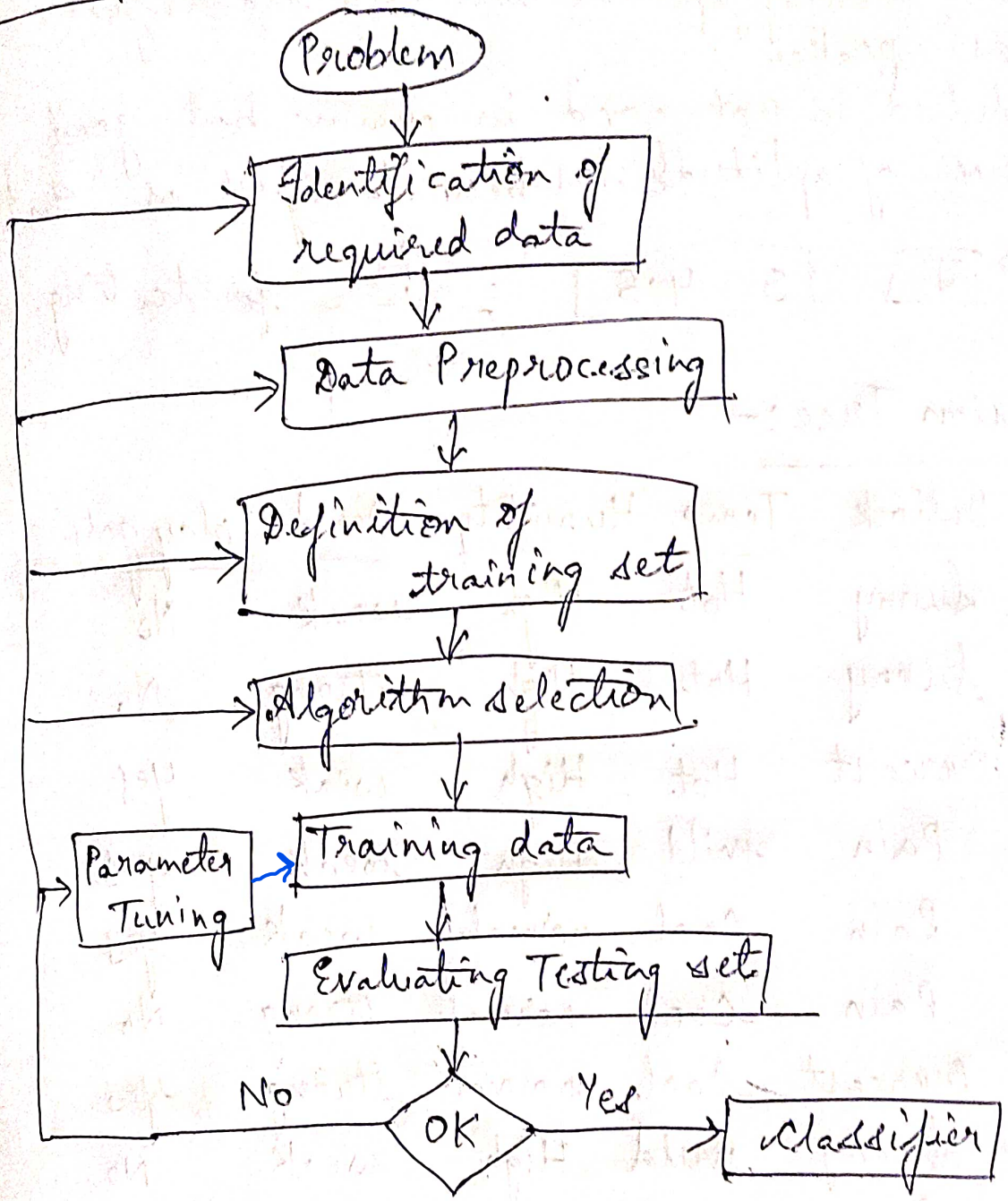
27/4/22.

Unit - 3

Classification.

Classification Learning steps:-



KNN:-

	Name	Aptitude	Communication	Class
Training	A	2	5	Speaker
	B	2	6	Speaker
	C	7	6	Leader
	D	5	3	Intel
	E	6	5.5	Leader
	F	6	4	Intel
	G	9	7	Leader
GT		5	4.5	Intel.

Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

(i) $9/_6$ students having good comm. skills as well as good level of aptitude, have been classified as "Leader".

(ii) Student having good comm. skills but not so good level of aptitude have been classified as "Speaker".

