1 What is Machine Learning?

- Machine Learning is a subset of artificial intelligence.
- It is focuses mainly on the designing of a systems, thereby allowing them to learn and make predictions based on some experience which is data in case of Machine.
- How does Machine learning works?

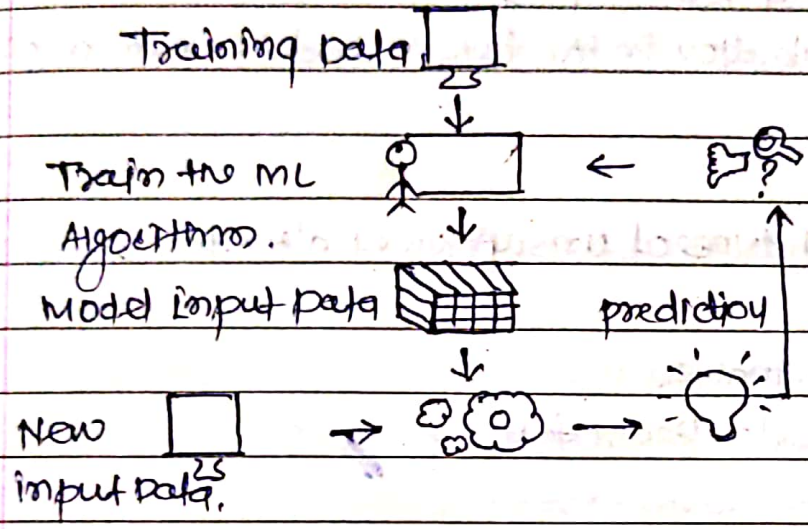


Diagram: How Machine learning works.

- There are different types of Machine learning Algorithm, Above are the types of Machine learning.
  (i) supervised learning.
  (ii) Unsupervised learning
  (iii) Reinforcement learning.

① Supervised learning:

- supervised learning Algorithm is input variable x and output variable y, and use an algorithm to learn mapping function from input to output, that is $y = fx$.
- Goal is to approximately mapping function so well that whenever new data (input x) good output predicted variable.

(i) classification
(ii) Regression.

(ii) Unsupervised learning Algorithm:
- Unsupervised learning algorithm only have input data & No corresponding output.
- Goal of unsupervised learning algorithm is model underlying structure or distribution in the data in order to learn about the data.

- There are different types of unsupervised algorithm,
  (i) clustering.
  (ii). Association Analysis
  (iii) Dimensionality Reduction.

(iii) Reinforcement Algorithm:
- Reinforcement algorithm software can machine automatically determine ideal behaviour, with in specific context to maximize its performance.

- There are different type of Reinforcement Algorithm.
  i) Model free - Q-learning
  ii) Model Based.

(1) Linear Regression:

- Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has constant slope.
- It is used to predict values within a continuous range, (e.g. sales, price), rather than trying to classify them into categories.
- There are two main types
  - (i) simple linear Regression.
  - (ii) Multiple Linear Regression.

i) simple Regression:

- simple linear Regression is used to estimate the relationship between two quantitative variables.
- use simple linear Regression when you want to know.
  - (i) How strong the relationship between two variables.
  - (ii) The value of the dependent variable at a certain value of independent variable.
- If you have more than one independent variable, use multiple linear Regression.

Assumption of simple Linear Regression:

- simple Linear Regression is a parametric test, meaning that it makes a certain assumption about the data.
- Assumption are,
  - (i) Homogenity of variance (Homoscedasticity):
    - size of Error in our prediction doesn't change significantly across the values of the independent variable.
  - (ii) Independance of observation:
    - observations in the dataset were collected using statistically valid sampling methods.
    - No hidden relationship among observations.

(iii) **Normality**:

- Data follows Normal Distribution.

- Linear Regression makes one ~~more~~ additional assumption.

(iv) **Linearity**

- Relationship between the independent and dependent variable is linear.
- the line of best fit through the data points is a straight line (rather than curve or some sort of grouping factor)

- If data do not meet the Assumption of Homoscedasticity or Normality, you may able to use a non-parametric test instead.

