UNIT - I

Introduction to Machine Learning

Evolution of Machine Learning

⇒ Machine learning is first started by Alan Turing and He was the first to propose that machines can learn and became artificially intelligent.

⇒ Arthur Samuel from IBM Laboratory has developed first machine learning program that could play checkers.

⇒ The first neural network program that can simulate human brain is designed by Frank Rosenblatt.

⇒ The first pattern Recognition algorithm i.e., Nearest neighbor algorithm is introduced.

⇒ The first self-driving car that can navigate and avoid obstacles in a room is developed by Stanford University students.

⇒ A variety of NN i.e., RNN - Recurrent Neural N/w is introduced

⇒ One of the learning strategies i.e., Reinforcement learning concepts were introduced.

⇒ Machine learning commercialization is started.

⇒ ML algorithms such as - Random forest and SVM were developed.

⇒ A program named Deep Blue was invented by IBM and it defeated world chess champion, Gary kasparov.

⇒ Netflix launched first ML competition.

⇒ Deep Learning concepts were introduced by Geoffrey Hinton.

⇒ A website for ML competitions, named Kaggle is initialed.

⇒ A new machine learning program named Google's AlphaGo program is introduced. Go is a board game which can be played by 2-players. It is a very complex game than chess, but it also defeated human player.

What is Human Learning?

✓ Learning is referred to as the process of gaining information through observation.

Why do we need to learn?

✓ If we learn more, we can do the tasks more efficiently.

✓ To do a task in a proper way, we need to have prior information on one (or) more things related to the task.

Ex: Information from past rocket launches helps in taking the right precautions and makes more successful rocket launch.

Types of Human Learning

There are 3 - ways in which a human being can learn.

1. Learning under expert guidance

✓ It is the process of gaining information from a person having sufficient knowledge based on past experience.

✓ In all phases of life of a human being, there is an element of guided learning.

Ex:

1. A baby calls his/her hand, as a hand, because that is the information he/she gets from his/her parents.
2. The baby says - the sky is blue - because the baby was taught by the parents.
3. When the baby starts going to school, learns alphabets and digits from teacher.
4.

2. Learning guided by knowledge gained from experts

✓ It is not like direct learning, it is some past information shared on some different situation, which is used as a learning to make decisions.

✓ It is applying knowledge which has been imparted by some expert at some point of time.

Ex:

1. A baby can group objects of same color even if his/her parents have not specifically taught.
2. A student can identify what is a noun and a verb by looking at a sentence, due to the knowledge he/she gained from English teacher long back.

3. Learning by self (or) self-learning

Humans can learn on their own in many situations.

Ex:

* A baby learning to walk through obstacles.

What is Machine Learning?

⇒ As per Tom M. Mitchell, machine learning can be defined as - A computer Program is said to learn from experience 'E' with respect to some class of tasks 'T' and performance measure 'P', if its performance at tasks in 'T', as measured by 'P', improves with experience 'E'.

Ex: In the context of learning to play chess,

E - represents - Experience of playing the game

T - represents - the task of playing chess and

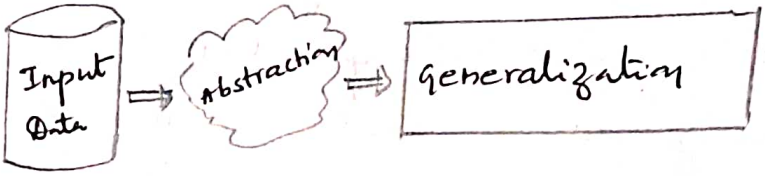
P - represents - the no. of times, that game is won by the player.

Machine Learning Process

ML process can be divided into 3-parts.

1. **Data Input** : Past data is utilized as a basis for future decision making. ↓ knowledge
2. **Abstract** : The input data is represented through an algorithm
3. **Generalization** : Algorithm is generalized to form a framework for making decisions.

The following fig. represents the machine learning process.



⇒ In the ML process, knowledge is fed in the form of input data. But, the data cannot be used in the original shape and form.

ABSTRACTION

⇒ A conceptual map (or) model is derived based on the I/p data.

⇒ This model is derived with the help of abstraction. The model can be in any one of the following forms:

* Conditional blocks like - if / else rules
* Mathematical Equations
* Specific data structures like trees / graphs.
* Logical grouping of similar observations.

⇒ The choice of the model used to solve a specific learning problem is a human task. This choice depends on multiple aspects, such as:

* The type of problem to be solved:
  Ex: Prediction (or) forecast problem,
  Trend analysis, grouping of objects etc,

* Nature of the I/p data :

Ex: Size of data,
whether data has missing values,
different data types, etc,

* Domain of the Problem :

Ex: fraud detection, Disease detection etc,

⇒ Once the model is chosen, the next task is to fit the model based on the I/p data.

Ex: Suppose a model is represented by a mathematical equation, say,

$y = mx + c$ ⇒ This model is known as linear Regression

⇒ Based on the I/p data, we have to find out the values of 'm' and 'c', otherwise the equation (or) the model is of no use.

⇒ Therefore, fitting the model in this case, means finding the values of the unknown coefficients (or) constants of the equation.

⇒ The process of fitting the model based on the I/p data is known as training.

⇒ The Input data based on which the model is being finalized is known as training data.

GENERALIZATION

⇒ Another key part of ML process is generalization.

⇒ In this phase, the model is applied on unknown data, known as test data and to take a decision.

