UNIT-1

Introduction to Machine Learning

→ It is the science of getting computers to learn, without being explicitly programmed

i)

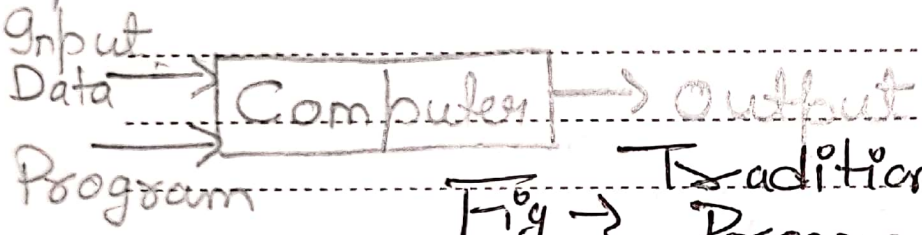


Fig → Traditional Programming

ii)

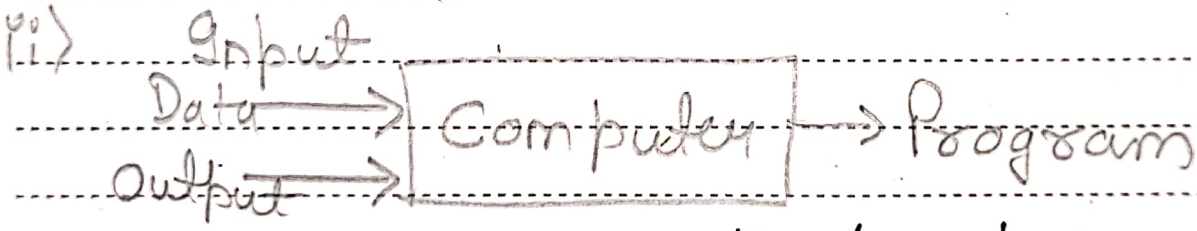


Fig → Machine Learning

→ Machine learning is an application of artificial intelligence that provides the systems the ability to learn automatically & improve from experience without being explicitly programmed.

→ Machine learning works on a simple concept i.e. understanding with experience

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

i) Ex Facebook continously notices
→ friends that you connect with
→ profiles that you visit
→ your interest etc.

On the basis of continous learning, a list of facebook users are suggested that you can become friends with.

ii) Tag Friends
When you upload a picture of you with a friend, facebook instantly recognizes that friend. This is possible with the help of machine learning.

iii) Ads Recommendation
When you shop any product online, after some days you keep receiving notifications for shopping suggestions.
The shopping website or the app recommend you some items that somehow matches with your interest

→ This is possible with the help of machine learning, on the basis of your behaviour with the website/app, past purchases, items liked or added to cart etc, the product recommendations are made.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

Primary aim of Machine Learning

→ It is allow the computers learn automatically without human intervention.

Scope of Machine Learning

1) Machine Learning in Education

Teachers can use machine learning to check how much of lectures students are able to consume, how they are understanding the lectures taught & whether they are finding it too much to consume.

This allows the teachers to help their students group the lectures. It also prevent the weak students from fail.

2) Machine learning in Search Engine

Search engine ranks page based on what you are most likely to click on.

Search engine depends on machine learning to improve their services such as voice recognition, image search & many more.

In future, many more interesting features will also come

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

3) Machine learning in Digital Marketing

ML can help significantly in Digital Marketing. It allows a more relevant personalization. Thus, companies can interact & ~~enage~~ engage with the customer.

Companies have information which ~~also~~ can be used to learn Customer behaviour & send messages according to their ~~interact~~ interest to increase company sales.

4) Machine learning in health care

The value of machine learning in healthcare is its ability to process huge datasets beyond the scope of human capability & then relatively convert analysis of that data into clinical insights that aid physicians in planning & providing care, ultimately leading to better outcomes, lower costs of care & increased patient satisfaction

5) Spam Detector:-

Our mail agent like Gmail does a lot of hardwork for us in classifying the mails & moving the spam mails to spam folder

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

6) Traffic Alert:-

Google Maps app is used to give assistance in directions & traffic also based on machine Learning

Everyone using maps app is providing their location average speed, the route in which they are traveling which in turn helps Google collect massive data about the traffic, which makes them predict the upcoming traffic & adjust your route according to it

7) Social Media

Social Media like facebook uses face detection & image recognition to automatically find the face of the person which matches its database & hence suggest us to try that person, this is because of ML.

8) Google Translate

Google translator helps us in translation of any sentence or words written in one language to another language

Google Neural Machine

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

translation (GNMT) is a neural ML that ~~works~~ works on thousand of languages & dictionaries, uses neural language processing to provide the most accurate translation of any sentence or words.

Limitations of ML

1) Accuracy depends on training & learning which is not always available

2) Required accurate details on many past projects.

3) Have large data requirement to learn about various topics which may be time taking & require various resources.

4) Performance cannot be audited or guranteed

5) Require large amount of hand-crafted, structured ~~train~~ training data.

6) A machine cannot learn if there is no data

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

7) A machine needs to have heterogeneity (variations) in dataset to learn meaningful insight

8) Lack of data or the variation in dataset leads to poor evaluation & prediction

Types of Machine Learning

- Supervised
- UnSupervised
- Reinforcement

1) Supervised Learning :- @ In this learning, given training examples of inputs & corresponding outputs, making predict outputs for new inputs.

→ We train the machine using data which is well labeled with correct output

→ As the dataset is labeled, meaning that the algorithm identifies the features explicitly

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

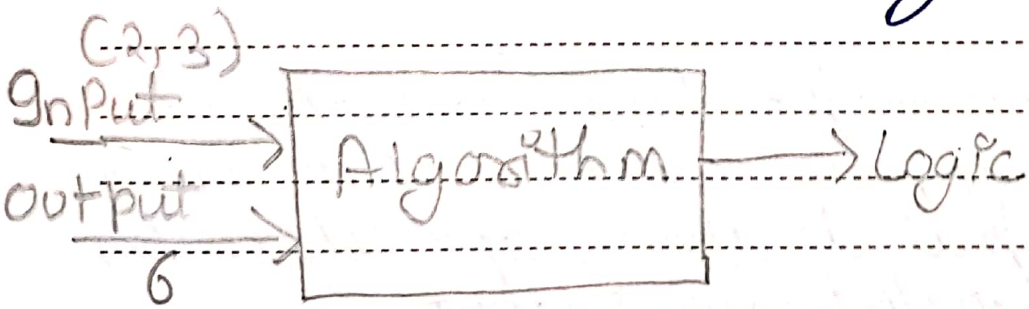
Scanned with OKEN Scanner

& carries out prediction or classification accordingly

→ As the training period progresses, the algorithm is able to identify the relationships between the two variables such that it can predict a new outcome.

→ It is widely used for data where there is a precise mapping between input & output data.

→ You can assume a dataset as a teacher or supervisor & its role is to train the machine / model. Once the machine / ~~model~~ gets trained it can start making a prediction or decision when ~~new~~ new data is given to it.



Training with datasets

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

$6, 2 \rightarrow \boxed{\text{Logic}} \rightarrow 12$

Prediction with new data

→ The term supervised learning comes from the idea an algorithm is learning from a training dataset, which can be thought of as the teacher.

→ SML systems provide the learning algorithm with known quantities to support future judgements.

→ SL occurs when a system is given input & output variables with the intentions of learning how they are mapped together or related.

→ The goal is to produce an accurate enough mapping function that when ~~an~~ new input is given, the algorithm can ~~pre~~ predict the output.

→ This is an iterative process & each time the algorithm makes a prediction, it is

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

Corrected or given feedback until it achieves an acceptable level of performance.

→ Supervised Learning algorithms try to model relationships & dependencies between the target prediction output & the input features such that machine can predict the output values for new data based on those relationships which it learned from the previous datasets.

* Types of Supervised Learning

- Classification
- Regression

1) Classification:-

→ This ~~a~~ problem is when the output variable is a category such as car or truck, red or blue.