- If your data violate the assumption of independence of observation (Ex. observation are repeated over time) you may able to perform a linear mixed-effects model that accounts addition structured data.

Perform a simple Linear Regression:

- the formula for a simple linear Regression is,

$$Y = B_0 + B_1 X + \varepsilon$$

Y - predicted value of depedent variable (Y) for given value of independent variable (X).

B_0 - intercept, predicted value of Y when X is 0.

B_1 - Regression coefficient - How much we expect Y to change as X increases.

X - Independent variable (variable we expect influencing Y)

e - e is the error of the estimate, How much variation there is in our estimate of the regression coefficient.

- Linear Regression finds line of best fit line through data by searching regression coefficient (B_i) that minimizes the total error (ϵ) of model.

① Error calculated in a Linear Regression model?

- Linear Regression most often uses mean-square-error (MSE) to calculate the error of the model. MSE is calculated by.
  (i) Measuring the Distance of the observed y values from predicted y-values at each value of x.
  (ii) squaring each of these distances.
  (iii) calculating mean of each of squared distances.

- Linear Regression fits a line to the data by finding the regression coefficient that results in smallest MSE.

② How to find mean?

- Mean or Arithmetic mean of a dataset is the sum of all values devided by total number of values.
- Also called as "Average".

③ What is Regression Model?

- Regression model is a statistical model that estimates the relationship between one dependent variable and one or more independent variable ~~is quantitative~~ using line. (or a plane in the case of two or more independent variables).

- A Regression model can be used when the dependent variable is a qualitative.

(ii) Multiple Linear Regression:

- Regression models are used to describe relationship b variables by fitting a line to the observed data.
- Regression allows you to estimate how a dependent va changes as independent variable changes.

- Multiple linear regression is used to estimate the relation between two or more independent variables. and one depend variables.
- Use multiple linear regression when you want to know,
  (i) How strong the relationship between two or more indep variable and one dependent variable:
  (Ex. How rainfall, temprature, and amount of fertilizer affect of crop growth).
  (ii) The value of dependent variable at a certain value of t independent variables.
  (Ex. Expected yield of a crop at certain levels of rainfall, temprature and fertilizers addition).

Assumption of multiple linear Regression:

- Multiple Linear Regression make all assumption as simple li Regression.
  (i) Homogeneity of variance (Homoscedasticity):
  - The size of the error in our prediction doesn't change significantly across the values of independent variable
  (ii) Independence of observation:
  - observation in the datasets were collected using stati -cally valid method
  - There is NO hidden relationship among ~~them~~ variable

- In multiple linear Regression, it is possible that some of independent variable highly correlated with another, so it is important to check these before developing Regression model.
- If two variables are too highly correlated ($R2 > \sim 0.6$), then only one of them should be used in Regression model.

(iii) Normality:
- The data follows a Normal Distribution

(iv) Linearity:
- The line of best fit through the data points is a straight line rather than a curve or some sort of grouping factor.

Perform a multiple linear Regression
- Formula for multiple linear Regression is,

$$Y = B_0 + B_1X_1 + B_2X_2 - \cdots + B_nX_n + \epsilon$$

Y = predicted value of dependent variable.
B_0 = Y-intercept (value of Y when all other parameters are set to 0).
B_1X_1 = Regression coefficient (B_1) of the first independent variable (X_1) [effect that increasing value of independent variable has on the predicted Y value]
..... = same for however many independent variables you are testing.
B_nX_n = Regression coefficient of the last independent variable.
ϵ = model error [how much variation there is our estimate of Y.