⇒ At this time, 2 problems may raise and are as follows :

(i) The trained model is trained with too much input data, hence, the model may not produce actual result/decision.

(ii) The test data may possess certain characteristics which may be seen as different as trained data.

⇒ Due to these problems, an exact reason-based decision-making is not possible.

WELL-POSED LEARNING PROBLEM

⇒ By using a machine learning frame work, a new problem can be defined.

⇒ This framework offers answers to the following 3 Questions:

- ✓ what is the problem ... ?
- ✓ Why does the problem need to be solved ... ?
- ✓ How to solve the problem ... ?

Step-1 : What is the Problem ... ?

⇒ In this step, the problem has to be described formally and informally. and a list of assumptions and similar problems has to be considered.

* Informal Description of the problem :

Ex: I need a program that will prompt the next word, when I type a word.

* formal Description:

using Tom Mitchell's machine learning definition, we can write formal desciption of the above problem

Ex: Task (T) ; Prompt the next word when I type a word

Experience (E) : A collection of commonly used English words and phrases.

Performance (P) : The no. of correct words prompted.

* Assumptions :

Creat a list of assumptions about the problem.

* Similar Problems :

List the problems that are similar to the problem trying to solve.

STEP-2 : Why does the problem need to be solved...?

⇒ In this step, the motivation behind this problem, the benifits of solving the problem and how that solution can be used have to be listed.

* Motivation

List out the reasons (or) facts to solve the problem.

Ex : ✓ Fraud transactions can be detected as it is a long standing issue.

✓ Suggesting movies to enjoy in weekend.

* Solution benefits

Consider the benefits of solving the problem and they can be articulated to sell the project.

* Solution use

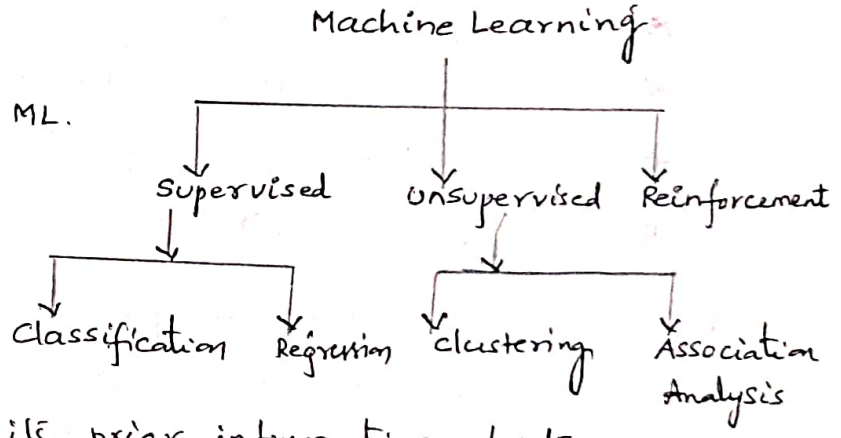
listing out, how the solution can be used.

STEP-3 : How would I solve the Problem...?

Describe how the problem would be solved manually. Detail out step-by-step data collection, data preparation and program design to solve the problem.

Types of Machine Learning

There are 3-categories of ML.



(i) Supervised Learning

✓ It is also called predictive learning. A machine predicts the class of unknown objects based on its prior information about similar objects.

(ii) Unsupervised Learning

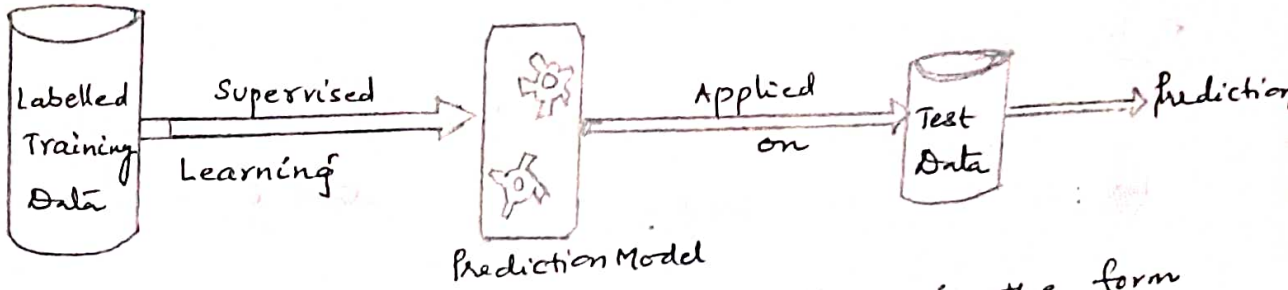
✓ It is also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects together.

(iii) Reinforcement Learning

A machine learns to act on its own to achieve the given goal.

1. Supervised Learning

⇒ The following fig. represents the Supervised learning process.



⇒ In ML, the basic I/p (or) the experience is given in the form of training data.

⇒ This training data is the past information on a given task.

⇒ This training data is called "Labelled Data" in Supervised Learning.

⇒ As shown in the fig, this Labelled data is given as I/p to the machine.

⇒ Then, the machine builds a predective model.

⇒ This model is applied on test data.

⇒ The model finally predicts a label and assigns the label for each record in the test data.

Areas where we can apply supervised learning are :

* Predicting the results of a game.
* Predicting whether a tumor is malignant (or) benign
* Predicting the price of domains like stock, real estate etc,
* Classifying mails as Spam (or) non-Spam.