→ This technique group the output into a class. If the algorithm tries to label input into two distinct classes, it is

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL DEVELOPERS

Scanned with OKEN Scanner

called binary classification

→ Selecting between more than two classes is known as multiclass classification

→ Ex ⇒ i) Determining ~~whatever~~ whether or not someone will be defaulter of the Loan.

ii) This mail is spam or not

→ Classification separates the data

Examples

Fraud detection, Email spam detection, Diagnostics, Image classification

2) Regression

→ A regression problem is when the output variable is a real value such as price or weight

→ This technique predicts a single output value using training data

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

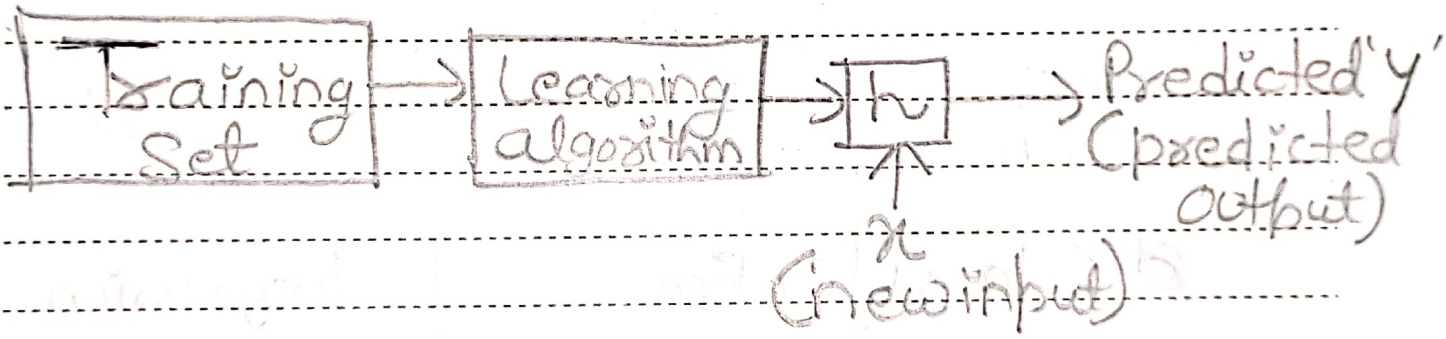
Scanned with OKEN Scanner

→ Ex:- You can use regression to predict the house price from training data. The input variables will be locality, size of house etc.

→ Ex:- Risk Assessment, Score prediction

→ Regression fits the data.

A Training set in Supervised learning



$x^{(i)}$ – input variable / input features
$y^{(i)}$ – output / target variable
$x^{(i)}, y^{(i)}$ – training example

i) Training set is the list of 'm' training examples $\{(x^{(i)}, y^{(i)}); i = 1 \ldots m\}$

ii) In supervised learning, our goal is given a training set, to learn a function

$$h : x \longrightarrow y$$

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

iii) So that $h(x)$ is a good predictor for the corresponding ~~of~~ value of y

iv) When target variable (y) that we are trying to predict is continuous then learning problem is a regression problem

v) When 'y' can take only a small ~~amount~~ number of discrete values, it is a classification problem.

* Examples of Supervised Learning algorithm

Classification	Regression
→ Support Vector Machine (SVM)	Linear Regression
→ Naive Bayes	Decision Trees
→ Nearest Neighbour	Neural Networks
→ Neural Networks	SVM.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS SIRT-S | SIRT-E | SIRT-P | SIRTS-P | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

2) Unsupervised Learning

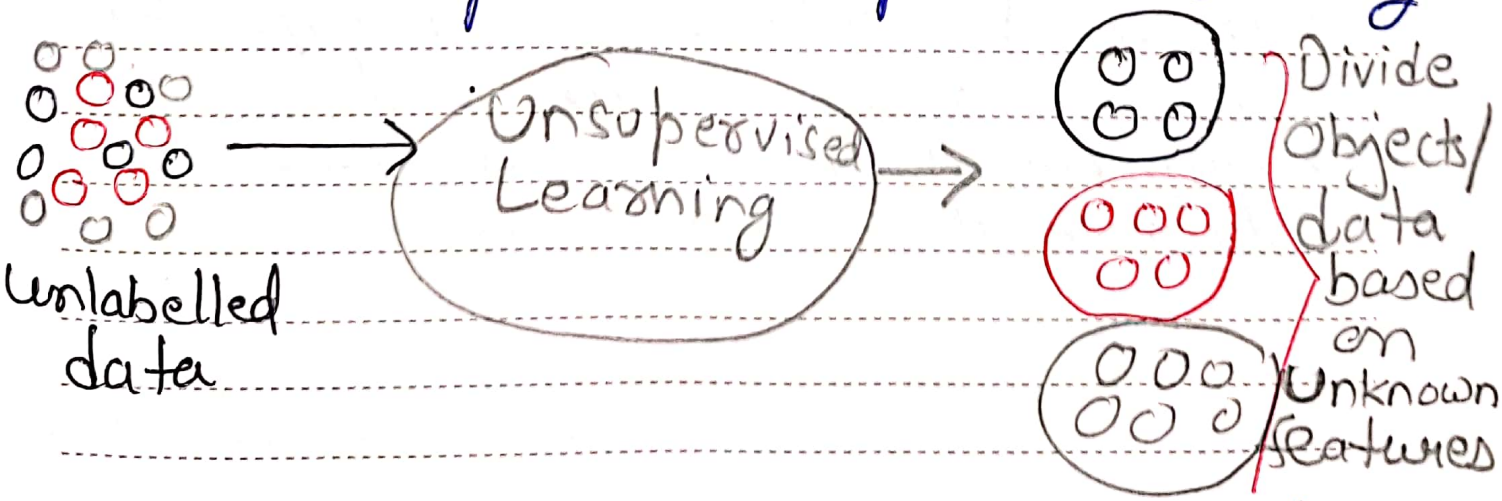
→ It deals with the unlabelled data

→ No teacher (training dataset) is provided that means no training will be given to the machine. Therefore machine work on its own to discover information

→ Here, the task of machine is to group unsorted information according to similarities, patterns & differences without any prior training of data

→ Machine is trained with unlabelled data.

→ Unsupervised Learning algorithms allows you to perform more complex processing tasks compared to Supervised Learning



M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

→ Unlabeled data might include photos, videos, audio recording etc. There is no explanations for each ~~place~~ piece of unlabelled data. It just contains the data & nothing else.

→ UML finds all kind of unknown patterns in data

→ Unsupervised method help you to find features which can be useful for categorization

→ It is easier to get unlabeled data from a computer than labeled data which needs manual intervention

→ Unsupervised learning algorithm try to use techniques on the input data to ~~mi~~ mine for rules, detect pattern & summarize & group the data points which help in deriving meaningful insight & describe the data better to the users.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

* TYPES OF UNSUPERVISED LEARNING

- CLUSTERING
- ASSOCIATION

1) CLUSTERING

→ A clustering problem is where you want to discuss the inherent grouping / categories in the data

→ Ex:- Grouping customers by purchasing behaviour.

→ It mainly deals with finding a structure or pattern in a collection of uncategorized data

→ Clustering algorithms will process your data & find cluster (groups) if they ~~ex~~ exist in the data.

→ It automatically split the dataset into groups based on their similarities

→ Clustering means grouping of Objects based on the information found in the data

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

describing the objects or their relationships.

→ The goal of Clustering is that object in one group should be similar to each other but different from objects in another group.

→ Examples of Clustering algorithms

i) k-Means Clustering
ii) Fuzzy / C-mean Clustering
iii) Hierarchical Clustering

→ Clustering used in real world problems such as anomaly detection recommending systems, document grouping or finding customers with common interests based on their purchases.

2) Association

→ An association rule learning problem is where you want to discover rules that describe large portions of your data

→ Ex:- People that buy X (milk) also tend

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

to buy Y (bread).

→ Association rules allow us to establish associations amongst data objects inside large databases.

→ Association algorithm will process your data & discover interesting relationships between variables in large databases.

→ Association mining identifies sets of items which often occur together in your dataset.

→ Use of association mining in market basket analysis & Recommender Systems.

→ Example of Association algorithms

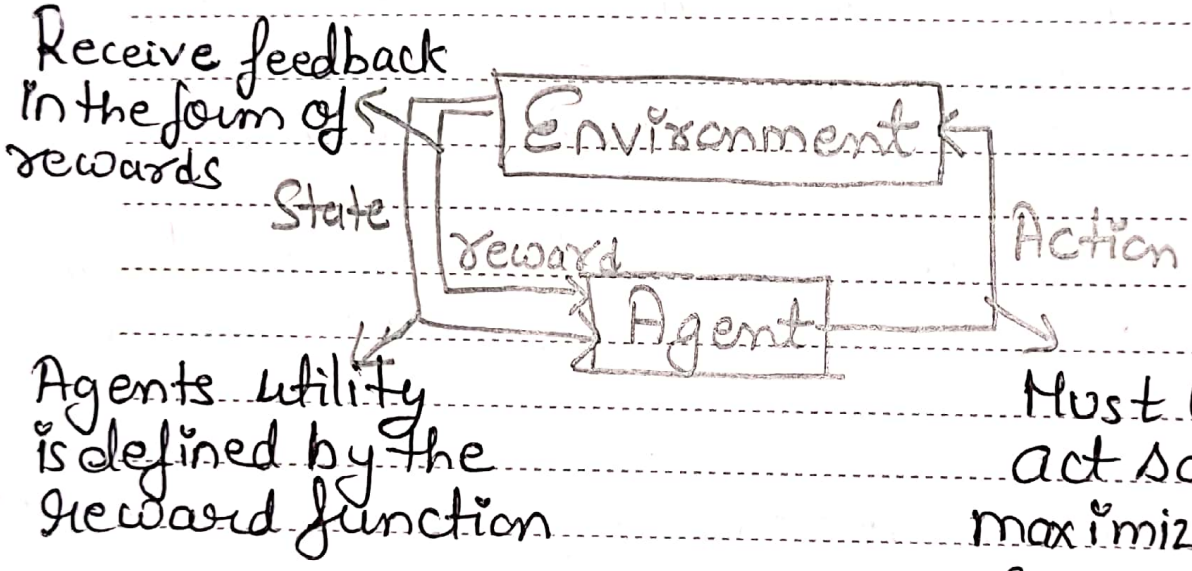
i) Apriori Algorithms
ii) FP-Growth

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

3) Reinforcement Learning

→ It is a type of machine learning in which a computer learns to perform a task through repeated trials & error interactions with a dynamic environment



→ It is different from supervised & unsupervised machine learning algorithms that require large datasets for training purposes.