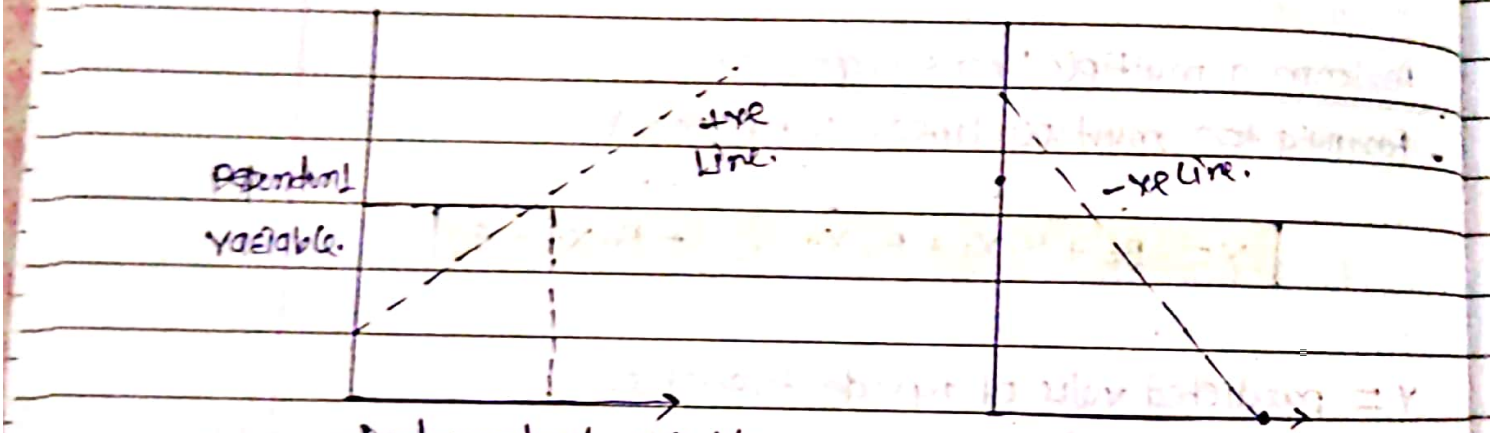
- To find best-fit line for each independent variable, multiple linear Regression calculates three things,

(i) Regression coefficient lead to the smallest overall model.
(ii) t-statistics of overall model.
(iii) Associated p-value (how likely it is that t-statistics would have occurred by chance if the Null hypothesis of no relationship between independent dependent variable was true).

- It then calculates t-statistics and p-value for each regression coefficient in the model.

Algorithm:

- suppose you have ~~dataset~~ independent variable on the x-axis, dependent variable on y axis.