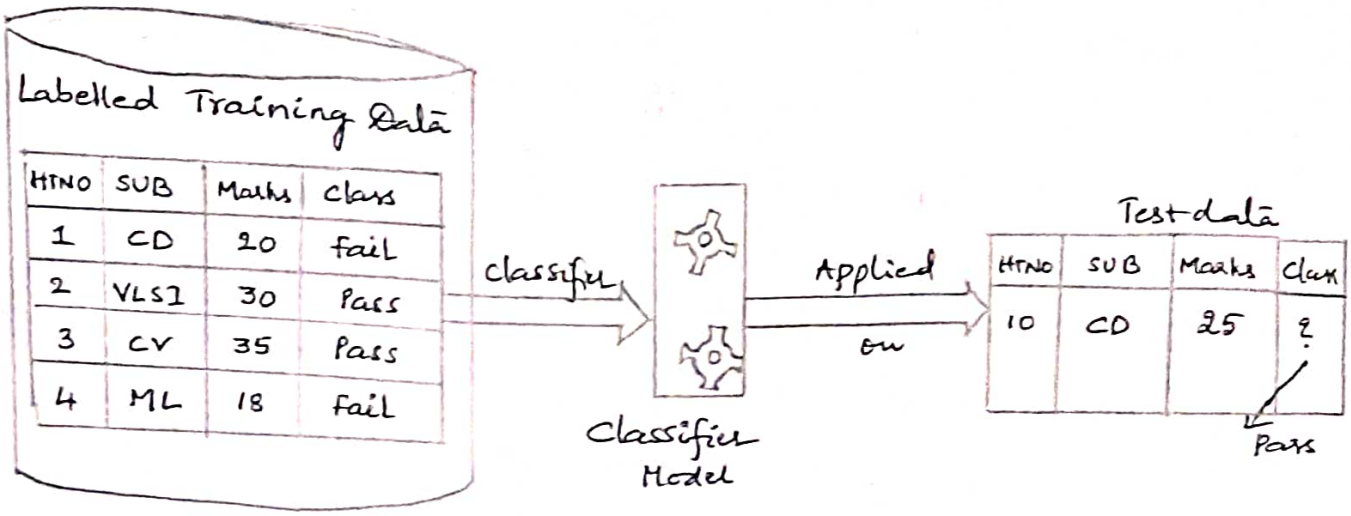
⇒ If we consider the example, predicting whether a tumor is harmful (or) not, we'll get the answer as either "YES" (or) "NO" When we are trying to predict a label and if that label is a "categorical" (or) "nominal variable", then the problem is known as a "classification Problem".

⇒ If we consider the example, predicting price of a real estate, we'll get the answer as a "real value variable". Then, this kind of problem is known as "Regression Problem."

Hence, classification and regression comes under supervised learning.

Classification

The following fig represents the classification process.



⇒ In classification, the whole problem revolves around assigning a label (or) category (or) class to a test data based on the label (or) category (or) class information that is given by the training data.

⇒ Few ML algorithms to solve classification problems are Decision Trees, K-NN and Naive Bayes etc,

⇒ Finally, the target categorical feature in classification is known as "class".

Areas where classification can be applied include :

* Image classification
* Disease Prediction
* Win-loss prediction of games
* Prediction of earthquake, floods etc,
* Recognition of handwriting

2. Regression

⇒ Prediction of a real valued variable comes under regression.

⇒ Regression can be classified as simple linear regression and multiple linear regression.

Linear Regression

⇒ It is one of the most well known algorithms in statistics and ML.

⇒ This algorithm assumes a linear relationship b/w the I/p variable(x) and the single o/p variable.

⇒ when there is a single I/p variable (x), then the method is known as simple linear Regression.

⇒ when there are multiple I/p variables, then the method is referred to as multiple linear regression..

⇒ In this algorithm, the I/p variable is also named as predictor variable and the o/p variable is named as target variable.

simple linear Regression

✓ This method is used to estimate the relationship b/w two quantitative variables.

✓ The value of the dependent variable always changes when there is a change in the independent variable value.

✓ This method uses a straight line.

For Ex: A linear regression model can be represented as:

$Y = mx + c$ ⇒ A straight line equation.

⇒ Here, Y is a target variable (or) o/p variable.

x is a predictor (or) I/p variable

m is a coefficient and

c is a constant.

✓ The variables predictor and target are continuous in value.

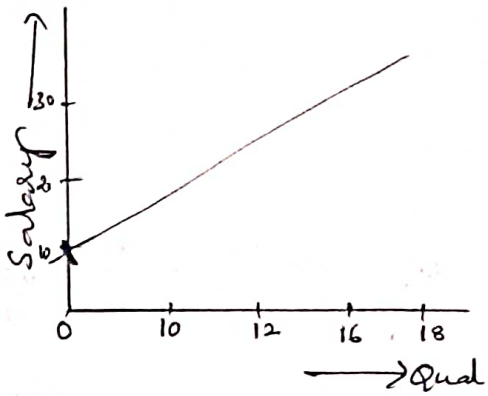
✓ A straight line relationship is "fitted" b/w I/p & o/p variables based on least squares method from statistics. (LSM)

using this LSM, the sum of square of error b/w actual and predicted values of the target variable is to be reduced.

Ex) The following fig: represents the linear regression relationship b/w Qualification and Salary.

✓ If the qualification increases, salary also increases.

