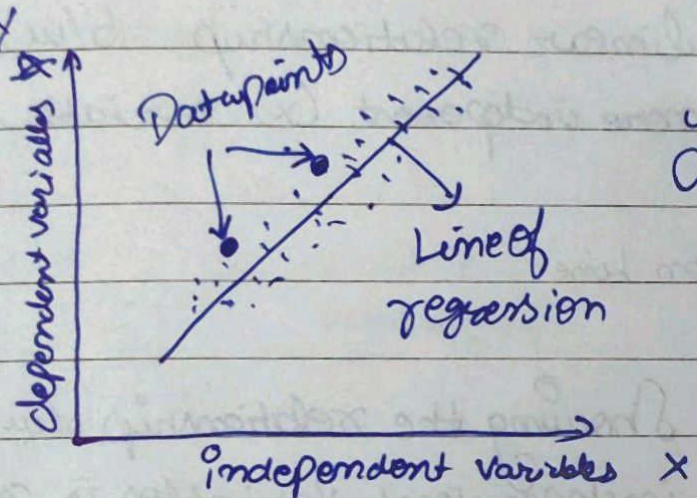
Unit-1

Linear Regression

- Dependent Variable is continuous in nature.



$$y = 0.9x + 1.2x_1 + 2x_2 + 4x_3 + 1x_4$$

Simple Linear equation

$$y = \alpha_0 + \alpha_1 x_1 \quad (y = c + mx)$$

Multiple Linear Equation

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 - - - - \alpha_m x_m$$

α_i = Reg. Coeff.
x_i = Independent variables.
y = Dependent Variable.

→ Linear equation makes predictions for continuous, real or numeric variables such as Sales, Salary, age, product price etc.

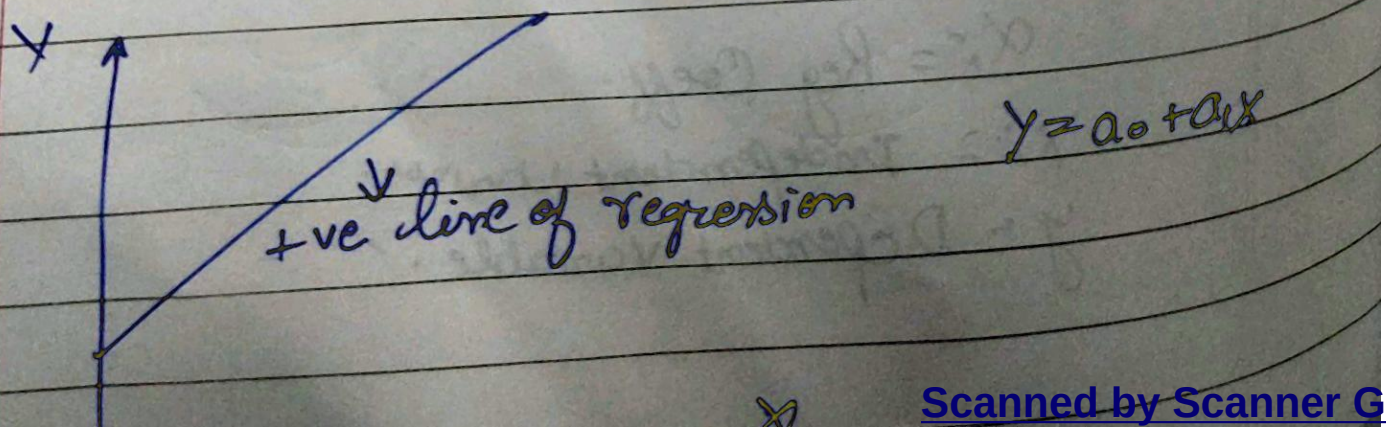
→ It shows a linear relationship b/w a dependent (y) and one or more independent (x) variables, i.e.

Linear Regression Line

A linear line showing the relationship b/w the dependent and independent variables is called a regression line.

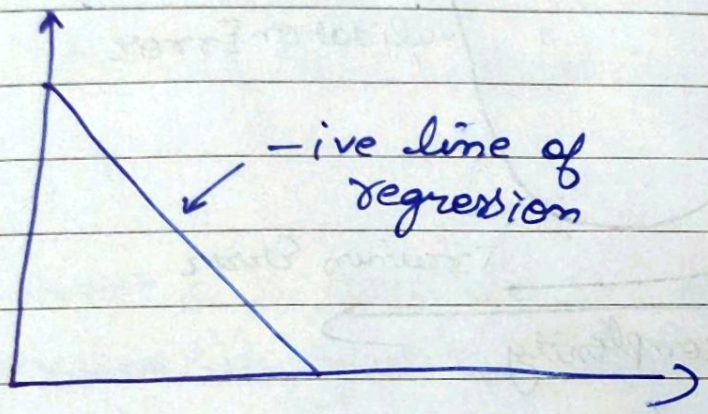
- Positive Linear Relationship

If the dependent variable increases on the Y axis and independent variable increases on X-axis, then such a relationship is termed as a positive linear relationship.



- Negative Linear Relationship

If the dependent Variable decreases on the Y-axis and independent Variable increases on the X-axis then such a relationship is called a negative linear relationship.