→ Reinforcement algorithm are designed to continually learn from experience rather than only from data

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

→ In reinforcement learning, an agent will have to interact with environment & find out what is the best outcome

→ Agent follows the concept of hit & trial method, agent is awarded or penalized with a point for a correct or a wrong answer

→ On the basis of the positive reward points gained the model train itself.

→ Once the agent gets trained, it is ready to predict the new data presented to it

→ This learning approach enables the computer (agent) to make a series of decisions that maximize a reward metric for the task without human intervention & without being explicitly programmed to achieve the task.

→ Reinforcement learning algorithm continously learns from the environment in an iterative fashion until it explores the full range of possible states.

→ Simple reward feedback is required for

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

the agent to learn its behaviour, this is known as the reinforcement signal.

* Basic Elements of an Reinforcement Learning

1) Environment - Physical world in which the agent operates.

2) State - Current situation of the agent

3) Reward - Feedback from the environment

4) Policy - Method to map agents state to actions

5) Value - Future reward that an agent would receive by taking an action in a particular state.

Example:-

Consider the task of parking a vechile using an automated driving system

Goal :- The goal of this task is for the vechile computer (Agent) to park the vechile in

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

the correct parking spot with the right ~~observation~~ orientation

Environment - Everything outside the agent like dynamics of the vechile, other vechiles that may be nearby, weather conditions etc.

→ During training, the agent uses readings from sensors such as camera's GPS & lidar (observation) to generate steering, breaking & acceleration command (actions).

→ To learn how to generate the correct actions from the observations (policy tuning), the agent repeatedly tries to park the vechile using a trial & error process.

→ In autonomous parking training is supervised by a training algorithm

→ The training algorithm is responsible for tuning the agents policy based on the collected sensor readings, actions & rewards.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ After training is complete, the vehiles computer should be able to park using only the ~~trained~~ tuned policy & sensor readings

Applications of Reinforcement Learning

1) Robotics for industrial automation
2) Business strategy planning
3) Aircraft control & robot motion control
4) Gaming

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Difference Between

Supervised Learning	Unsupervised Learning	Reinforcement Learning
1) In this, we teach the machine using labelled data	In this, the machine is trained on unlabelled data	In this, an agent interacts with its environment by producing errors or rewards. There is no predefined data
2) There is external supervision	There is no external supervision	There is no external supervision
3) Aim of this learning is to predict output	Aim of this learning is to discover unknown patterns & discover output	Aim of this learning is to learn series of actions.
4) It needs to learn first follow it by a test	It is always looking for a pattern in the data	It tries to figure out data on its own by

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

Supervised learning	Unsupervised learning	Reinforcement learning
		interacting with the environment
5) Uses training dataset	Uses just input dataset	Works on interaction with the environment
6) Data is classified based on training dataset	Uses properties of given data to classify it	learner interacts with the environment to extract the output or make decisions
7) Used for prediction	Used for Analysis	It works as a reward and action system
8) It is relatively complex as it requires labelled data	It is less complex as there is no need to understand & label data	It is more complex as learning agent itself creates data on its own by interacting with the environment

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

Supervised Learning	Unsupervised Learning	Reinforcement Learning
9) It maps labelled input to known output	It understand pattern & discover unknown output	It follows trial & error method
10) It is used when "I know how to classify this data, I just need you (algorithm) to sort it	10) It is used when "I have no idea how to classify this data, can you (algorithm) create a classifies for me.	It is used when "I have no idea how to classify two data & I will give you a reward if to convert or I will punish you if its not correct
11) Algorithm learn from the association between input & output	Algorithm derive the structure from input just looking at relation between input themselves	Learning through delayed feedback by interacting with environment

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS SIRT-S | SIRT-E | SIRT-P | SIRTS-P | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Supervised Learning	Unsupervised Learning	Reinforcement Learning
12) Ex of this learning are → Linear Regression → Logistic Regression → Super Vector Machine → k-Nearest Neighbour	Ex of this learning are- → k-Means → A'prim → C-Means	Ex of this learning are- → Q-Learning → SARSA (State Action Reward ~~Action~~ State Action) Algorithm
13) Applications → Risk Evaluation → Forecast Sales → Fraud detection → Email Spam detection	Applications → Recommendation Systems → Document grouping	Applications → Self driving Cars → Gaming
14) It has two main tasks called Regression & Classification	It has two main tasks called Clustering & Association	It has different tasks such as exploitation & exploration, Markov decision process, Policy learning.

M/s Agrawal Construction Co. | SIRT | SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

Linear Regression (Linear Model)

→ It is a machine learning algorithm based on Supervised Learning

→ It is the easiest & most popular Machine Learning algorithm

→ It is used for predictive analysis

→ It makes predictions for continous (real) variable such as price of product / house, salary etc

→ Regression models a target prediction value based on independent variables

→ It shows a linear relationship between a dependent variable & one or more independent variables, for this reason it is called Linear regression

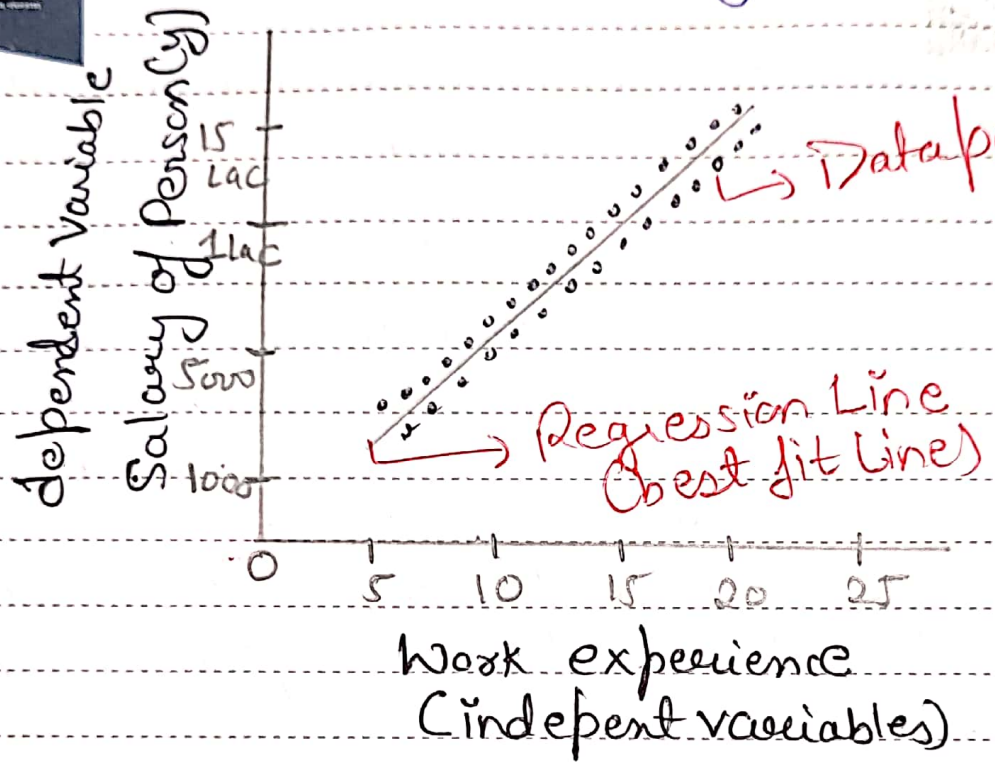
→ It finds how the value of dependent variable (y) is changing according to the value of independent variable (x)