✓ In this example, qualification is an I/p variable and salary is an O/p variable.



few of Regression Applications are:

* Demand forecasting in retails
* Sales prediction for managers
* Price prediction in real estates
* Weather forecast etc;

2. Unsupervised Learning

⇒ In this learning strategy, there is no labelled training data and no prediction.

⇒ The model simply takes the data as I/p and try to find groupings (or) patterns within the data elements (or) records.

⇒ Unsupervise learning is also known as "Descriptive Model"

⇒ The process of unsupervised learning is referred to as "pattern discovery" (or) "knowledge Discovery"

1. Clustering

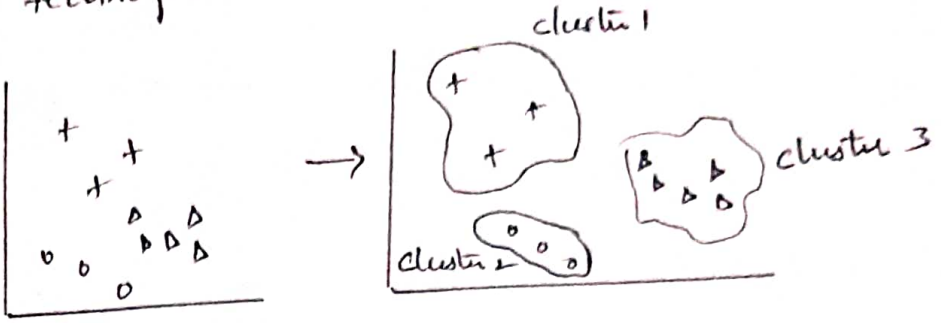
✓ This technique is used to group similar objects together.

✓ Similarity among objects can be measured by in different ways.

✓ one measure is distance. 2 Data items are considered a part of the same cluster if the distance b/w them is less.

✓ If the distance b/w data items is high, the items do not belong to the same cluster.

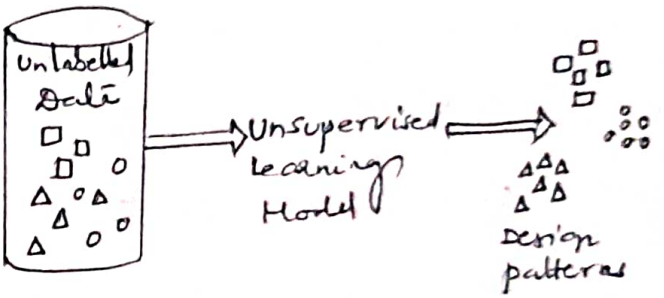
✓ Hence, this technique is also known as "distance-based-clustering".



2. Association Analysis

⇒ It is an another type of unsupervised learning.

⇒ The association (or) relationship b/w data elements is identified by using this technique.



Ex: Market-Basket Analysis

TransID	Items Bought
1	{Butter, Bread}
2	{Diaper, Bread, Milk, Beer}
3	{Bread, butter, Diaper, Beer}
4	{Bread, Diaper, Chicken, Beer}
...	.....

Frequent itemsets ⇒ (Diaper, Beer)

Possible Association ⇒ Diaper → Beer

Applications of Association Analysis include → Market Basket Analysis, Recommender Systems.

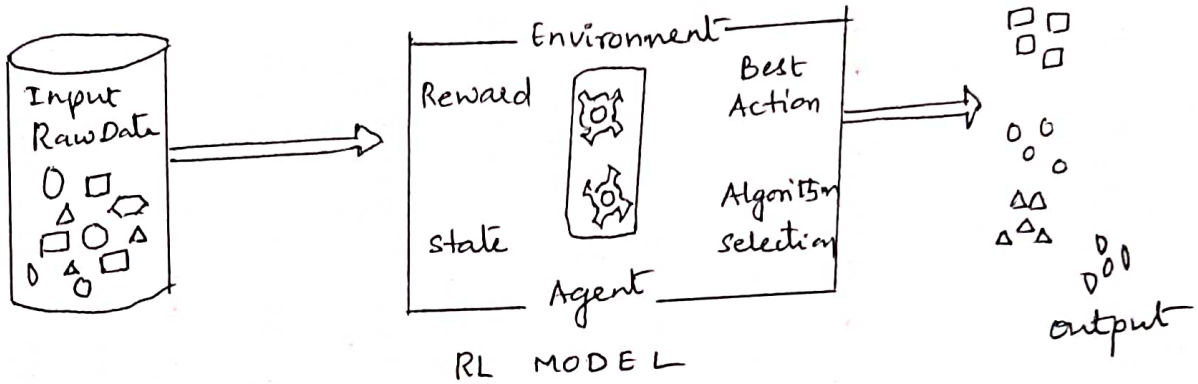
Reinforcement Learning

⇒ The model in this learning strategy has to do "classification", But it's not provided with any idea about the class (or) label of a particular data.

⇒ The model will get a "Reward" if the classification is correct, else it will get a "Punishment".

⇒ The model learns and updates itself through reward/punishment.

⇒ RL is more complex to understand and apply.



RL MODEL

Example of RL is Self-Driving car. It needs speed limit in different road segments, traffic conditions, road conditions, weather conditions etc; The tasks that have to be done by the car are start/stop, accelerate/deaccelerate, turn to left/right etc;

Problems not to be Solved using Machine Learning

✓ ML should not be applied to those tasks in which humans are very effective (or) there is a very frequent human interaction.

Ex: Air traffic control is very complex & it needs human involvement.