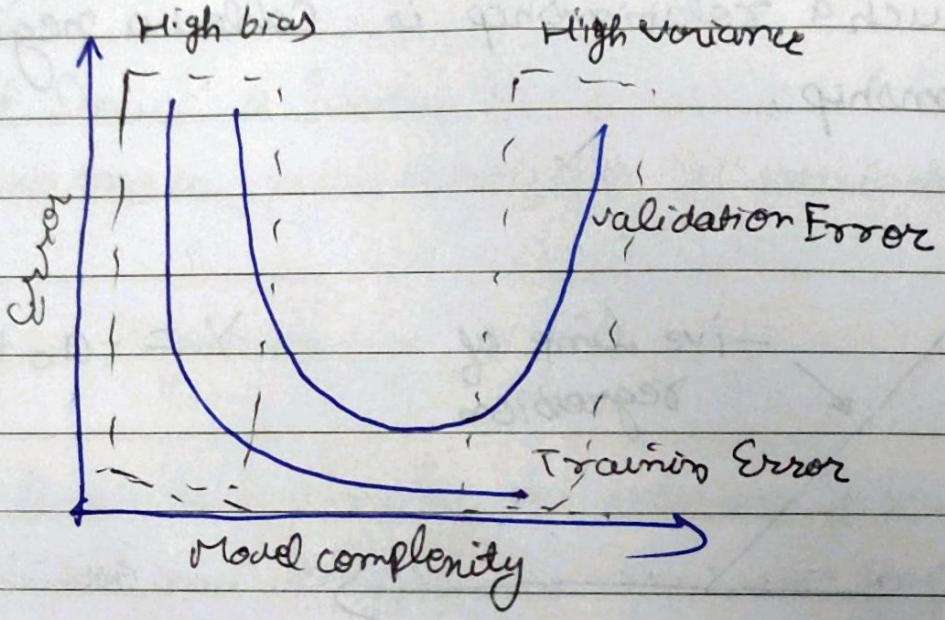
$Y = -a_0 + a_1 x$

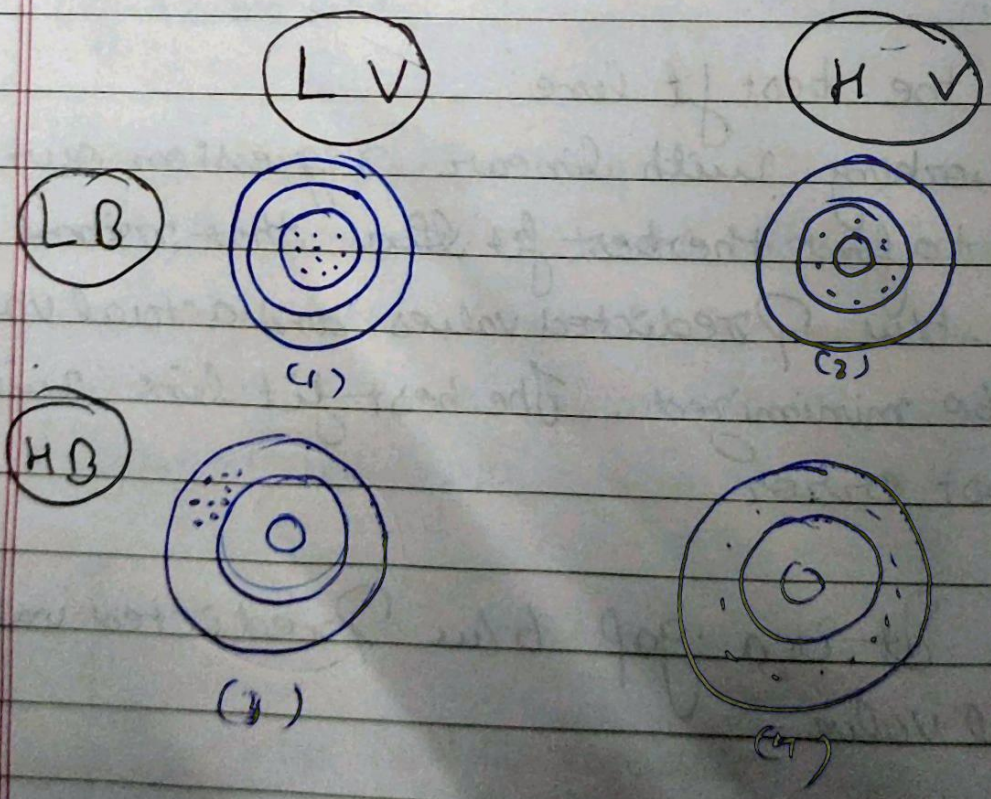
Finding the best fit line:

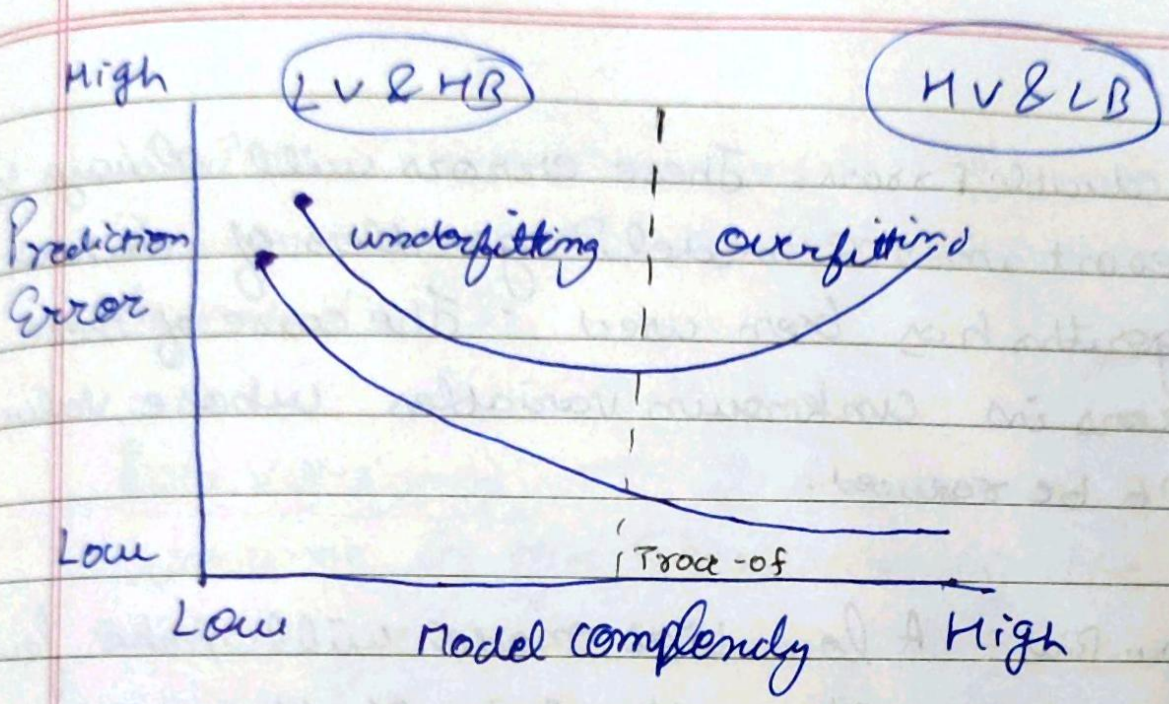
When working with linear regression our main goal is to find the best fit line that means the error b/w Predicted values and actual values should be minimized. The best fit line will have the least error.