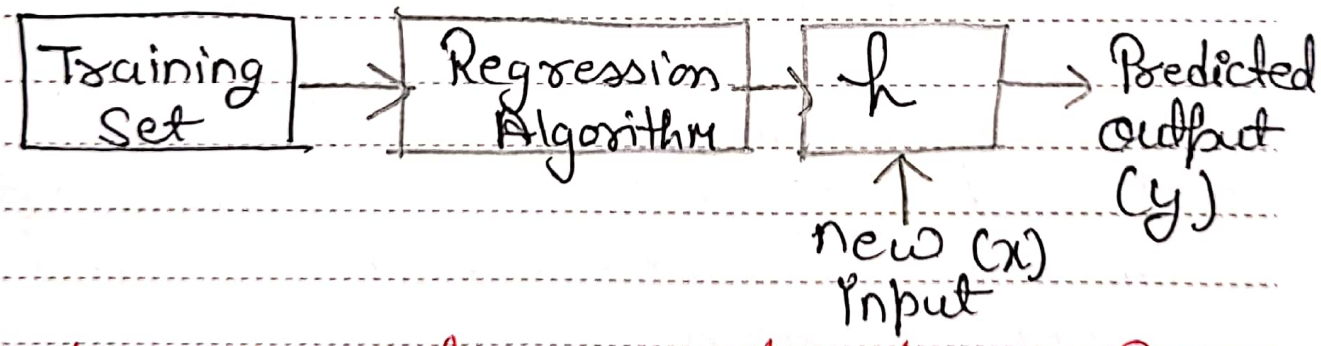
→ This technique finds out a linear relationship

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

between x (input) & y (output)





Hypothesis function for Linear Regression

$$y = b_0 + b_1 x + \epsilon$$

where,

x = input training data (independent variable)

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

y - label ~~of~~ to data (Dependent Variable)
b_0 - intercept of the line
b_1 - Coefficient of x / Linear Regression Coefficient
ε → Random error

→ While training the model we are given input training data (x) & labels to data (y) as it is Supervised Learning

* Regression line - A linear line showing the relationship between the dependent & independent variables is called regression line

* Best fit line - It is a straight line that best represents the data on a scatter plot

→ When working with linear regression our main goal is to find the best fit line that means the error between predicted value & actual value should be minimized

→ The best fit line will have the least error.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ The different values of b_0, b_1 gives a different line of regression

→ Once we find the best b_0 & b_1 values, we get the best fit line, then our model will predict the value of y for the input value of x.

→ To find the best fit line, Cost function is used

* Cost function :-

→ It is used to find the accuracy of the mapping function, which maps the input variable to output variable. This mapping function is also known as hypothesis function.

→ By achieving the best fit regression line, the model aim to predict 'Y' value such that the error difference between predicted value & true value is minimum

→ So, values of b_0 & b_1 are updated to reach the best value that minimize the error between predicted y value & actual y value

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS SIRT-S | SIRT-E | SIRT-P | SIRTS-P | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ For linear regression, we use the root mean squared error cost function which is the average of squared error occured between predicted y value & actual y values

Cost function

$$J = \frac{1}{n} \sum_{i=1}^{n} (y_{i\,actual} - y_{i\,Pred})^2$$

Where

n → the total number of observation

→ The goal of this model is to minimize cost function

Types of Regression

- Simple Line Regression
- Multiple Line Regression

i) Simple Line Regression → Single independent variable is used to predict the value of dependent variable.

ii) Multiple Line Regression :- → Multiple Independent

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

variables are used to predict the value of dependent variable

* For simple Line Regression

$$y = b_0 + b_1 x$$

$y \rightarrow$ dependent variable
$x \rightarrow$ Independent variable
$b_0, b_1 \rightarrow$ Coefficient

* For Multiple Line Regression

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots a_n x_n$$

$a_i \rightarrow$ regression coefficient
$x_i \rightarrow$ independent variable
$y_i \rightarrow$ Dependent variable

* Numerical on Least Square Regression Line

The Least Square regression line for the set of 'n' data points is given by the equation of line in slope intercept form

$$y = ax + b$$

where,

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right)$$

Q-1 Consider the following set of points

$$\{(-2,-1), (1,1), (3,2)\}$$

A) Find the least square regression line for the given data points

B) Plot the given points & the regression line in the same rectangular system of axes

Soln

x	y	xy	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
$\Sigma x = 2$	$\Sigma y = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

$\boxed{n = 3}$

As, we know the least square regression

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

line for n data points is given by

$$y = ax + b$$

Where

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

i) $a = \frac{3\times 9 - 2\times 2}{3\times 14 - (2)^2} = \frac{23}{38}$

$$b = \frac{1}{n}(\Sigma y - a\Sigma x)$$

ii) $b = \frac{1}{3}\left(2 - \frac{23}{38}\times 2\right)$

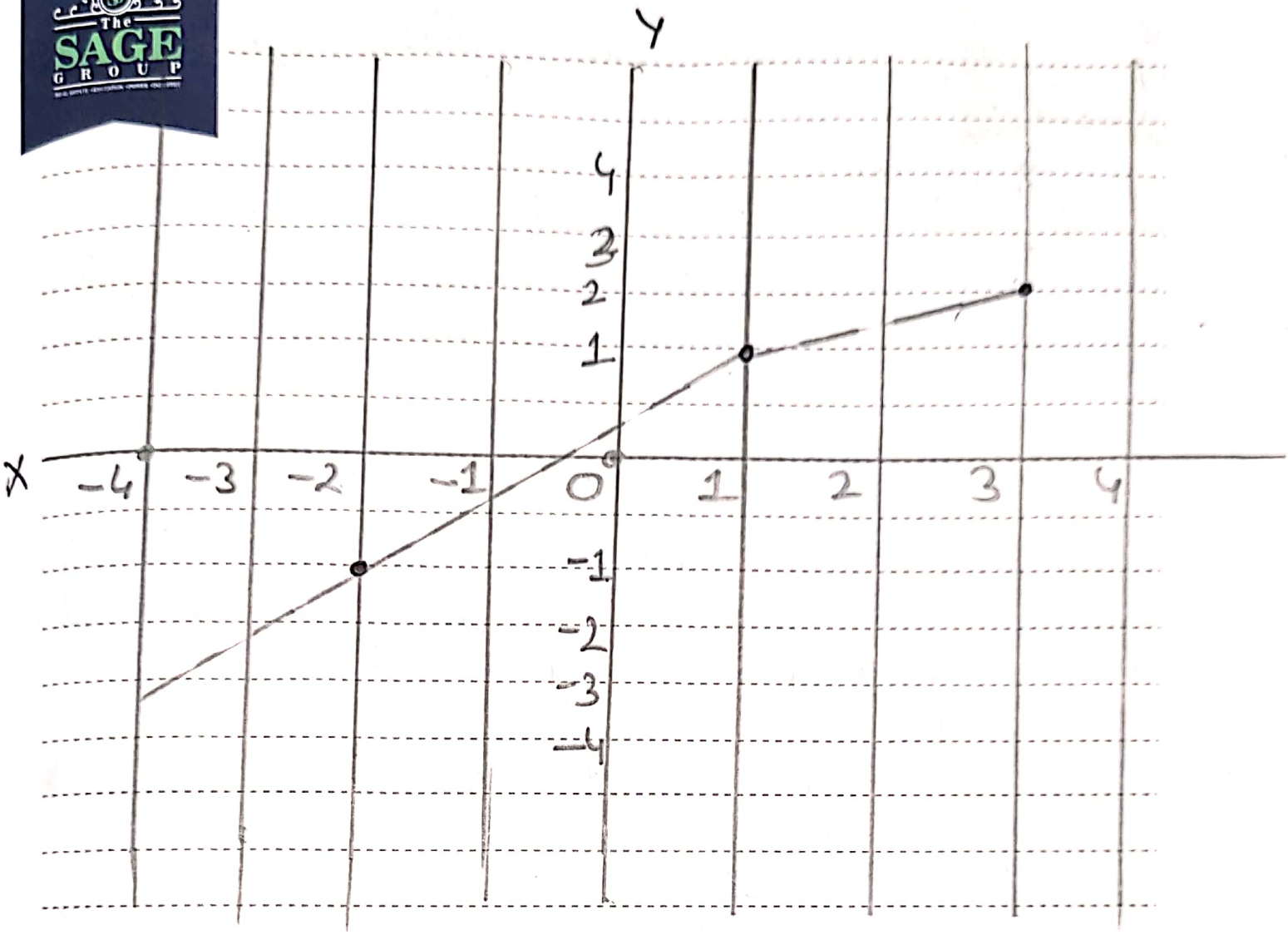
$$b = \frac{5}{19}$$

A) $\therefore \; y = \frac{23}{38}x + \frac{5}{19}$ Ans

B) Graph of given points & Regression Line

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner



Q-2 The sales of a company (in million dollars) for each year are shown in the table below

x (years)	y (sales)
2005	12
2006	19
2007	29
2008	37
2009	45

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

i) Find the least square regression line $y = ax + b$

ii) Use the least square regression line as a model to estimate the sales of the company in 2012

