms too.

4. ~~Preprocessing~~

4. Presentation layer:-

It has mechanisms such as dashboards & applications that display the results of analytical engines & ML algorithms
Thus, the big data processing cycle involves data management that consists of the following steps

1. Data collection
2. Data preprocessing
3. Applications of ML algorithm
4. Interpretation of results & visualization of ML algorithm.