Bias → It is a gap b/w Predicted value & actual value.

Bias is a Phenomenon that Skews the result of an algorithm in favour or against an idea.

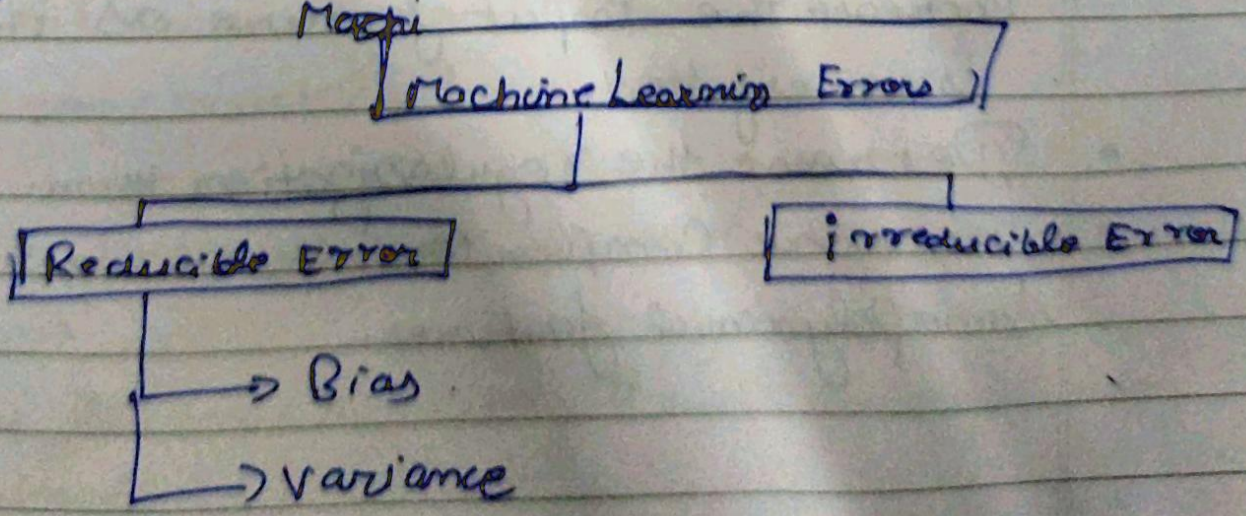






An error is a measure of how accurately an algorithm can make predictions for the previously unknown dataset.

Reducible errors: These errors can be reduced to improve the model accuracy. Such errors can further be classified into bias & variance.



Irreducible Errors: These errors will always be present in the model regardless of which algorithm has been used. The cause of these errors is unknown variables whose value can't be reduced.

Low Bias: A low bias model will make fewer assumptions about the form of the target function.

High Bias: A model with a high Bias makes more assumptions, and the model becomes unable to capture the important features of our dataset.

Ways to reduce high bias:

- increase the input feature as the model is underfitted.
- Decrease the regularization term.
- Use more complex models, such as including some polynomial features.

Variance:- Variance tells that how much a random variable is different from its expected value.

Low variance → It means there is a small variance in the prediction of the target func with changes in the training data set.

High variance → It shows a large variation in the prediction of the target func with changes in the training dataset.

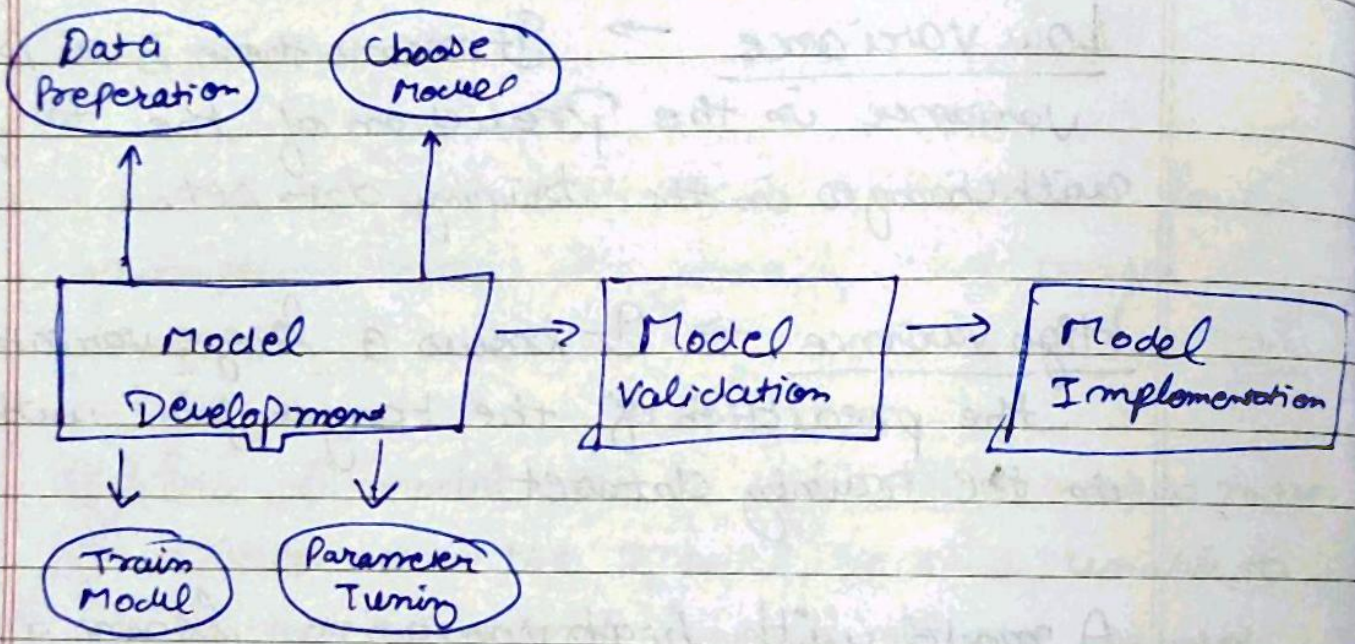
A model with high variance learns a lot and perform well with the training dataset, and does not generalize well with the unseen dataset.