Sol^n

i) First assum 2005 as 0, 2006 as 1, 2007 as 2, 2008 as 3, 2009 as 4, this makes values of x smaller easily manageable

So, the values become

x	y	xy	x^2
0	12	0	0
1	19	19	1
2	29	58	4
3	37	111	9
4	45	180	16
$\Sigma x = 10$	$\Sigma y = 142$	$\Sigma xy = 368$	$\Sigma x^2 = 30$

Here $n = 5$

i) As, we know that least square regression line for n data points is given by

$$y = ax + b$$

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Where,

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \frac{5 \times 368 - 10 \times 14}{5 \times 30 - (10)^2}$$

$$a = 8.4$$

$$b = \frac{1}{n}(\Sigma y - a\Sigma x)$$

$$= \frac{1}{5}(142 - 8.4 \times 10)$$

$$b = 11.6$$

∴ Least Square regression line is

$$y = 8.4x + 11.6 \quad \text{—(1)}$$

ii) Sales of the Company in 2012

	x
2010	5
2011	6
2012	7

for 2012, $x = 7$

Substituting value of x in eqn —(1)

$$y = 8.4 \times 7 + 11.6 = 70.4$$

∴ The estimated sales in 2012 is 70.4 million dollars

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Data Preprocessing

→ It is a process of preparing the data & making it suitable for a machine learning model

→ As quality of data affects the ability of model to learn, hence it is extremely important that we preprocess data before giving as input to model

→ Whenever the data is gathered from different sources it is collected in raw format which is not suitable for the analysis. So data preprocessing is a technique that is used to convert the raw data into a clean dataset

→ Data in raw format contain noises, missing values, unusable format

→ Data preprocessing clean the data & making it suitable for machine learning model which also increases the accuracy & efficiency of a machine learning model

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

* Steps involved in preparing data for ML are:

Data Collection → Data Preprocessing → Data transformation

Step-1

i) Data Collection:-

This step involves gathering data from various sources such as databases, files & external repositories.

→ Before starting the data collection process, first we have to find out the problem solved by an ML model.

→ Knowing the objective by algorithm will assist in determining the type of data required. Also it helps in getting a clear picture of all the data available, required & missing.

Data Augmentation:-

→ In some ~~class~~ cases, data augmentation might be required to expand the size of the existing dataset without gathering more data.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ If we have more data, the better will be our ML models.

→ But every data collection process is associated with a cost. This cost can be in terms of dollars, human effort, computational resources & time consumed in the process.

→ Therefore, we may need to augment existing data to increase the datasize that we feed to our ML models.

→ For ex,- if a dataset of images were collected they can be augmented by

i) Rotating the original image
ii) Crop the original image differently
iii) Altering the light conditions

So for one image you can generate different sub samples.

Data labeling:-

Data labeling in supervised machine learning models, might be also a part of the data preparation process.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ It can be done manually by Crowd workers or automatically using specialized framework (eg ~~Snot~~ Snorkel).

Step-2

ii) Data Preprocessing

As the data collected in step-1 may be in undesired format, unorganized or extremely large so further steps are needed to enhance the quality of data

Steps for preprocessing data

A) Formatting

It is required to ensure that all variables with the same attribute are consistently written, For ex:- All phone numbers, addresses or sums of money should be written in the same format to avoid values in the same column like 26.05, $80, 420 dollars, 2 cents.

B) Data Cleaning

It is applied to remove messy data & manage missing values.

→ At this step you can remove duplicates,

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Smooth noise data & identify or remove outliers if through exploratory data analysis accompained by domain ~~experience~~ expertise shows that this data is not important for the model.

* Ways to handle missing data cleaning

1) Using attribute mean to fill in the missing value

2) Predict the missing value by using a learning algorithm like decision tree, regression etc.

3) Using global constant or dummy values
Like N.A
Eg → 0 (Absence of value)
−1 (Presence of value)

4) Ignore the tuple

* Techniques used to identify outliers & smooth out noisy data

1) Binning:- a) Sort the attribute values &

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

partition them into bins

b) Then smooth by bin means, bin median or bin boundaries

Ex → 15, 10, 5, 7, 13, 12, 9, 11

Sort → 5, 7, 9, 10, 11, 12, 13, 15

5, 7, 9, 10 11, 12, 13, 15

Binsize = 2

Bin mean

$\frac{5+7+9+10}{4}$: $\frac{11+12+13+15}{4}$ { Replace all values of bin by their bin mean

$= 7.75$ $= 12.75$

7.75, 7.75, 7.75, 7.75 12.75, 12.75, 12.75, 12.75

Bin median

for n = odd values $= \frac{n+1}{2}$

for n = even values $= n/2$ { Replace all values by 2 value

$= \frac{4}{2} = 2$

7, 7, 7, 7 12, 12, 12, 12

Bin Boundary

5, 7, 9, 10 11, 12, 13, 15

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

min = 5 | min = 11
max = 10 | max = 15

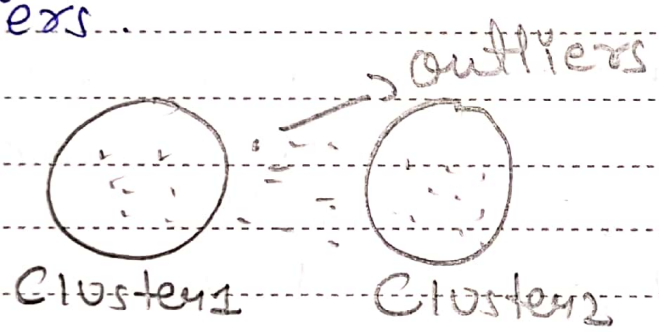
replace value by closet bin boundary

5, 5, 10, 10 | 11, 11, 11, 15

ii) Clustering

Clusters (groups) are formed from the data having similar values

So to remove noise & outliers, we group values in clusters & then dataset & remove outliers.



iii) Regression:-

Smooth by fitting the data into regression functions.

$y = ax + b$

x → independent variable

y → Predicted value [dependent variable]

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

c) Sampling:-

If we have to much data, then we can use sampling

During exploring & prototyping, a smaller representative sample can be fed into the model to save time & cost.

iii) Data transformation

→ This is the last stage in preparing data for machine learning tasks.

→ It involves transforming the preprocessed data into forms that are more suitable for a specific machine learning algorithm.

→ The data can be transformed through normalization/scaling, decomposition or aggregation.

→ I^st Normalization (Normalizing datasets)

→ Since most datasets contain features that vary greatly in terms of range, units or magnitude, input such unstandardized data into machine learning algorithms produces wrong predicted output

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

For ex

For prediction of diabetes risk we are using human height in meters & human weight in kilograms then weight attribute will the height attribute because of large numerical values.

(Years) Age	Salary (Rs)
24	10000
45	50000
36	30000

→ Therefore, scaling is necessary to decrease this effect by normalizing all features to a similar level of magnitude

→ In normalization, we scale attribute values to fall within a specified range

Techniques of Normalization

1) Min-Max Normalization

For ex →

Transform V in {min, max} to V' in {0, 1}

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Apply $V' = \frac{V - Min_A}{Max_A - Min_A}$ A → Attribute

	A	V'
V	2	0
	100	0.197
	5	0.006
	500	1

V' in [0,1] range

$V = 2$ $Min_A = 2$, $Max_A = 500$

$$V' = \frac{2-2}{500-2} = \frac{0}{498} = 0$$

2) Z-Score Normalization (Zero Mean)

→ In this method, Scaling is done by using mean & standard deviation

→ It is useful when min & max are unknown or when there are outliers

Here $V' = \frac{V - Mean_A}{Stand\ Dev_A}$ Stand Dev = 15000

A

V → 58000 Mean → 40000

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

$$V' = \frac{58000 - 40000}{15000} = 1.2$$

It reduce the effect of outliers

3) Decimal Scaling

$$V' = \frac{V}{10^J}$$ → Greater value

$J = 2$

	A
1	-80
2	8

V

$$V' = \frac{-80}{10^2} = -0.8$$

$$V' = \frac{8}{10^2} = 0.08$$

IInd Aggregation

City / Street → address

→ Combining two or more attribute into a single attribute

→ It reduces the number of attributes

→ Features aggregation can be performed to bring related features together & decrease

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

the dimensionality of ~~the~~ an input set

→ In some cases, combining multiple features into a single feature can be more useful for an algorithm.

III rd Decomposition:-

→ If some values in the data set are complicated, decomposing them into various constituent parts may be more meaningful to machine learning model.

Ex → If a date has a day component & a time component, & only the time part is relevant to the problem being addressed. So splitting day component can capture more specific date.