(iii) Student is not good in comm. but good level of aptitude has been classified as "Intel".

→ [I] [5 4.5] ? for testing.

Decision Tree:-

Day	Outlook	Temp	Humidity	Wind	play golf
D_1	Sunny	Hot	High	weak	No
D_2	Sunny	Hot	High	strong	No
D_3	Overcast	Hot	High	weak	yes
D_4	Rain	mild	High	weak	yes
D_5	Rain	cool	normal	weak	yes
D_6	Rain	cool	normal	strong	No
D_7	Overcast	cool	normal	strong	yes
D_8	Sunny	mild	High	weak	No
D_9	Sunny	cool	normal	weak	yes
D_{10}	Rain	mild	normal	weak	yes
D_{11}	Sunny	mild	normal	strong	yes
D_{12}	Overcast	mild	High	strong	yes
D_{13}	Overcast	Hot	normal	weak	yes
D_{14}	Rain	mild	High	strong	No

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$\text{Information Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

yes = 9, No = 5

① $$\text{Entropy}(S) = \frac{-9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= \frac{-9}{14}\left(\frac{\log\left(\frac{9}{14}\right)}{\log 2}\right) - \frac{5}{14}\left(\frac{\log\left(\frac{5}{14}\right)}{\log 2}\right)$$

$$= 0.94$$

Attribute: Outlook (Sunny, Overcast, Rain)

$S_{(\text{Sunny})} = 5$ (yes = 2, No = 3)

$$\text{Entropy} = \left[\frac{-2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$

$$= 0.97$$

$S_{(\text{overcast})} = 4$ (4, 0)

$$\text{Entropy} = \left[\frac{-4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right]$$

$$= 0$$

$S_{(\text{Rain})} = 5$ (yes = 3, No = 2)

$$\text{Entropy} = \left[\frac{-3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 0.97$$

IG (S, outlook)

$= 0.94 - \frac{5}{14}(0.97) - \frac{4}{14}(0) - \frac{5}{14}(0.97)$

$= 0.2464$

Attribute: Temp (Hot, mild, cool)

$S_{(Hot)} = 4$ (yes = 2, No = 2)

Entropy $= \left[\frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$

$= 1$

$S_{(mild)} = 6$ (yes = 4, No = 2)

Entropy $= \left[\frac{-4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right]$

$= 0.9183$

$S_{(cool)} = 4$ (yes = 3, No = 1)

Entropy $= \left[\frac{-3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

$= 0.8113$

IG (S, Temp)

$= 0.94 - \frac{4}{14}(1) - \frac{6}{14}(0.9183) - \frac{4}{14}(0.8113)$

$= 0.94 - 0.2857 - 0.3935 - 0.2318$

$= 0.029$

Attribute: Humidity (High, normal)

$S_{(High)} = 7$ (yes = 3, No = 4)

Entropy $= \left[\frac{-3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$

$=$

$S_{(normal)} = 7$ (yes = 6, No = 1)

Entropy $= \left[\frac{-6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right]$

$=$

IG (S, Humidity)

$= 0.94 - \frac{7}{14}(\quad) - \frac{7}{14}(\quad)$

$=$

Attribute: Wind (Strong, Weak)

$S_{(strong)} = 6$ (yes = 3, No = 3)

Entropy $= \left[\frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$

$= 1$

$S_{(weak)} = 8$ (yes = 6, No = 2)

Entropy $= \left[\frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right]$

$=$

IG (S, Wind)

$= 0.94 - \frac{6}{14}(1) - \frac{8}{14}(\quad)$

$=$

.... cont ....

Q. (2)

Cgpa	Communication	Aptitude	Prog. skill	Job offer
high	good	high	good	yes
medium	good	high	good	yes
low	bad	low	good	no
low	good	low	bad	no
high	good	high	bad	yes
high	good	high	good	yes
medium	bad	low	bad	no
medium	bad	low	good	no
high	bad	high	good	yes
medium	good	high	good	yes

yes = 6, no = 4

$$\text{Entropy}(S) = \frac{-6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= 0.97$$

Attribute : Cgpa (high, medium, low)

$S_{(High)} = 4$ (yes = 4, No = 0)

$$\text{Entropy} = \frac{-4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$$

$$= 0$$

$S_{(medium)} = 4$ (yes = 2, no = 2)

$$\text{Entropy} = \frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 1$$

$S_{(low)} = 2$ (yes = 0, no = 2)

$$\text{Entropy} = \frac{-0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$$

$$IG(S, \text{Cgpa}) = 0.97 - \frac{4}{10}(0) - \frac{4}{10}(1) - \frac{2}{10}(0)$$

$$= 0.57$$

Attribute : Communication (good, bad)