- A high variance model leads to overfitting
- Increase model complexities.

- If we decrease the variance, it will increase the bias
- If we decrease the bias, it will increase the variance.

Model Validation

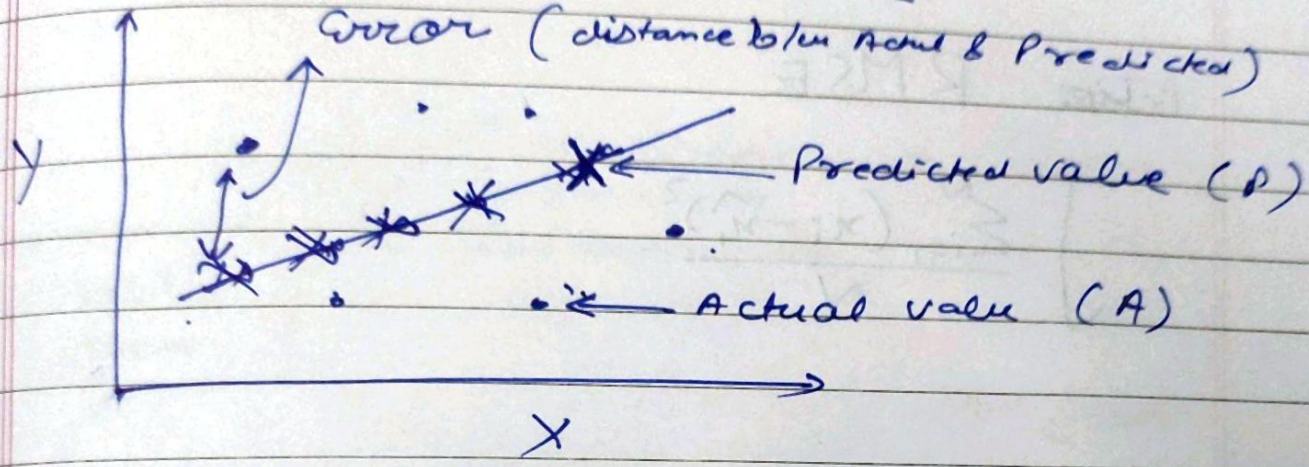
- A model should perform well in the new data



Model validation is the process that is carried out after Model Training where the trained model is evaluated with a testing data set.

The testing data may or may not be a chunk of the same data set from which the training set is procured

Mean Squared Error



$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Model A

Actual	Predicted	Error	Error Sqrd
10	9	1	1
13	14	-1	1
14	12	2	4
9	10	-1	1
17	15	2	4
		TE(7)	11
		MSE	2.2