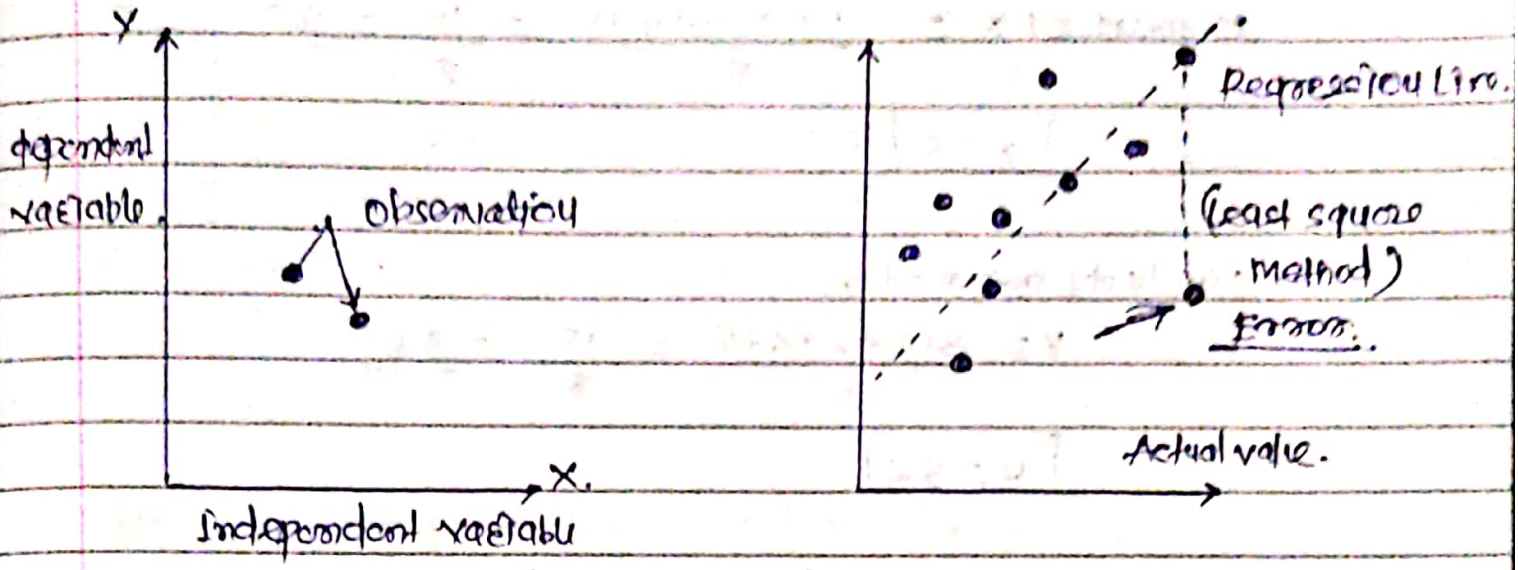


- Independent variable on the x-axis (increasing) and does the dependent variable on the y axis you would get positive linear regression.
- Independent variable on the x axis (increasing) and other hand the dependent variable on the y axis (decreasing) that is Negative Independent as a slope of the line is Negative and the particular line that is

$$y = mx + c$$

of linear expression, which shows the relationship between independent variable and dependent variable and this only known as line of linear regression.

- some data points on to our graph.

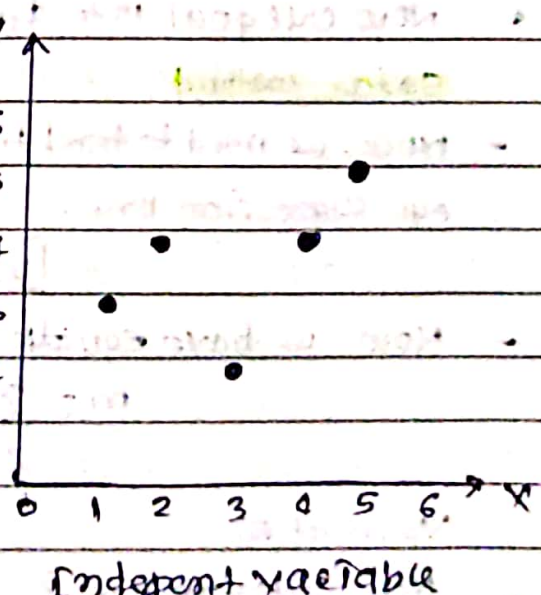


- All are data points are plotted Now our task to make Regression line. of the best fit line.
- Now we get estimated, predicted and Actual value.
- Our main goal is to reduce this error that is distance between estimated value and Actual value. or predicted value.

Mathematical Implementation:

- consider Datapoints $x = \{1, 2, 3, 4, 5\}$ on the x axis and $y = \{3, 4, 2, 4, 5\}$ on the y axis.

x	y
1	3
2	4
3	2
4	4
5	5



- Now calculate mean of x and y and plot it on graph,

so mean of $x = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

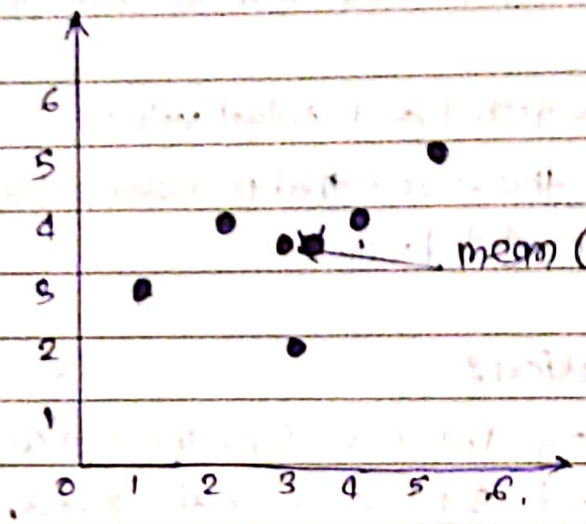
$$\boxed{x = 3}$$

- similarly mean of y.

$$y = \frac{3+4+2+4+5}{5} = \frac{18}{5} = 3.6$$

$$\boxed{y = 3.6}$$

- Now plot a graph using mean value,



- Now our goal is to find or predict best fit line using Least square method.
- Now, we need to first find equation of line, so lets find the equation Regression line.

$$\boxed{y = mx + c}$$

- Now, we have equation of line, so we need to find the,

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

value of m.