Data Visualization

→ It is the representation of data or information in a graph, chart or other visual format

→ It is important in machine learning because

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

it allows trends & patterns to be more easily seen.

→ With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data.

→ Machine learning makes it easier to conduct analysis & predicted which can then serve as helpful visualization to present

* Need of Data Visualization

→ We need because a visual summary of information makes it easier to identify patterns & trends then looking through thousands of rows on a spreadsheet

→ It will be more difficult to communicate the meaning of data without visualization

→ Without a visual representation, it can hard for the audience to understand the true meaning of the findings.

→ It is extremely difficult to determine the relationship between two variables

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

without a visualization.

→ Data visualization can be used to examining a market research

→ Marketing professionals need to know which audience to target with their message, so they analyze the entire market to identify audience clusters, influencers within clusters & outliers.

→ When planning out a schedule or timeline for a complex project, things get confusing.

→ A Gantt chart solves that issues by clearly illustrating each ~~test~~ task within the project & how long it will take to complete.

→ Data visualization can be used in risk analysis in which color coding a formula to show which opportunities are valuable & which are risky.

* Types of Data Visualization Chart

→ There are numerous tools available

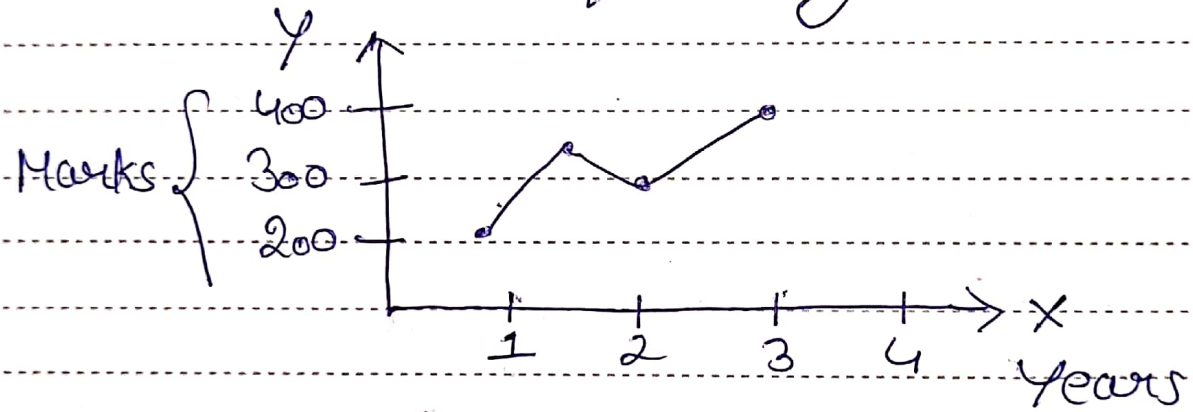
M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

to help create data visualization Some are non manual & some are automated, but either way they should allow you to make any of the following types of vizualization

i) Line Chart

It illustrates changes over time X-axis is usually a period of time while Y-axis is Quantity



ii) Area Chart

It represents the change in one or more quantities over time

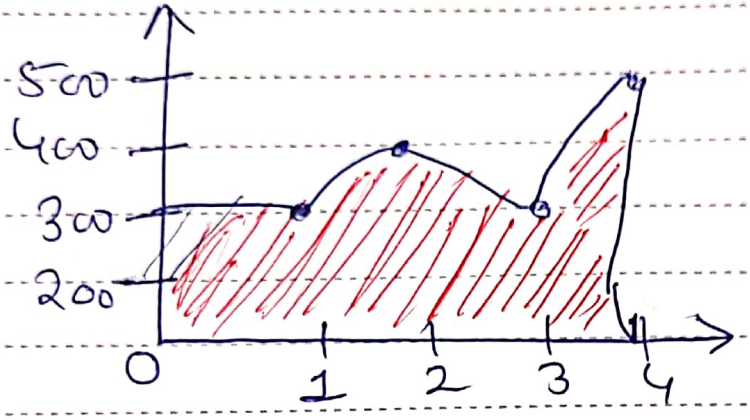
→ It is similar to a line graph

→ In both area chart & line graphs, data points are plotted & then connected by line segments to show the value of a quantity at several different times.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ Area Chart are different from line graphs because the area between X-axis & the line is filled with color or shading.



→ It is used when we want to show a trend over time, but not concerced with showing exact values.

iii) Bar Chart :-

→ It also illustrates changes over time.

→ If there is more than one variable a bar chart can make it easier to compare two data for each variable at each moment in time

→ A bar chart is a graph with rectangular bars (vertical / horizontal)

→ In horizontal bar chart, X-axis

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

represents the Categories & Y-axis represents a value for those Categories

→ A bar graph is useful for looking at a set of data & making comparisons.

iv) Histogram

→ It looks like a bar chart, but measures frequency rather that trends over time

→ It groups numbers into ranges

→ Bar chart & histogram ~~&~~ plot different types of data.

→ To plot discrete data, bar chart is used

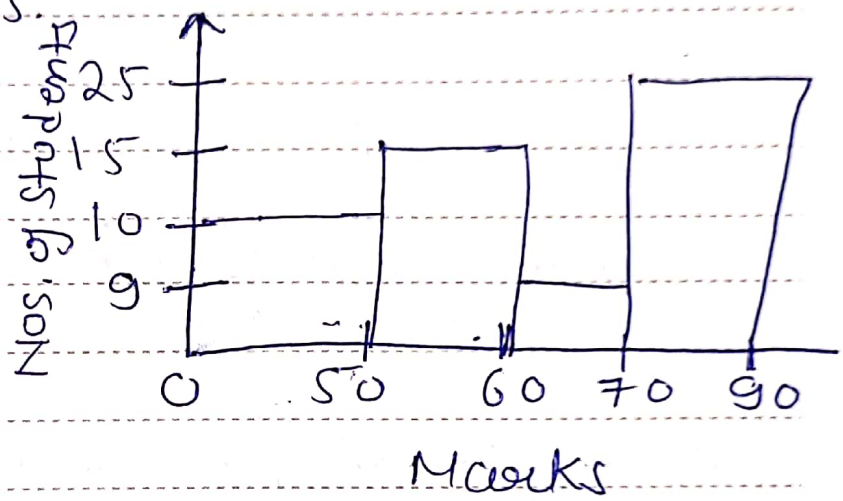
M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ To plot continuous data (range) histogram is used.

→ In bar chart, its a good choice when we want to compare things between different groups.

0-50 : 10
50-60 : 15
60-70 : 9
70-80 : 25



The height of each bar shows how many falls into each range.

V) Scatter plot :-

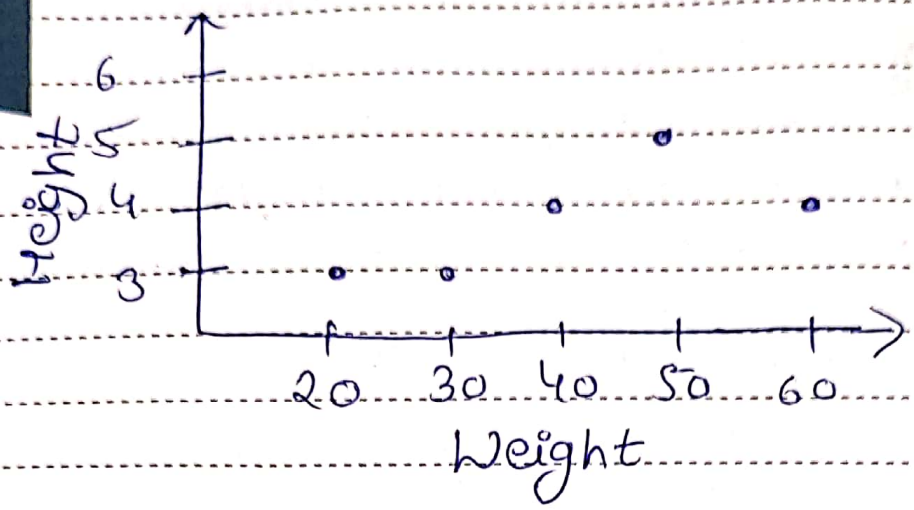
A scatter plot has points that show the relationship between two sets of data.

→ Each point on a scatter plot means when x equals this then y equals this.

→

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner



Each dot show one person's weight versus their height.

→ If the points trend a certain way like upward to the left, downward to the right etc there is a relationship between them.

→ If the plot is truly scattered with no trend at all, then two variables do not affect each other at all.

Other ex → Pie chart, Bubble chart, Map frame diagram etc.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

Hypothesis Testing

→ Hypothesis are statement about the given problem. It is a statistical method that is used in making a statistical decision using experiment data.