Model B

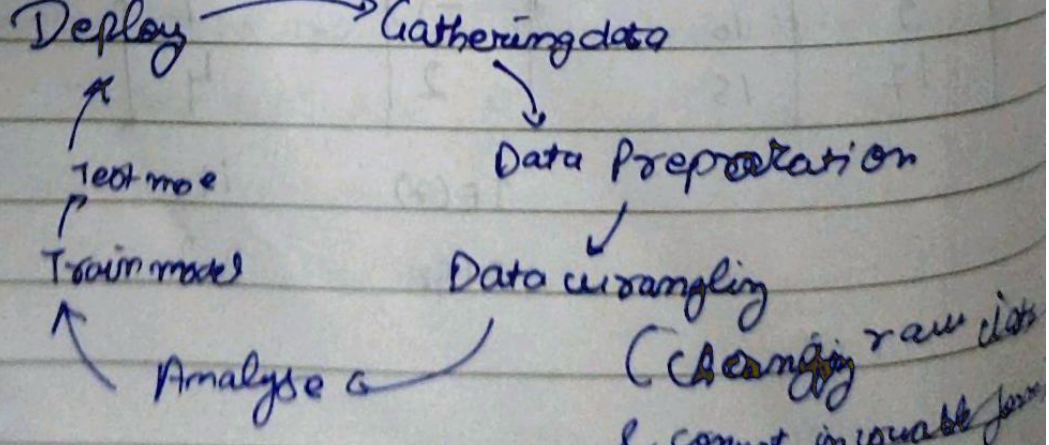
Actual RMSE

$$\sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x_i})^2}{N}}$$

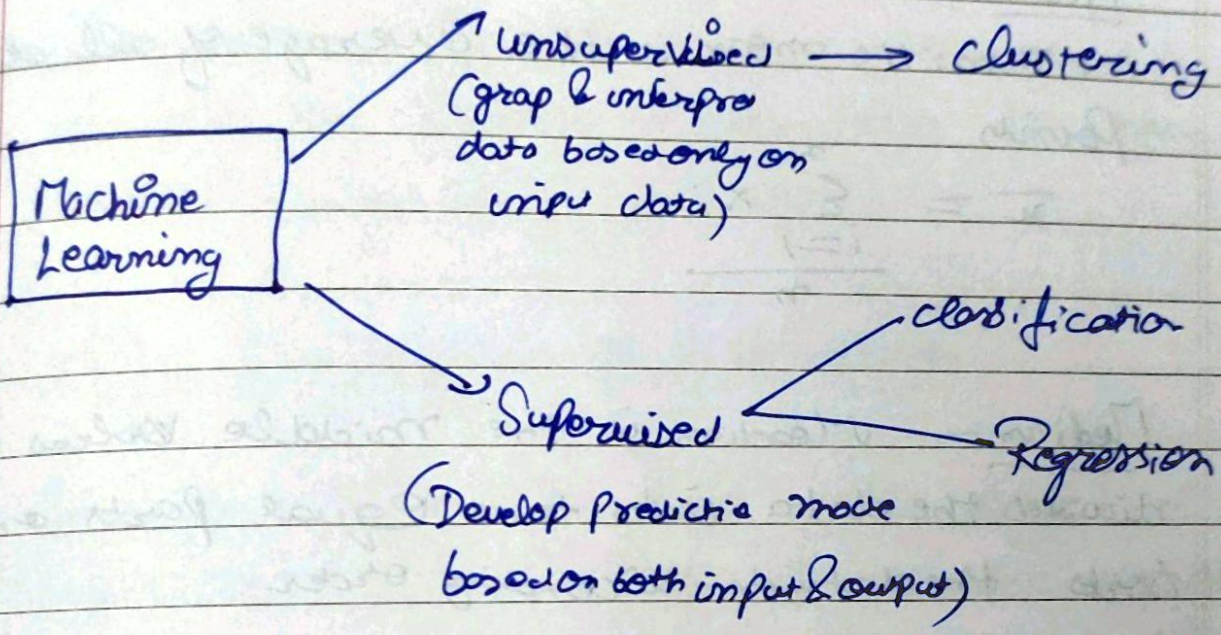
Types of ML:-

i) Supervised - labelled data for training
ii) unsupervised - unlabelled data for training
iii) Reinforcement - Algo discovered data through a proc of trial & error then decides what action results in higher rewards

ML life cycle:



ML techniques



Regression

Help predictor interpret a particular numerical value based on prior data.

Classification

Predict or explain a class value.

Clustering → Aim to group or group observations with similar characteristics

Statistics & Exploratory Data Analysis

- **Mean** -
Arithmitic mean is the average of all data points

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Median** – Median is the middle values that divides the data into two equal parts once it sorts the data in ascending order.

If total number of data points (n) is odd, the median is the value at position $\left(\frac{n+1}{2}\right)$.

- **Mode** – Mode is value that occurs most frequently in the data set.

* **Correlation** - Explains how one or more variables are related to each other.

i) +ve correlation - when value of one variable increases then the value of other variable also increases.

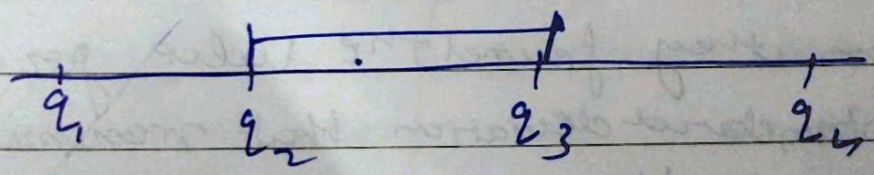
ii) -Ve Correlation - When value of one variable increase the value of other variable decrease.

* Covariance - Is a measure for how two variables are related to each other i.e. how two variables vary with each other.

Diff. b/w Covariance & correlation -

Covariance only tells about the directional relation b/w two variable but correlation along with direction also tells its strength of that relation

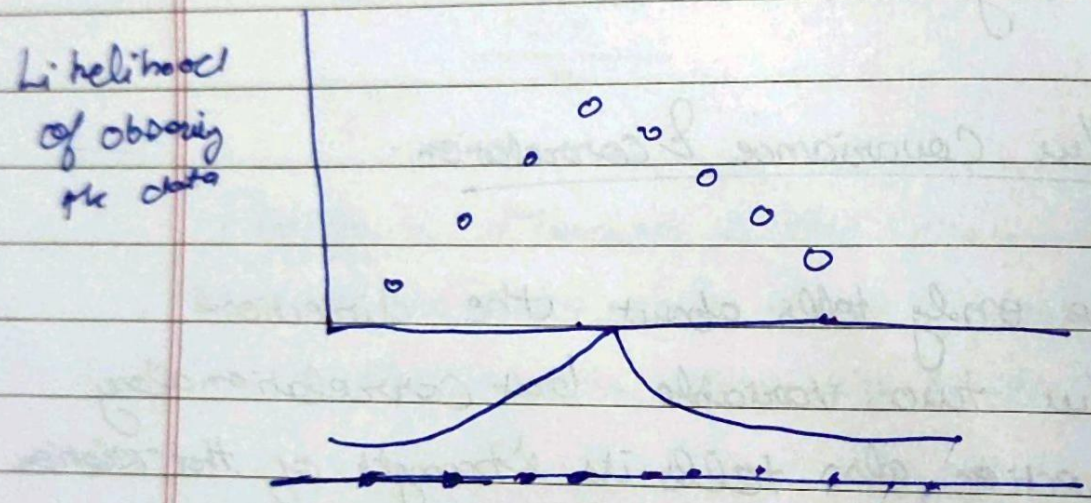
* Quartile: Quartile are position indicators that divides a sequence into a equal parts.