$S_{(good)} = 6$ (yes = 5, no = 1)

$$\text{Entropy} = \frac{-5}{6}\log_2\left(\frac{5}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$$

$$= 0.868$$

$S_{(bad)} = 4$ (yes = 1, no = 3)

$$\text{Entropy} = \frac{-1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$= 0.811$$

$$IG(S, \text{Communication}) = 0.97 - \frac{6}{10}(0.868) - \frac{4}{10}(0.811)$$

$$= 0.1248$$

Attribute : Aptitude (high, low)

$S_{(high)} = 6$ (yes = 6, No = 0)

$$\text{Entropy} = \frac{-6}{6}\log_2\left(\frac{6}{6}\right) - \frac{0}{6}\log_2\left(\frac{0}{6}\right)$$

$$= 0$$

$S_{(low)} = 4$ (yes = 0, No = 4)

$$\text{Entropy} = \frac{-0}{4}\log\left(\frac{0}{4}\right) - \frac{4}{4}\log_2\left(\frac{4}{4}\right)$$

$$= 0$$

IG(S, Aptitude)

$= 0.97 - \frac{6}{10}(\ 0\ ) - \frac{4}{10}(\ 0\ )$

$= 0.97$

Attribute : Prog. skill (good, bad)

$S_{(good)} = 7$ (yes = 5, no = 2)

$\text{Entropy} = \frac{-5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right)$

$= 0.862$

$S_{(bad)} = 3$ (yes = 1, no = 2)

$\text{Entropy} = \frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.918$

IG(S, Prog. skill)

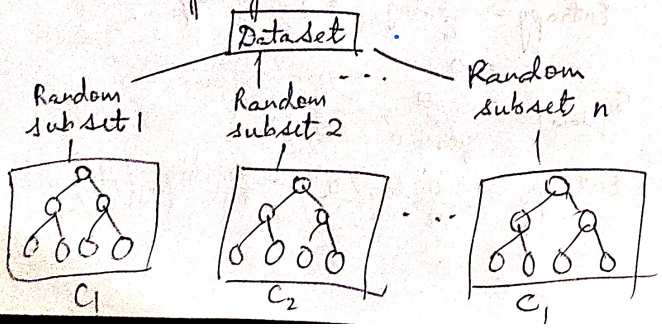
$= 0.97 - \frac{7}{10}(0.862) - \frac{3}{10}(0.918)$

$= 0.0912$

⋮

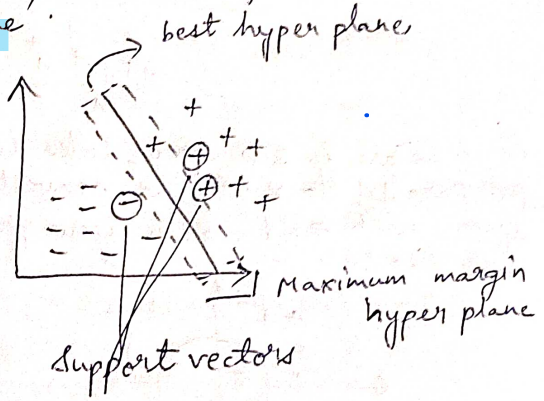
Random Forest :-

OOB – Out of Bag Error Rate.



Support Vector Machine (SVM) :-

* SVM is a eager learner which is used for both classification & regression problems.
* Major applications of SVM are face detection, image classification, text characterisation or hand writing recognition.
* The main goal of SVM is to create best line or decision boundary so that it can segregate n dimensional space into classes so that we can easily put new data point in the correct category feature.
* The best decision boundary is called as 'Hyper plane'.



Types of SVM :-

1) Linear SVM → linearly separable data (having 2 features only like ↑)

* It is used for linearly separable data which means if a dataset can be classified into 2 classes with straight line.

$c_0 + c_1 x_1 + c_2 x_2 = 0$

Random forest :- It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

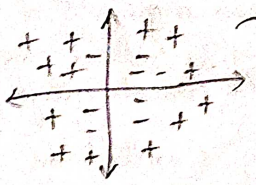
The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

2) Non linear SVM :-

It is used for non-linearly separable data which means if a dataset cannot be classified by using a straight line.