✓ Simple tasks which can be implemented using traditional programming paradigms, ML should not be used.

Ex: Simple rule (or) formula based applns like price calculation, dispute tracking appln, do not need ML techniques.

✓ When training data is not sufficient, then, ML is not effectively.

COMPARISION B/W SUPERVISED, UNSUPERVISED & RL

SNO	CRITERIA	Supervised Learning	Unsupervised Learning	Reinforcement Learning
1	Definition	The machine learns by using Labelled Data.	The model has to find patterns in the data.	An agent interacts with the Environment by performing Actions & learning from errors(or) Rewards.
2	Type of problems	Classification & Regression	Clustering & Association Analysis	Reward - Based
3.	Type of Data	Labeled Data	Unlabeled Data	No Predefined Data
4.	Training	External Supervision	No Supervision	No Supervision
5.	Approach	Maps the labeled Inputs to the test data	Understands patterns & Discovers the output	follows the trial-and-Error Method
6.	Implementation	Simplest one to understand.	More difficult to understand and implement	Most Complex to understand & apply.
7.	Algorithms	Naive Bayes, KNN, Decision Tree, Linear Regression, Logistic Regression, SVM etc.,	k-Means, self-organizing Map (SOM), Principal Component Analysis (PCA), DBSCAN etc.,	Q-Learning SARSA
8	Applications	Handwriting Recognition Stock Market Prediction Disease Prediction Fraud Detection, etc.,	Market-Basket Analysis Recommender Systems Customer Segmentation etc.,	Self-Driving Cars Intelligent Robots AlphaGo

CHAPTER-II : PREPARING TO MODEL

Machine Learning Activities

⇒ The first step in ML starts with data.

* In Supervised ML ⇒ there is training & test data
* In Unsupervised ML ⇒ finding patterns in data.

⇒ Therefore, to perform any ML activity, it is necessary to perform multiple pre-processing activities on the input data.

⇒ The following fig: ~~represents~~ represents the four-step process of machine learning

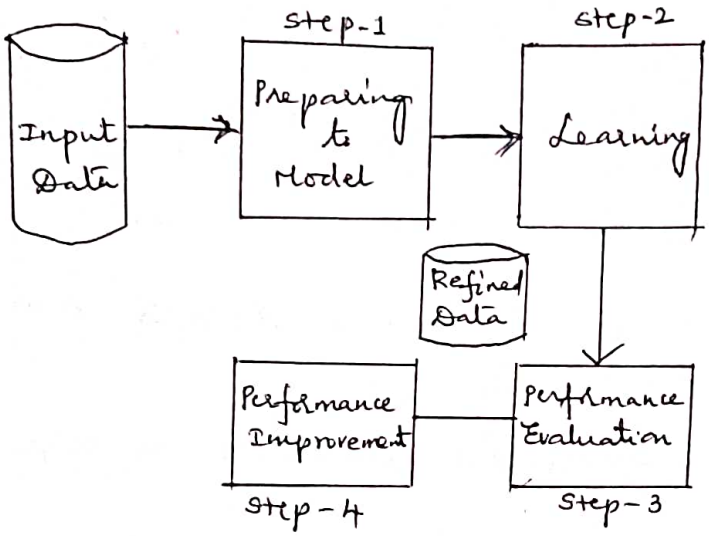


Fig : Detailed process of ML

⇒ The follg. are the pre-processing (or) preparation activities done on I/p data. in step-1

* Understand the type of data
* Nature & Quality of data
* Relationship amongst data elements.
* Find Issues in data
* If any missing values are there, fill them etc.,

⇒ In step-2 ⇒ The follg. activities to be done during learning

* Divide I/p data into training & test data (supervised)
* Consider different models/algorithms for selection
* Train the model based on the training data (supervised) (or) Directly apply the chosen model (unsupervised) on data

⇒ In Step-3 ⇒ Estimate the model performance

⇒ In Step-4 ⇒ Based on options availability, certain actions can be taken to improve the performance of the model.

Types of data in ML

✓ Dataset

⇒ It is a collection of related information (or) records.

⇒ Each record will have multiple attributes.

⇒ Attributes are also named as feature, dimension (or) field.

⇒ A row (or) record represents a point in a data space.

Ex:

Rollnumber	Maths	Science	%
01	90	75	82.03
02	80	65	70.50
03	75	50	55

⇒ In this example, there are 4 features/attributes, Hence, this is called "4-Dimensional Data Space"

⇒ Each dimension/feature provides information on a specific characteristic

⇒ There are different types of data in ML and can be divided into two types:

(i) Qualitative (ii) Quantitative

(i) Qualitative ⇒ Provides information about the quality of an object which can not be measured.

Ex: Good, Average, Poor

⇒ Qualitative Data is also called "categorical data."

⇒ Qualitative Data is also divided into 2 types:

(a) Nominal Data ⇒ It has no numeric value, but, a named value

⇒ It is used for assigning named values to attribut

Ex: Blood group : A, B, O, AB, etc,

Nationality : Indian, British, American etc,

Gender : Male, female, Other

⇒ A special case of nominal data is when only two labels are possible, i.e., for ex: pass / fail.

⇒ This type of nominal data is called "dichotomous"

(b) Ordinal Data ⇒ It has named value and also be ordered.

⇒ Ordinal data also assigns named values to attributes.

⇒ Ordinal data can be arranged in a sequence of increasing (or) decreasing value.