IQR = inter quartile range

Maximum likelihood

Goal of maximum likelihood is to find the optimal way to fit a distribution to the data.



When someone says that they have the maximum likelihood estimates for the mean are the standard deviation or something else then they found the value for the mean the standard deviation that maximizes the likelihood that you observed the things you observed.

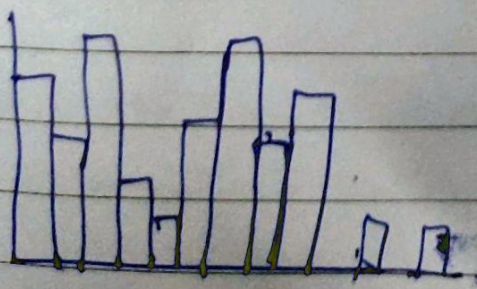
Bayesian Inference

Bayesian Inference technique specify how one should update one's beliefs upon observing data.

⇒ Machine learning is mainly concerned with prediction and prediction is very much concerned with probability.

→ The frequentist definition of probability is based on frequencies of events whereas the Bayesian definition of probability is based on our knowledge of events.

Ex: we have 50 data sample $x_1, x_2 \cdots x_{50}$ is exponential distribution $(\lambda e^{-\lambda x})$



What is estimating λ?

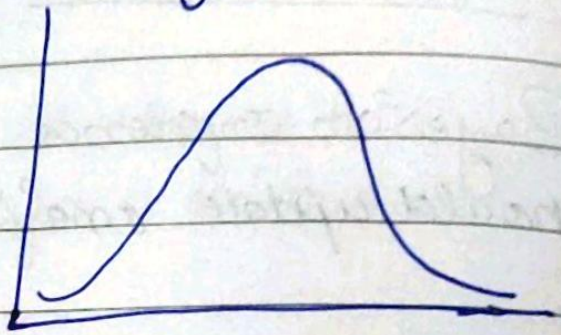
Point estimation (Eg. – maximum likelihood estimation)

Single best estimation

$$\lambda = \frac{1}{(x_1 + x_2 + \cdots x_n)/50}$$

Interval estimation

eg. - Bayesian Inference



So,

→ Bayesian inference is an interval estimation

→ It is a process of improving distribution function -

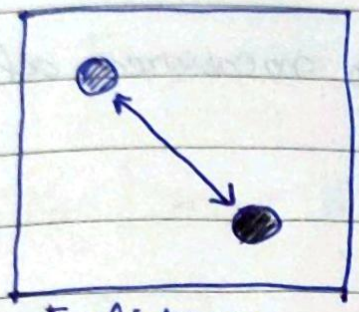
$$Pr(\lambda) \Rightarrow Pr(\lambda \mid \text{observed data})$$

Distance Matrices

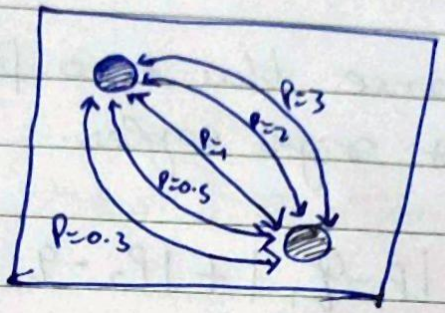
1. Euclidean Distance
2. Manhattan Distance
3. Minkowski Distance
4. Hamming Distance

Shadow Paging Deadlock CScan C-look elevator

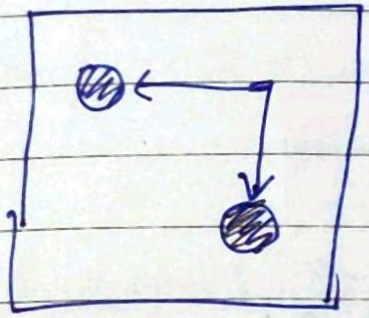
Date
Page

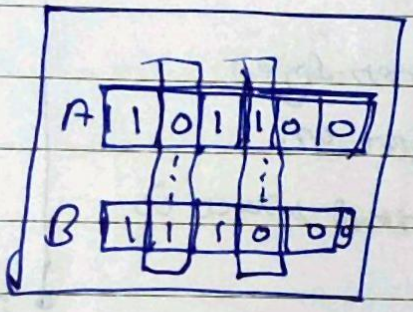


Euclidean



Minkowski





Hamming

⇒ Euclidean Distance

Euclidean Distance b/w ~~these~~ represents Shortest distance b/w two points.

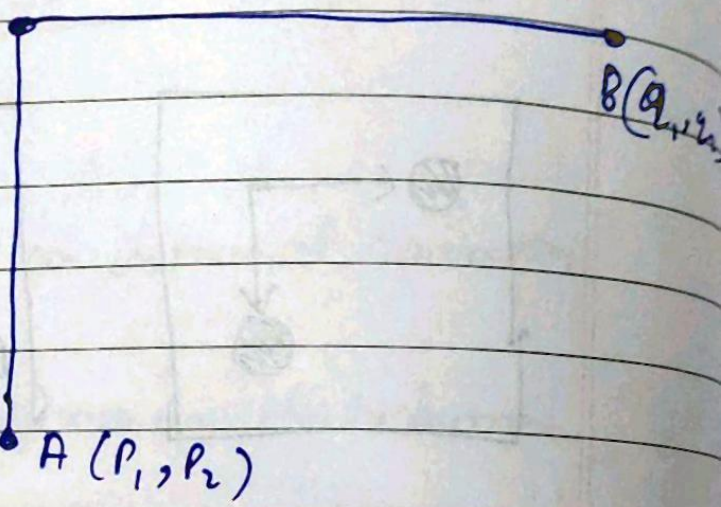
$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

⇒ Manhattan Distance

- It is the sum of absolute differences b/w points across all the dimensions.