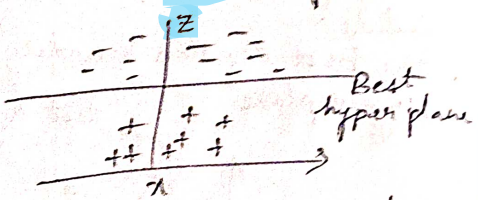
$$c_0 + c_1 x_1 + c_2 x_2 + \dots + c_n x_n = 0$$

$$\vec{c} \cdot \vec{x} + c_0 = 0$$

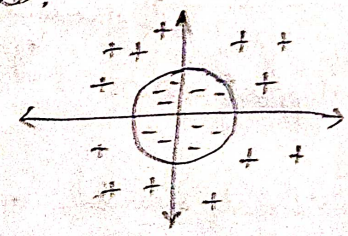


→ In this case we need to add one more dimension z.

$$z = x^2 + y^2$$



Since we are in 3D space, looks like a plane parallel to x. If we convert 3D to 2D space, with z=1, then your graph will look like;



Hyper plane :-

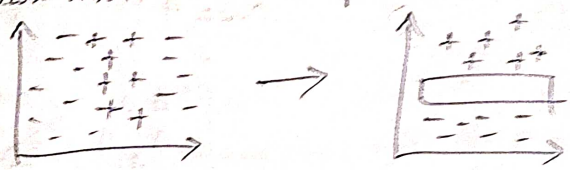
* There can be multiple lines or decision boundaries to segregate the data in n-dimensional space but we need to find the best decision boundary that helps to classify the data points.
* We always create maximum margin which means maximum distance b/w data points.

Support Vectors :-

The data points which are nearest to the margin is hyper plane.

Kernel Trick :-

It deals with non-linear separable data.



* These are the functions which can perform lower dimensional input space to higher dimensional space.
* In this process, linearly non-separable data is converted to linear separable data.
* These functions are called as Kernels.

1) Linear Kernel :

$$K(\vec{x_i}, \vec{x_j}) = \vec{x_i} \cdot \vec{x_j}$$

2) Polynomial Kernel :

$$K(\vec{x_i} \cdot \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d$$

3) Sigmoid Kernel :

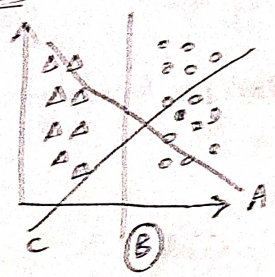
$$K(\vec{x_i} \cdot \vec{x_j}) = \tanh(K\vec{x_i} \cdot \vec{x_j} - \delta)$$

4) Gaussian RBF Kernel :

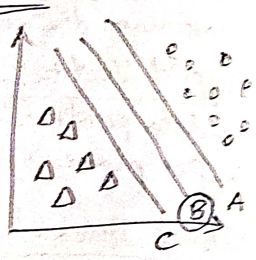
$$K(\vec{x_i} \cdot \vec{x_j}) = e^{\frac{-\vec{x_i} - \vec{x_j}}{2\sigma^2}}$$

Identifying the correct hyper plane in SVM :-

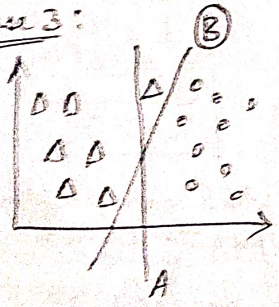
Case 1:



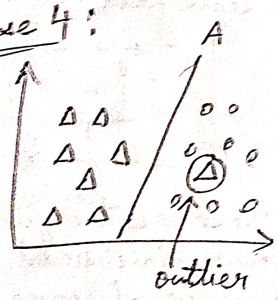
Case 2:



Case 3:



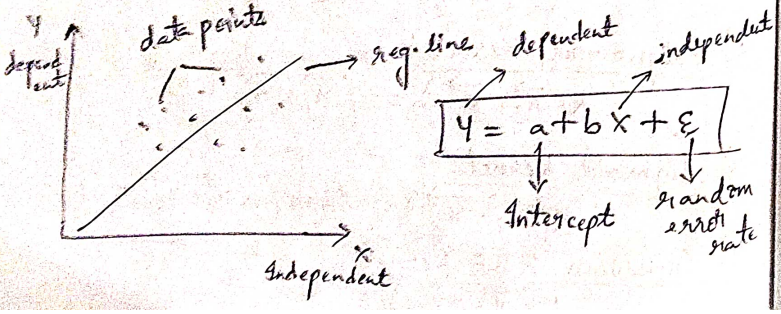
Case 4:



Regression :-

Linear Regression :-

It shows a linear relationship b/w dependent variable (Y) and one or more independent variables.