- Now calculate $x - \bar{x}$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$
1	3	$(1-3) = -2$	$(3-3.6) = -0.6$	4
2	4	$(2-3) = -1$	$(4-3.6) = 0.44$	1
3	2	$(3-3) = 0$	$(2-3.6) = -1.6$	0
4	4	$(4-3) = 1$	$(4-3.6) = 0.4$	1
5	5	$(5-3) = 2$	$(5-3.6) = 1.04$	4
mean	3	3.6	$\Sigma = 4$	$\Sigma = 10$

- Now we need product of $(x-\bar{x})$ and $(y-\bar{y})$

$$(x-\bar{x}) \cdot (y-\bar{y})$$

$(-2) \cdot (-0.6) = 1.2$
$(-1) \cdot (0.4) = -0.4$
$(0) \cdot (-1.6) = 0$
$(1) \cdot (0.4) = 0.4$
$(2) \cdot (1.4) = 2.8$

- Now calculate summation of last two columns.

$$m = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})} = \frac{4}{10} = 0.4.$$

- Now, we have $\bar{x}$ as 3.6 mean of $m = 0.4$.

Now $y = mx + c$
$\bar{y}$ 3.6

$3.6 = 0.4 * 3 + c$
$3.6 = 1.2 + c$
$3.6 - 1.2$
$c = 2.4$.

Now, $m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$
so this is the Regression line.

Now, $X = \{1, 2, 3, 4, 5\}$.

$m = 0.4$

$c = 2.4$.

Now, calculate predict value for Y. $\boxed{Y = mx + c}$

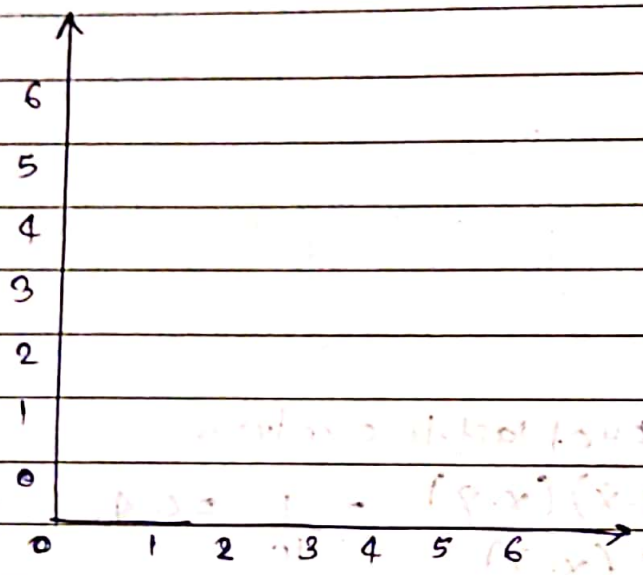
$Y = 0.4 * 1 + 2.4 = 2.8$

$Y = 0.4 * 2 + 2.4 = 3.2$

$Y = 0.4 * 3 + 2.4 = 3.6$

$Y = 0.4 * 4 + 2.4 = 4.0$

$Y = 0.4 * 5 + 2.4 = 4.4$.



- Now calculate distance between Actual value and predicted value and Reduce Distance. (Least square Method).
- So it will perform N number of iteration for different value of it will calculate equation of line.

$$\boxed{Y = mx + c}$$

- So value of m changes the line is changing. If iteration from one and it will perform n number of iteration, after each iteration it will calculate the predicted value.

- According to line and compare distance of Actual value to the predicted value. and value of M for which the distance between Actual and predicted value is minimum will be selected as best fit line.