⇒ Hence, we can perform mode, median and Quartiles operations on this type of data, but, mean cannot be performed.

Ex: Grades: A, B, C etc, Customer satisfaction: Happy, Very Happy, Unhappy etc,

(ii) Quantitative Data ⇒ It relates to information about the quantity.

⇒ It can be measurable.

⇒ It is also named as Numeric data.

Ex: Marks — It has a scale of measurement.

⇒ There are 2 types of Quantitative Data.

(a) Interval Data ⇒ It is numeric data, it has order and the difference between values is also known.

Ex: Temparature, date, time etc,

The difference between 15.5°c and 21.5°c is 6°c.

⇒ We can perform addition, subtraction, mean, median, mode, standard deviation.

⇒ Interval data do not have "true zero" value.

Ex: There is nothing called "no temparature".

(b) Ratio Data ⇒ Numeric data for which exact value can be measured.

⇒ Absolute zero is considered as ratio data.

Ex: Height, age, weight, Salary etc,

⇒ Addition, subtraction, multiplication, Division, mean, median, mode, standard deviation operations can be performed on this data.

⇒ Attributes can also be categorized into 2-types based on a number of values that can be assigned to them.

* Discrete * Continuous

* Discrete attributes ⇒ can take finite (or) countably infinite values
  * Nominal attribute such as DoorNo, StreetNo, PhoneNo, Pincode etc., can have a finite no. of values.
  * Numerical attribute such as rank of students can have countably infinite values.
  * Binary attribute — which assume two values only
    Ex: Male/Female, Yes/No, Positive/Negative etc.,
* Continuous attributes ⇒ can assume real numbers.
  Ex: length, height, weight, price etc.,

Exploring structure of Data

⇒ There are different approaches to deal with numeric and categorical data.

⇒ We can have a data dictionary with a standard data set.

⇒ A data dictionary is a meta data repository i.e repository of all information related to the structure of each element contained in the data set.

⇒ Data Dictionary provides detailed information on each of attribute.

⇒ In case, the data dictionary is not available, a standard library function of ML tool can be used to get the details.

Exploring Numerical Data

⇒ There are 2 mathematical plots which will explore numerical data effectively.

* Box Plot * Histogram

(i) Understanding central tendency - Understand nature of numerical data

⇒ If we apply mean and median, the measures of central tendency from statistics on data, we can better understand the nature of numeric variables.

Mean ⇒ Sum of all data values divided by the count of data elements.

Ex: Mean = $\frac{1+2+5+4+6}{5} = 18/5 = 3.6$

Median ⇒ The value of the element appearing in the middle of an ordered list of data elements.

Ex: If we consider the above list in an order like,

1, 2, 4, 5, 6 → then the median value is → 4.

(ii) Understanding Data Spread

To understand the spread (or) deviation of data, we can use the follg methods.

(a) Measuring data dispersion

⇒ Consider, the data values of 2 attributes,

* Attribute 1 values : 44, 46, 48, 45, 47.
* Attribute 2 values : 34, 46, 54, 39, 52

⇒ Both the set of values have a mean & median of 46.

⇒ If we observe, the first set of values of attribute 1 is clustered around the mean/median value.

⇒ But, the 2nd set of values of attribute 2 is dispersed.

⇒ To measure, this dispersion of data, Variance method can be used. The formula is as follows:

$$\text{variance}(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2,$$

where, x ⇒ is the variable/attribute whose variance is to be measured.

n ⇒ the no. of observations/values of variable x.

Standard deviation of a data can be measured by using variance and the formula is,

$$\text{Standard deviation}(x) = \sqrt{\text{variance}(x)}$$

⇒ In the above example, calculate the variance of attribute 1 & 2.

⇒ For attribute 1,

$$\text{Variance} = \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44^* + 46^* + 48^* + 45^* + 47^*}{5}\right)^2$$

$$= \underline{\underline{2}}$$

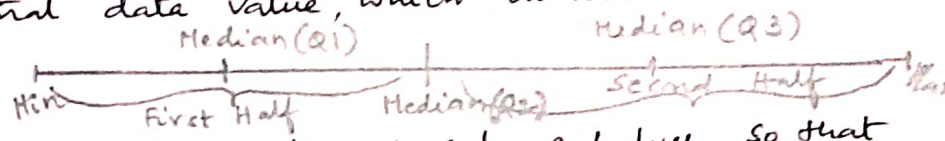
⇒ for attribute 2,

$$\text{Variance} = \underline{\underline{79.6}}$$

⇒ Therefore, we can conclude that - attribute 1 values are clustered around the mean whereas attribute 2 values are extremely dispersed.

(b) Measuring Data value position

⇒ When attribute values are arranged in an increasing order, Median gives the central data value, which divides the entire data set into 2-halves.



⇒ If the first half of the data is divided into 2-halves, so that each half consists of one quarter of the data set, then the median of the first half is known as first Quartile, Q1.

⇒ Similarly, if the 2nd half of the data is divided into 2-halves, then the median of the 2nd half is known as third Quartile, Q3.

⇒ The overall median is known as 2nd Quartile, Q2.

⇒ Therefore, any data set has 5 values -

minimum, Q1, median (Q2), Q3, maximum

Plotting and Exploring Numerical Data : Box Plots

⇒ A box plot is an effective mechanism to understand the nature of the data.

⇒ The box plot also called box and whisker plot gives a standard visualization of the 5 values of data set, namely, minimum, Q1, Q2, Q3, maximum.