→ Hypothesis testing is basically an assumption that we make about a population parameter

→ It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data

Ex

You say an average student in the class is 30 or a boy is taller than girls All those are an example in which we assume or need some statistic way to prove those. We need some mathematical conclusion whatever we are assuming is true

Need for Hypothesis testing

→ It is an important procedure in statistics.

→ It evaluates two mutually exclusive population statement to determine which statement

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

is ~~more~~ most supported by sample data. When we say that the findings are statistically significant, it is thanks to hypothesis testing.

Parameters of hypothesis testing

i) Null Hypothesis (H_0):-

In statistics, the null hypothesis is a general given statement or default position that there is no relationship between two measured cases or no relationship among groups.

In another words, it is basic assumption or made based on the problem knowledge. Ex→ A Company production is =50/day etc.

ii) Alternative Hypothesis (H_1):-

→ It is contrary to the null Hypothesis

Ex→ A Company production is not equal to 50 per day etc.

iii) Level of Significance

It refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting a hypothesis, so, we

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

therefore, select a level of significance i.e usually 5%. This is normally denoted with "α" & generally it is 0.05 or 5% which means your output should be 95% confident to give similar kind of result in each sample.

iv) P-value:-

The p-value, or calculated probablity is the probablity of finding the observed/extreme results when the null Hypothesis (H_0) of a study given problem is true.

→ If your P-value is less than the chosen significance level then you reject the null hypothesis. i.e. accept that your sample claims to support the alternative hypothesis.

Example:-

Given a coin & it is not known whether that is fair or tricky, so let's decide null & alternate hypothesis.

→ Null hypothesis (H_0): A coin is a fair coin
→ ~~Null~~ Alternative hypothesis (H_1): A coin is a tricky coin.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

→ $\alpha = 5$

Now, let's toss the coin & calculate p-value (probablity value)

→ Toss a coin 1st time & assume that result is head P-value = 50 (as head & tail have equal probablity)

→ Toss a coin 2nd time & assume that result again is head, now p-value $= 50/2 = 25$.

and similarly, we toss 6 consecutive time & got the result as all heads, now P-value $= 1.5$. But we set our significance level as 95, error rate we allow & here we see we are beyond that level i.e. our null-hypothesis does not hold good so we need to reject & propose that this coin is a tricky coin which is actually because it gives us 6 consecutive heads.

Error in hypothesis testing

i) Type I error :- When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

ii) Type II errors: When we accept the null hypothesis but it is false. Type II error are denoted by beta.

Probablity in Machine Learning

→ It is the bedrock of ML, which tells how likely is the event to occur.

→ The value of probablity always lies between 0 to 1. It is the core concept as well as a primary prerequisite to understanding the ML models & their applications.

→ Probablity can be calculated by the number of times the event occurs divided by the total number of possible outcomes.

→ Ex:- Let's suppose we tossed a coin, then the probablity of getting head as a possible outcome can be calculated as below formula:

$$P(H) = \frac{\text{Number of ways to head occur}}{\text{total number of possible outcomes}}$$

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

$P(H) = 1/2$
$P(H) = 0.5$

Where,
$P(H)$ = Probablity of occuring head as outcome while tossing a coin

* Types of Probablity

It can be categorized further in different types as follows:-

i) Empirical Probablity:-

It can be calculated as the number of times the event occurs divided by the total number of incidents observed

ii) Theoretical Probablity

It can be calculated as the number of ways the particular event can occur divided by the total number of possible outcomes.

iii) Join Probablity

It tells the probablity of simultaneously occuring two random events.

$$P(A \cap B) = P(A) \cdot P(B)$$

Where,

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

$P(A \cap B)$ = Probablity of occuring events A & B both
$P(A)$ = Probablity of event A
$P(B)$ = Probablity of event B

iv) Conditional Probablity:-
It is given by the probablity of event A given that event B occured.

The probablity of an event A conditioned on an event B is denoted & defined as
$$P(A/B) = P(A \cap B) / P(B)$$
Similarly, $P(B/A) = P(A \cap B) / P(A)$.
We can write the join probablity of as A & B as $(A \cap B) = P(A) . P(B/A)$. which means: The chance of both things happening is the chance that the first one happens, & then the second one is given when the first thing happened.
~~We have~~.

Statistics in Machine Learning
Statistics is also considered as the base foundation of machine learning. which deals with finding answers to the questions that we have about data. In general, we can define statistics as:

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

Statistics is the part of applied Mathematics that deals with studying & developing ways for gathering, analyzing, interpreting & drawing conclusions from empirical data. It can be used to perform better-informed bussiness decisions.

Statistics can be categorized into two major parts.

1) Descriptive Statistics
2) Inferential Statistics.

* Use of Statistics in ML

Statistics methods are used to understand the training data as well as interpret the results of testing different machine learning models. Further Statistics can be used to make better informed bussiness & investing decisions.

* With this basic understanding, it's time to dive deep into learning all the crucial concepts related to statistics for machine learning

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

1) Inferential Statistics

It involves mathematical estimates that allow us to infer on a pattern or trend based on the sample data sets of a larger population. Helps to generalize conclude & predict a bigger population

2) Descriptive Statistics:-

It refers to methods for summarizing & organizing the information in a data set.

We will use below table to describe some of the statistical concepts.

Characteristics of 10 loan applicants

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
1	Single	Y	38,000	3	2009	Good
2	Married	Y	32,000	7	2010	Good
3	Other	N	25000	9	2011	Good
4	Other	N	36000	3	2009	Good
5	Other	Y	33000	4	2010	Good
6	Other	N	24000	10	2008	Bad
7	Married	Y	25100	8	2010	Good
8	Married	Y	48000	1	2007	Good
9	Married	Y	32100	6	2009	Bad
10	Married	Y	32,200	5	2010	Good

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

* Some terminology used in Statistics for ML

1) Elements :-

The entities for which information is collected are called the elements. In above table, the elements are the 10 applicants. Elements are also called cases or subjects

2) Variables:

The chacteristics of an element is called a variable. It can take different values for different elements. e.g. marital status, mortgage, income, rank, year, & risk. Variables are also called attributes

It can be either qualitative & quantitative

→ Qualitative :-

A qualitative variable enables the elements to be classified or categorized according to some characteristic. The qualitative variables are marital status, mortgage, rank & risk. Its also called categorical variables.

→ Quantitative :-

A quantitative variable takes numeric value & allows arithmetic to be meaningful performed on it. Ex→ Income & year. It is also called numerical values

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

3) Discrete Variable

A numerical variable that can take either a finite or a countable number of values is a discrete variable, for which each value can be graphed as a separate point, with space b/w each point. 'year' is an example

4) Continuous Variable:-

A numerical variable that can take infinitely many values is a continuous variable, whose possible values form an interval on the number line, with no space between the points "income" is an example of a continuous variable.

5) Population:-

It is the set of all elements of interest for a particular problem. A parameter is a characteristic of a population.

6) Sample:-

It consist of a subset of the population. A characteristic of a sample is called a statistic

7) Random Sample:-

When we take a sample for which

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

each element has an equal chance of being selected.

* Measures of Center: Mean, Median, Mode, Mid-Range

⇒ Indicate where on the number line the central part of the data is located

8) Mean:-

It is the arithmetic average of a data set. To calculate the mean, add up the values and divide by the number of values. It is denoted by μ

9) Median:-

It is the middle data value, ~~where~~ when there is an odd ~~value~~ number of data values, & the data have been sorted into ascending order.

If there is an even number, the median is the mean of the two middle data values.

10) Mode:-

It is the data value that occurs with the greatest frequency. Both

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

quantitative & categorical values can have modes, but only quantitative variables can have means or medians. Each income value occurs only once, so there is no mode. Ex→ The mode of the year is 2010, with a frequency of 4.

11) Mid-range:-
It is the average of the maximum & minimum values in a data set. The mid-range income is

mid-range (income) = (max(income) + min (income))/2 = $\frac{(48000+24000)}{2}$

= $36000

Measures of Variability:- Range, Variance, Standard Deviation

⇒ Quantify the amount of variation, spread or dispersion present in the data

12) Range:-
The range of a variable equals the difference between the maximum & minimum values. Ex→ The range of income is:

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

Scanned with OKEN Scanner

range (income) = max(income) − min(income)
= 48000 − 24000 = $24000

Range only reflects the difference between largest & smallest observation, but it fails to reflect how data is centralized.

13) Variance :-
Population variance is defined as the average of the squared differences from the Mean, denoted as σ^2 ("sigma-squared"):

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

Larger variance means the data are more spread out

The sample variance S^2 is approximately the mean of the squared deviations, with N replaced by n−1.
This difference occurs because the sample mean is used as an approximation of the true population mean

$$S^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner

14) Standard Deviation:-

The sd of a bunch of numbers tells you how much the individual numbers tend to differ from the mean.