- Now we have calculated the best fit line, now its time to check goodness of fit or check how good our model performing.
- In ordered to do that we have a method called R-squared method.
[R squared method use for check goodness of fit in Regression line]

What is R squared ?

- R- squared value is a statistical measure of how close data are to the fitted regression line.
- It also known as coefficient of determination or coefficient of multiple determination.

- In general it is considered as that High-square value model is a good model but you can also have a lower squared value for a good model as well as higher squared value for a model that does not fit at all.

calculation of R^2 :

- Distance = Actual - mean vs Distance = predicted - mean.
this is nothing but R^2.

$$R^2 = \frac{\Sigma (Y_p - \bar{Y})^2}{\Sigma (Y - \bar{Y})^2}$$

Where,

Y is Actual value.

Y_p is predicted value.

$\bar{Y}$ is mean value of Y (3.6).

X	Y	Y-Ȳ	(Y-Ȳ)²	Yp	(Yp-Ȳ)	(Yp-Ȳ)²
1	3	-0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	-1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
	3.6		Σ(5.2)		Σ(1.6)	Σ 1.6

$$R^2 = \frac{\Sigma(Yp-\bar{Y})^2}{\Sigma(Y-\bar{Y})^2} = \frac{\Sigma 1.6}{\Sigma 5.2} = 0.3$$

- this is not a good fit. It suggest that the data points are far away from Regression line.
- When increase value of R^2 to 0.7 the actual value would likely closer the Regression line, $R^2 \approx 0.7$
- When it comes to $R^2 \approx 0.9$ it comes to more close, and value approximately equals to 1, then actual values lies on Regression line itself.

For Example:

- In this case if you get very low value of R squared suppose $R^2 = 2$ so in that case, actual value are far away from Regression line
- there are outliers present in the data.

RSS (Residuals).

- Residual for a point in the data is the difference between the actual value and the value predicted by our linear regression model.

$$\text{Residual} = \text{actual} - \text{predicted} = y - \hat{y}$$

- Residual plot tell us whether regression model is right fit or not (data).
- It is Actually an assumption of ~~data~~ Regression model that there is no trend in Residual plot.

- using Residual value, we can determine sum of square of Residuals, also known as Residual sum of square or RSS.

$$RSS = \sum (y_i - \hat{y}_i)^2$$

- lower value of RSS, the better is model prediction. OR
- Regression line is a line of best fit it minimizes the RSS. value.
- The value depends on a scal of the target variable.

Total sum of squares:

- TSS or total sum of squares gives the total variation in Y.
- similer to variance of Y.
- variance is a average of sum of difference between actual values and data points.

calculate R-squared:

$$\text{R-squared} = \frac{(TSS - RSS)}{TSS}$$

$$= \frac{\text{Explained variation}}{\text{total variation}}$$

$$= 1 - \frac{\text{Unexplained variation}}{\text{total variation}}$$

Adjusted R-squared :

- Adjusted R-squared takes into account the number of ind variable used for predicting the target variable.
- so, determine whether adding new variable to the model actu increases the model fit.

$$\text{Adjusted } R^2 = \left\{1 - \left[\frac{(1-R^2)(n-1)}{(n-k-1)}\right]\right\}$$

Where,

$n \rightarrow$ number of data points in our Dataset.

$k \rightarrow$ number of Independent variable.

$R \rightarrow$ R-squared value determined by model.

- IF R-squared does not increase significantly on the add new independent variable, then value of Adjusted R-squa will Actually Decrease.

$$\text{Adjusted } R^2 = \left\{1 - \left[\frac{(1-R^2)(n-1)}{(n-\uparrow k-1)\downarrow}\right]\uparrow\right\}\downarrow$$

- If on adding new independent variable we see signifi increase R-squared value, then Adjusted R-squared value wi increased.

$$\text{Adjusted } R^2 = \left\{1 - \left[\frac{\downarrow(1-\uparrow R^2)(n-1)}{(n-k\uparrow-1)\downarrow}\right]\uparrow\right\}\uparrow$$