⇒ The following is a detailed summary of a box plot.

- ✓ The rectangle / box spans from Q1 to Q3
- ✓ Median is given by the line within the box.
- ✓ The values beyond min. & max. boundaries are called outliers.

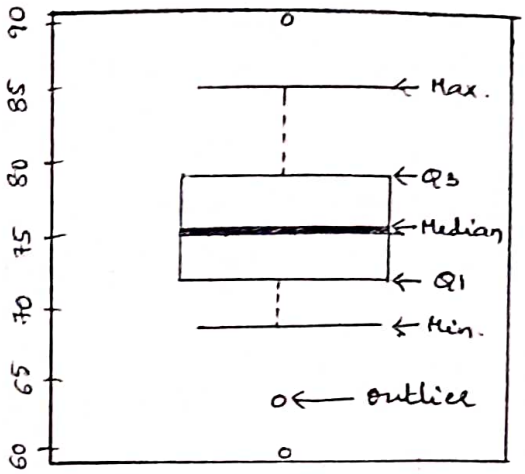


Fig: Box Plot

⇒ The Box Plot is an excellent visualization tool / medium for numeric data

* Histogram

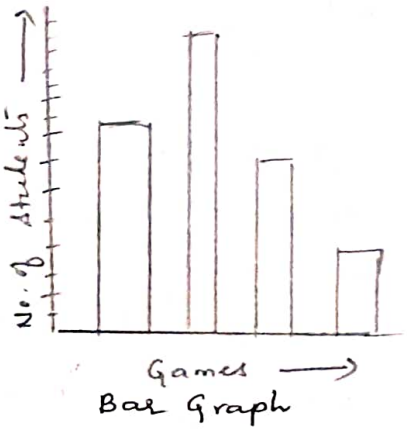
⇒ It is another visualization tool to represent numeric attributes.

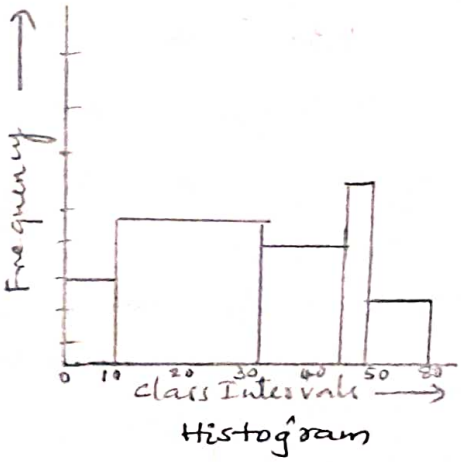
⇒ To understand the distribution of a numeric data into "series of intervals" also called "Bins", this tool works effectively.

⇒ The histogram is represented by a set of rectangles (Bar graph), adjacent to each other, where each bar represent a kind of data.

⇒ Histogram is different from a bar graph.

Histogram	Bargraph
It is a 2-dimnl plot	It is a one dimensional plot
The frequency is shown by the area of each rectangle.	The height shows the frequency.
Rectangles touch each other	Rectangles separated from each other

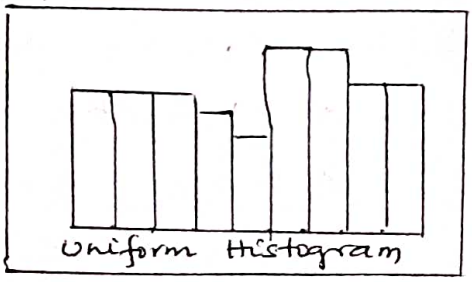




Types of Histogram

⇒ Based on the frequency distribution of data, the types of Histogram are as follows:

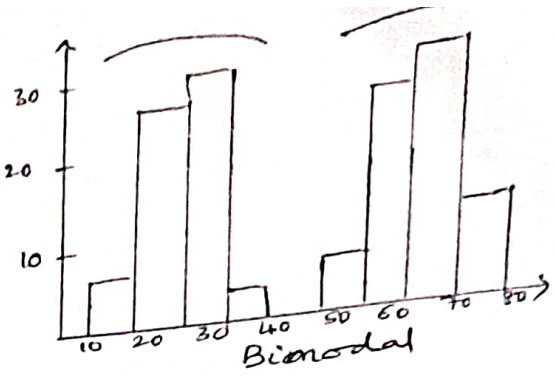
- ✱ Uniform histogram
  - ✓ The no. of classes is too small & each class has the same no. of elements



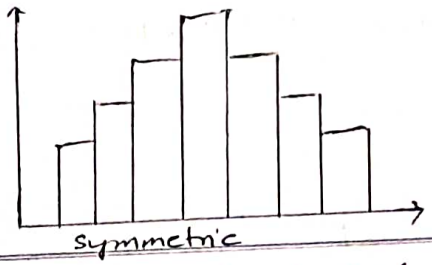
* Bimodal Histogram

⇒ If a histogram has two-peaks, then it is known as bimodal.

⇒ It occurs when the dataset has values on two different kinds of individuals.



* Symmetric Histogram



⇒ It is also called a Bell-shaped plot.

⇒ When we draw the vertical line down the center of the histogram and the 2-sides are identical in size and shape, then it is known as Symmetric.

⇒ The Asymmetric histograms are named as Skewed histograms → Left & Right Skewed.

Exploring Categorical Data

⇒ There are not many options to deal with categorical data.

⇒ one of the statistical measure "mode" is applicable on categorical data.