The distance b/w two points measured along axes at right angles.

$$d = |P_1 - q_1| + |P_2 - q_2|$$



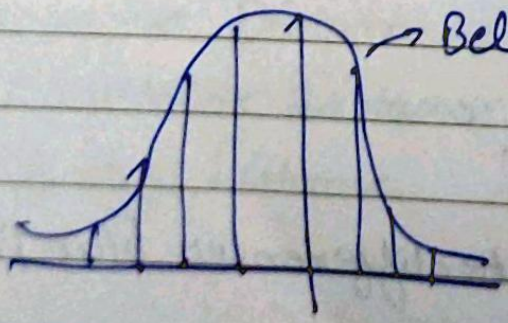
And the generalized formula for an n-dimensional space is given as

$$D_m = \sum_{i=1}^{n} |P_i - q_i|$$

Normal / Gaussian Distribution

$x \sim G.D(\mu, \sigma)$ σ^2 → Standard deviation

mod



Emperical Formula

(1) $Pr[\mu - \sigma \leq x \leq \mu + \sigma] \simeq 68\%$

(2) $Pr[\mu - 2\sigma \leq x \leq \mu + 2\sigma] \simeq 95\%$

(3) $Pr[\mu - 3\sigma \leq x \leq \mu + 3\sigma] \simeq 99.7\%$

Hypothesis Testing -

Evaluates 2 or more mutual exclusive statement on population using sample data

→ Null hypothesis - It is regarding the assumption that there is no anomely pattern or believing according to the assumption made

→ Alternate hypothesis - Contrary to the null hypothesis it shows that observation is the result of real effect.

Steps:-

i) Make initial assumption (H_0 - null hypothesis)
ii) Collect data (evidence)
iii) Gather evidence to reject or not reject null hypothesis.

	H_0	H_1
Do not Reject	OK	Type 2 error
Reject	Type 1 error	OK

Type 1 error: False Positive (Reject H_0 when it's true)

Ex - The test says you have covid but you actually don't

Type 2 error:- False negative (Accept H_0 when it is false).

Ex - The test result says you don't have covid but you actually do.

* P-value - Significance value.
It tells us how likely our data could have occured under the null hypothesis.

→ Correctly it is taken as 5%

→ If p value is 0.05 that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hype was true.

Dealing with missing values

→ Ignore the tuple
→ Filling Manually
→ Using Global Constants "NULL" / "N/A"
→ use Attribute Mean / Median
→ Mean Normal distribution
Median Skewed distribution (Asymmetric)
→ use of data mining Algorithm
→ inference tool, decision tree, clustering Algorithm.

Regression Equation:-

Q) From the given ~~equation~~ data, calculate regression equations taking deviation of items from the mean of X and Y series

X	1	2	3	4	5
Y	2	5	11	8	14

Theory Introduction to Regression Equation:-

i) Regression eq^n of Y on X
ii) Regression eq^n of X on Y

Y on X → Variation in Y when the changes in X is given.

X on Y → Variation in X when the changes in Y is given.

i) Regression equation of X on X

$$x - \bar{x} = \frac{r \cdot \sigma_x}{\sigma_y}(y - \bar{y})$$

$\frac{r \cdot \sigma_x}{\sigma_y}$ → regression coff

$$b_{xy} = \frac{\gamma \cdot \sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}$$

ii) Regression eq^n of y on x

$$y - \bar{y} = \frac{\gamma \sigma_y}{\sigma_x}(x - \bar{x})$$

$$b_{yx} = \frac{\gamma \sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}$$

γ = correlation coff

$$\gamma = \sqrt{b_{xy} \cdot b_{yx}}$$

Sol^n

X	$x = (x - \bar{x})$	x^2	Y	$y = (y - \bar{y})$	y^2	XY
1	-2	4	2	-6	36	12
2	-1	1	5	-3	9	3
3	0	0	11	3	9	0
4	1	1	8	0	6	0
5	2	4	14	6	36	12
$\Sigma X = 15$		$\Sigma x^2 = 10$	$\Sigma Y = 40$			$\Sigma xy = 27$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{15}{5} = 3 \qquad \bar{Y} = \frac{\Sigma y}{N} = \frac{40}{5} = 8$$

1) Regression equation of X on Y

$$(x-\bar{x}) = \frac{\gamma \sigma_x}{\sigma_y}(y-\bar{y})$$

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

$$(x-3) = \frac{27}{90}(y-8)$$

$$(x-3) = 0.3(y-8)$$

$$x-3 = 0.3y - 2.4$$

$$x = 0.3y + 0.6$$

2) Regression equation of Y on X

$$(y-\bar{y}) = \frac{\gamma \sigma_y}{\sigma_x}(x-\bar{x})$$

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

$$(y-8) = \frac{27}{10}(x-3)$$

$$y-8 = 2.7x - 8.1$$

$$y = 2.7x - 0.1$$

$$r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{\frac{\Sigma xy}{\Sigma y^2} \times \frac{\Sigma xy}{\Sigma x^2}}$$

I-test

A t-test is a type of inferential statistic used to determine the significant diff b/w the means of two groups which may be related to certain features.

→ A t-test is only used when comparing the means of two groups also known as a pairwise comparison

Assumption

1. A t-test assumes your data are independent
2. Assumes your data are normally distributed.
3. Assumes data have a similar amount of variance within each group.