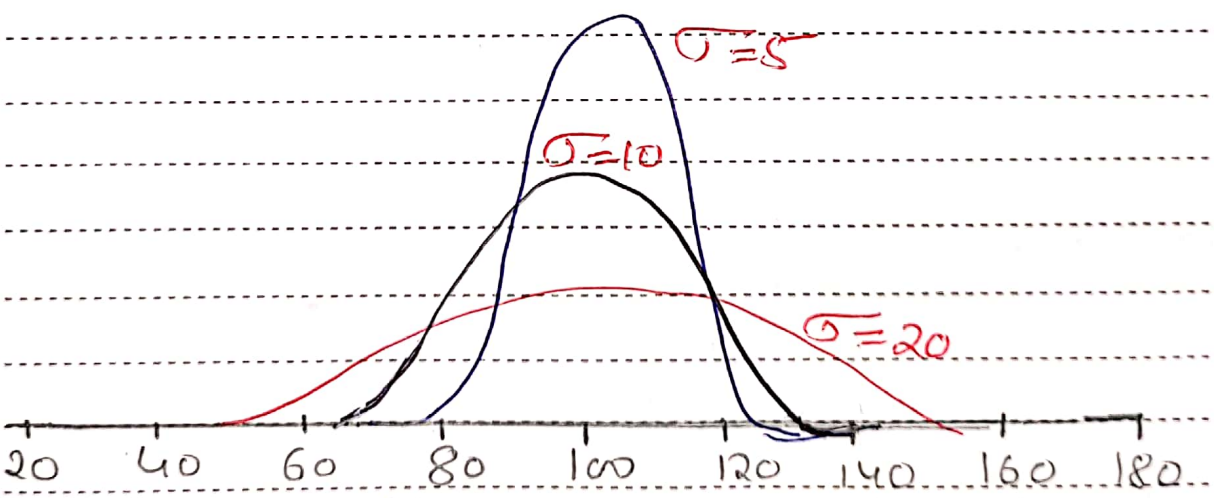
The sample standard deviation is the square root of the sample variance:

$$Sd = \sqrt{s^2}$$

For ex → income, deviate from their mean by $7201.

The population standard deviation is the square root of the population variance.

$$Sd = \sqrt{\sigma^2}$$



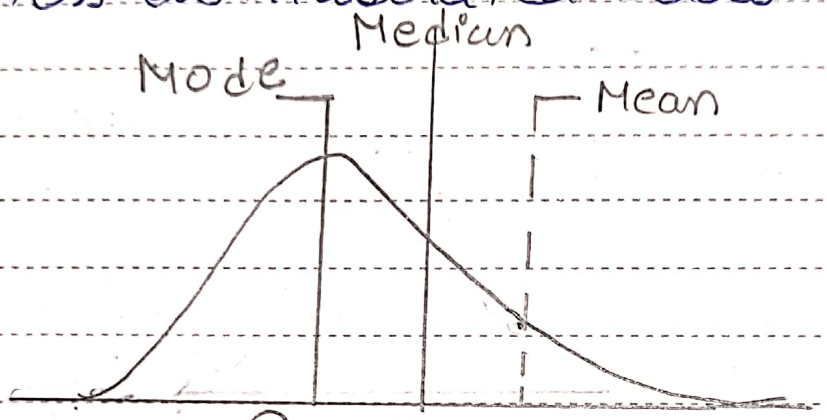
The smaller the standard deviation, narrower the peak, the data points are closer to the mean. The further the data points are from the mean, the greater the standard deviation.

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

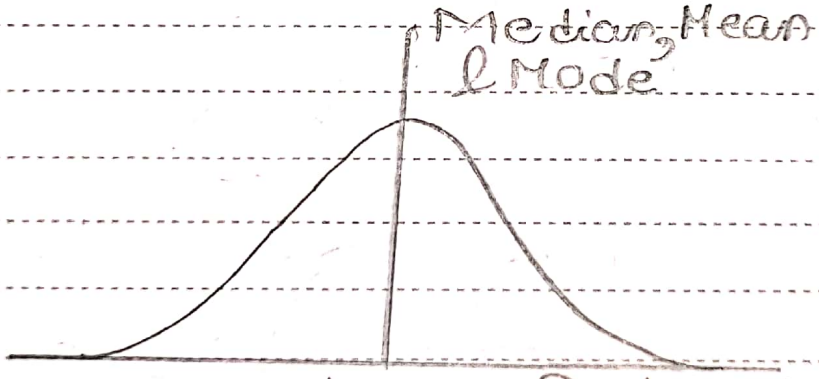
Scanned with OKEN Scanner

15) Skewness:-

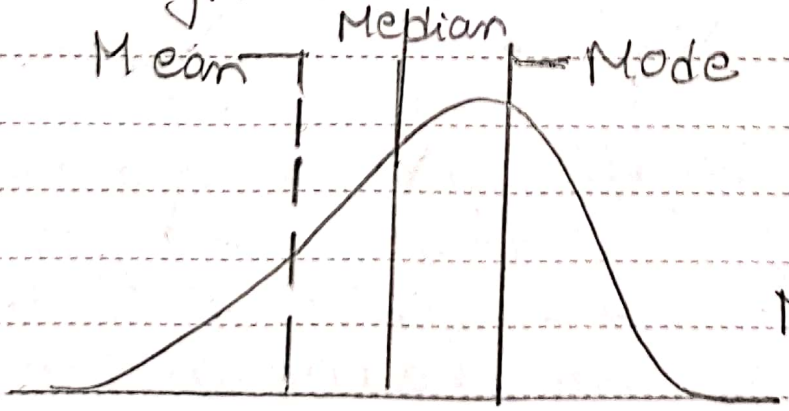
It gives us a measure of distortion from symmetry (skew). Depending on whether the left or right tail is skewed for give data distributions, skewness is classified into positive & Negative Skewness as illustrated below



Positive Skew



Symmetrical Distribution (Normal)



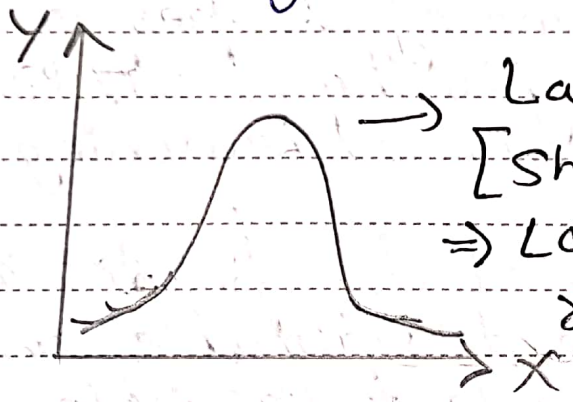
Negative Skew

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL

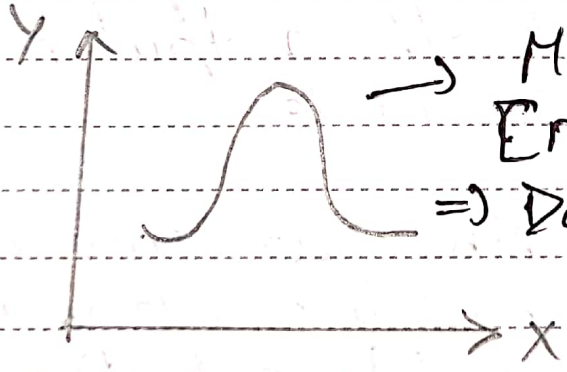
Scanned with OKEN Scanner

16) Kurtosis:-

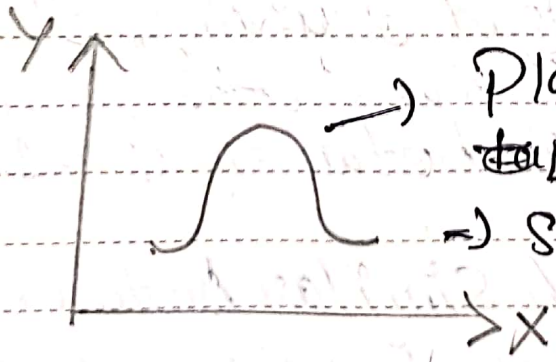
It relates bulliginess or tailedness of normal distribution. It is a statistical measure that defines heavily the tails of a distribution differ from the tails of a normal distribution. It generally defines whether the tails of extreme values.



→ Lapto kurtic
[sharp tails]
⇒ Larger level of risk investment



→ Mesokurtic
[normal tails]
⇒ Data follows normal distribution



→ Platykurtic
[small tails or flat
⇒ Small level of risk investment

M/s Agrawal Construction Co. | SIRT SAGAR GROUP OF INSTITUTIONS | SAGAR INTERNATIONAL SCHOOL | SAGE UNIVERSITY INDORE | SAGE UNIVERSITY BHOPAL | AGRAWAL

Scanned with OKEN Scanner