* Mode ⇒ Most frequently occuring value in the given data set.

Ex: Data = { "Bat", "car", "Bat", "Bike", "Bat", "Ball", "car", "Bat" }

Mode = Bat

⇒ An attribute may have one (or) more than one mode values.

⇒ If the frequency distribution of an attribute have single mode, then, it is called "unimodal"

⇒ If the attribute has 2-modes, then it is called "Bimodal"

⇒ Multiple modes are called "Multi-modal".

Exploring relationship between Variables

⇒ We know that, there exists a relationship between attributes.

⇒ But, this relationship can be visualized effectively using

* Scatterplot
* Two-way cross-tabulations

Scatterplot

⇒ It helps in visualizing "bivariate relationships", i.e., relationship b/w "2-variables"

⇒ It is a two-dimensional plot in which points (or) dots are drawn on coordinates provided by values of the attributes.

⇒ In a dataset, there are 2-attributes, attr-1 and attr-2.

⇒ we want to understand the relationship between attr-1 & attr-2.

⇒ If we change the value of attr-1, how does the value of attr-2 changes....?

⇒ we can draw a scatterplot with attr-1 on x-axis and attr-2 on y-axis.

⇒ Every point in the plot represent the values of attr-1 and attr-2.

⇒ In a 2-dimensional plot, attr-1 is called independent variable(x) and attr-2 is called dependent variable(y)

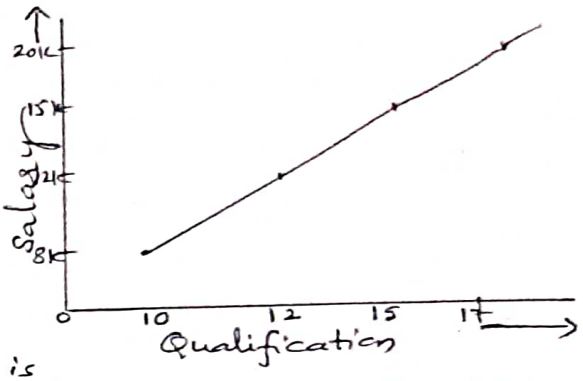


fig: Scatterplot

Ex: As shown in the plot, when the value of Qualification is increasing, the value of attribute Salary is also increasing

2. Two-way cross tabulations

⇒ It is also called "cross-tab" (or) "contingency table."

⇒ It is used to understand the relationship between two categorical attributes.

⇒ It has a matrix format that represents a summarized view of the bivariate frequency distribution of data.

⇒ It is much like a scatter plot, which helps to understand how the values of one attribute changes with the change in data values of another attribute.

Ex: If a teach asks all the students about the toughness and ease of subject named compute vision. This comparision can be shown in the given contingency Table.

Contingency Table

	YES	NO
GIRLS	50	10
BOYS	60	-

DATA QUALITY AND REMEDIATION

1. Data Quality

⇒ ML success depends on the quality of data.

⇒ There are always 2-types of problems with the data:

- ✓ Missing values
- ✓ Outliers

⇒ The reasons for above problems include:

* Incorrect sample set selection:

This data may not reflect normal (or) regular quality due to incorrect selection of sample set.

Ex: If we are selecting a sample set of sales transactions from a festival period and trying to use that data to predict sales in future. This leads to wrong prediction.

* Errors in data collection:

It results in outliers and missing values.

Outliers ⇒ Data elements which are different from other elements

Missing values ⇒ A value that is missing for an attribute in a record.

2. Data Remediation

The above said data problems can be solved, if the right amount of efficiency has to be achieved in the learning activity

* Handling outliers

⇒ Outliers are data elements with an abnormally high/low value which may affect prediction accuracy. The following measures can be taken to deal with outliers

(i) Remove outliers ⇒ If the count of outlier records are few, then they can be simply removed.

(ii) Imputation ⇒ The value of the most similar data element may be imputed (or) impute the value with mean (or) median (or) mode.

~~(iii) Capping~~

Data Preprocessing ⇒ It is a process of preparing the raw data & making it suitable for an ML model.

(i) Dimensionality Reduction

⇒ High dimensional data sets need a high amount of computational space and time.

⇒ But, not all features/dimensions/fields are useful, because they degrade the performance of ML algorithms.

⇒ Most of the ML algorithms perform well, if the number of features in a dataset is reduced.

⇒ Irrelevant and redundant features can be reduced by using dimensionality reduction method.

⇒ It is easy to understand a model, if the no. of features involved in the learning activity is less.

⇒ The most commonly used technique for dimensionality reduction is "Principal Component Analysis" (PCA)

*PCA ⇒ It is a statistical technique.

⇒ It is used to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components.

⇒ These principal components are a linear combination of the original variables.

⇒ They are orthogonal to each other.

⇒ As principal components are uncorrelated, they capture the maximum amount of variability in the data.

⇒ Another commonly used dimensionality reduction technique is "Singular value Decomposition" (SVD).

(ii) Feature subset selection

- ✓ It is also called feature selection
- ✓ It is also used in both supervised as well as unsupervised ML.
- ✓ It tries to find out the optimal subset of the entire feature set.
- ✓ It reduces the computational cost without any impact on the learning accuracy.
- ✓ A feature subset may lead to loss of useful information, this is because, certain features may be excluded from final set of features used for learning.
- ✓ Irrelevant and redundant features may be selected for elimination.

Applications of Machine Learning