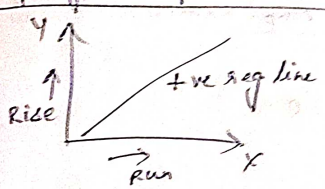
• Simple LR :-

In Simple LR, we have only 1 independent variable which is used to predict the value of numerical dependent variable.

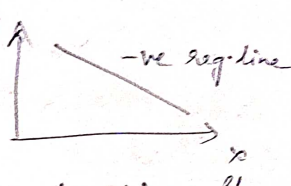
ex: $Y = a + bX$

$Emp_{sal} = a + b\,Emp_{exp}$.

Slope of a simple LR model :-



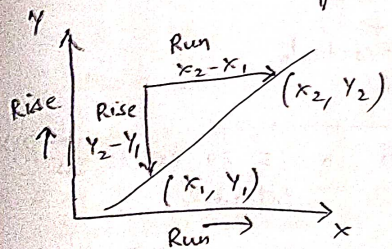
Positive Linear Slope



Negative Linear Slope.

Slope of a straight line represents how much the line in a graph changes in the vertical direction (Y-axis) over the changes in horizontal direction (X-axis).

$$\text{Slope} = \frac{\text{change in } Y}{\text{change in } X} = \frac{Y_2 - Y_1}{X_2 - X_1}$$



Change in Y = Rise

Change in X = Run

$$\text{Slope} = \frac{\text{Rise}}{\text{Run}}$$

Q: Find the slope of a graph where the lower point on the line represents $(-3, -2)$ and the higher point of line is $(2, 2)$.

Soln:

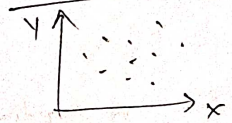
$$\text{Slope} = \frac{2-(-2)}{2-(-3)} = \frac{4}{5} = 0.8$$

* There are 2 types of slopes in LR model.

(i) Positive slope ⇓ $x \uparrow y \uparrow$ (ii) Negative slope ⇓ $x \uparrow y \downarrow$

$y = a_0 + a_1 x$

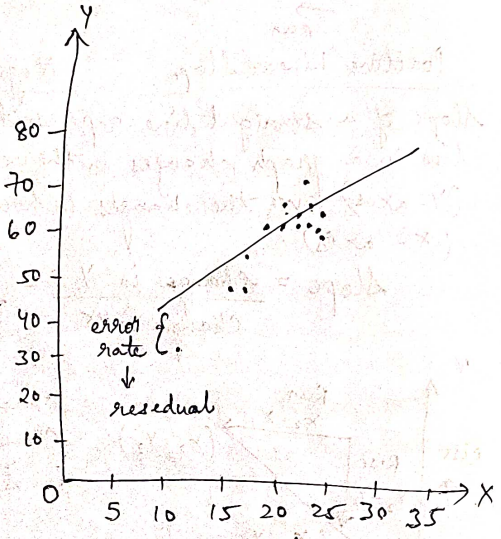
No relationship graph:-



ex: Finding emp. salary by emp. name.

Finding error in Simple LR:-

Internal (x)	External (y)
15	49
23	63
18	58
23	60
24	58
22	61
22	60
19	63
19	60
16	52
24	62
11	30
24	59
16	49
23	68
299	852



mean $\bar{y}(Y) = 56.8$; mean $\bar{x}(X) = 19.93$

SSE = Sum of Squares / Sum of products

$y = a + bx$

$$b = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

a=mean(y)-b(mean(x))

X - mean(x)	Y - mean(y)	$(x-\bar{x})(y-\bar{y})$	$(x_i-\bar{x})^2$
-4.93	-7.8	38.454	24.3049
3.07	6.2	19.034	9.4249
-1.93	-1.2	-2.316	3.7249
3.07	3.2	9.824	9.4249
4.07	1.2	4.884	16.5649
2.07	-4.2	8.694	4.2849
2.07	3.2	6.624	4.2849
-0.93	6.2	-5.766	0.8649
-0.93	3.2	-2.976	0.8649
-3.93	-4.8	18.864	15.4449
4.07	5.2	21.164	16.5649
-8.93	-26.8	239.324	79.7449
4.07	2.2	8.954	16.5649
3.93	-7.8	30.654	15.4449
3.07	11.2	34.384	9.4249
		429.8	226.9335