Type of t-test

i) One Sample t-test

we compare the average of one group against the set average.

$$t = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

t = t-statistic
m = mean
μ = population mean
s = standard deviation
n = group size

ii) Unpaired or independent t-test -

Compare the means of two different groups of sample

$$t = \frac{m_A - m_B}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$$

m_A & m_B - means of two diff groups
n_A & n_B - sample sizes

$$s^2 = \frac{\Sigma(x - m_A)^2 + \Sigma(x - m_B)^2}{n_A + n_B - 2}$$

Degree of freedom

Refers to the values in a study that can vary and are essential for assessing the null hypothesis importance & validity.

Paired t-test

we measure one group at two different times. we compare different means for a group at two different times or under two different conditions.

$$t = \frac{m}{\frac{s}{\sqrt{n}}}$$

Chi - Square Test :-

This helps to determine whether there is a notable difference b/w the normal frequencies and observed freq in one or more classes or categories.

Date
Page

Properties

→ Two times the number of degree of freedom is equal to the variance.

→ The number of degree of freedom is equal to mean distribution

→ Chi-square distribution curve approaches the normal distribution when the degree of freedom increases.

Formula

$$\chi^2 = \frac{\Sigma(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

ANOVA → Analysis of variance test to generate the T-tests for multiple groups.

Assumptions-

1) Data for every level of the factor is distributed generally.

2. Case independent - The sample cases must be independent from each other.

3. Variance b/w the group needs to be around equal.

Types:-

i) One-way → Analysis of Variance Test which that only one independent varible

ii) Two-way → ANOVA that has two independent variables. Also known as factorial anova test.

iii) n-way - use more than two independent variables.

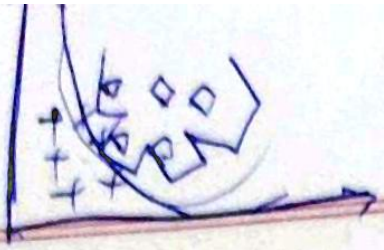
Missing Value Treatment -

Mainly missing values can be treated in two ways.

i) Deletion -

To either delete the entire row or column in which missing value is present.

(Not recommended).



ii) Imputing

→ Replacing with an arbitary value
→ Replacing with mean
→ Replacing with mode
→ Replacing with median
→ Replacing with prewous values

For Categorical data impur most frequent value.

Outlier detection

- whether user-labeled examples of outliers can be obtained
  - Supervised, Semi-Supervised and unsupervised methods
- Assumptions about normal data and Outliers
  - Statistical, Proximity-based and clustering-based methods

→ among the data objects, on which does not obey general behaviour.

* Outlier Detection

Process of detecting outliers and subsequently removing them

1. Statistical Approach :

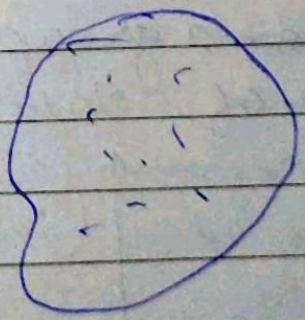
based on probability of the data points low probability - outliers

- Parametric methods
- Non Parametric methods

2 Proximity Approach :

based on location of data points.

- Density based approach
- Distance " "
- Grid " "
- Deviation " "

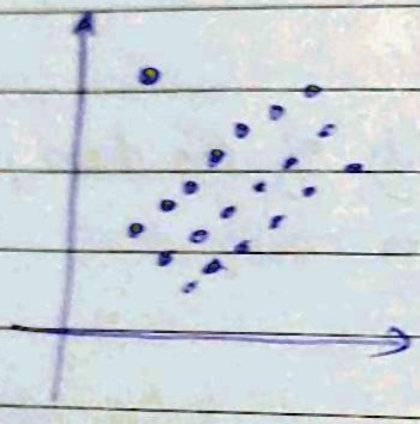


- 2.2276 -

* Types of outliers

① Types of outliers

1. Global / Point outliers

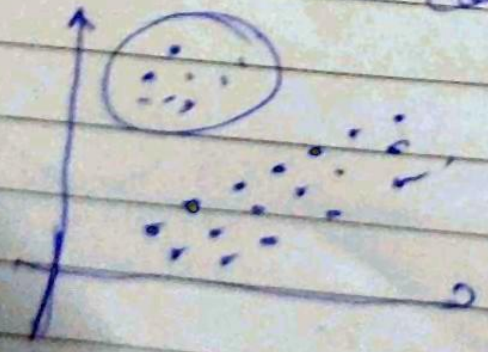


When a single data object deviates from the rest of data.

Point → Global / Point outliers

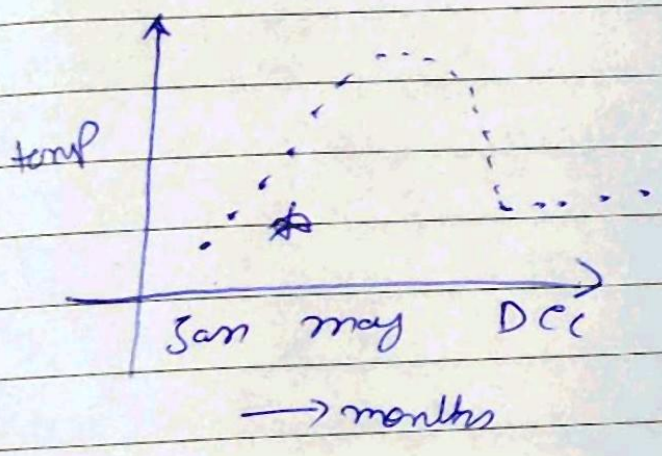
2. Collective outliers:

When a group of data points deviates from the rest of data called collective outliers.



3. Contextual / Conditional outliers:

Data objects deviates from others because of any specific condition - called contextual outliers.



Feature engineering

It is the process of selecting, manipulating, and transforming raw data features that can be used in supervised learning.

It is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and

Speeding up data transformations while also enhancing model accuracy.

Feature creation: Creating features in values
Creating new variables which will be most helpful