$$b = \frac{429.8}{226.9335} = 1.89$$

$a = \bar{y} - b\bar{x} = 56.8 - (1.89)19.93 = 19.084$

$y = a + bx$

$y = 19.04 + 1.89x$

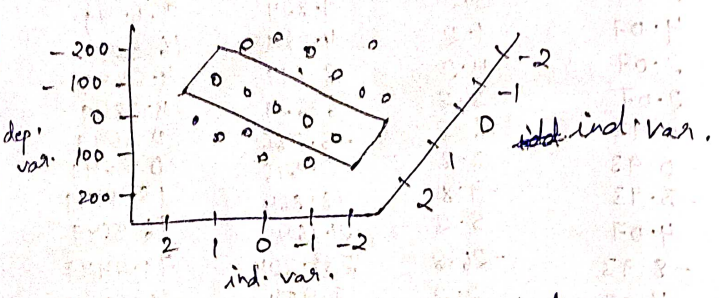
∴ Marks of external = 19.04 + 1.89 marks of Internal.

Multiple LR:-

* It is also like LR but with more than one independent variable i.e it gives linear relationship b/w 1 dep. variable & more ind. variables.
* Here the dep. variable must be continuous or real but ind. variables may be continuous or categorical.
* Multiple LR tries to fit a regression line through a multi-dimensional space of datapoints.

* y is dep. variable; $b_0, b_1, b_2 \ldots$ are coefficients of model; $x_1, x_2, \ldots$ are feature variables.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$



Assumptions in Linear Regression Analysis :-

* Regression is a parametric approach i.e it makes assumptions about data for the purpose of analysis.
* It fails to deliver good results with given dataset which doesn't fulfill assumptions.

Assumptions :-

(i) There should be linear and additive relationship b/w dep. & ind. variable.

Linear & additive:

If you fit a linear model to non-linear, non-additive dataset, the regression algorithm fails to capture the performance of a model.

(ii) There should be no correlation b/w residual terms. So the absence of this phenomenon is called Auto-correlation.

Auto-correlation:

This usually occurs in the time-series models where the next instance depends on previous instance. The presence of correlation in error terms drastically reduces the model accuracy.

(iii) The ind. variable should not be correlated. The absence of this phenomenon is called as Multi-collinearity.

Multicollinearity:

This phenomenon exists when the ind. variables are found to be highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of predictors with response variable.

(iv) The error terms must have constant variance. This is called as Homoskedasticity. The presence of non-constant variances is called as Heteroskedasticity.

Heteroskedasticity:

Non-constant variance arises in presence of outliers or extreme leverage values. These values have high weight that leads to reduce the performance of a model.

(v) The error term must be normally distributed.

Logistic Regression :-

* It is a classification model used to predict categorical dependent variable using a given set of ind. variables.
* The categorical dep. variable can be 0 or 1, yes or no, true or false. This is also called as Binomial Logistic Regression.
* In log. reg. instead of fitting line we fit S-shaped logistic function which predicts 2 maximum values.
* We need this Sigmoid function to get probability of a given variable x.

$$f(x) = \frac{1}{1+e^{-x}}$$

(∴ Sigmoid fn. ranges 0 to 1)

* This log. reg. uses the concept of predictive modelling of regression.

Assumptions in Log. Reg. :-

(i) The dep. variable must be categorical.
(ii) The ind. variable should not have multicollinearity.

Types of Log. Reg. :-

(i) Binomial Log. Reg.
(ii) Multinomial log. reg.
(iii) Ordinal log. reg.

If line, $y = mx + c$

- $P(y) = \frac{1}{1+e^{-y}}$ (or) $\frac{1}{1+e^{-(mx+c)}}$
- $P(x) = \frac{1}{1+e^{-(mx+c)}} = \frac{1}{1+e^{-y}} = \frac{1}{1+\frac{1}{e^y}}$

$$\Rightarrow P(x) = \frac{e^y}{1+e^y}$$

- $y = h(\theta)$

$$P(x) = \frac{e^{h(\theta)}}{1+e^{h(\theta)}}$$

$$1 - P(x) = 1 - \frac{e^{h(\theta)}}{1+e^{h(\theta)}}$$

$$\Rightarrow \frac{P(x)}{1-P(x)} = \frac{e^{h(\theta)} / 1+e^{h(\theta)}}{1 - \frac{e^{h(\theta)}}{1+e^{h(\theta)}}}$$

$$= e^{h(\theta)}$$

$$= e^{mx+c}$$

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \log\left(e^{mx+c}\right)$$

$$\therefore \boxed{\log\left(\frac{P(x)}{1-P(x)}\right) = mx+c}$$