Nov-11:

Machine Learning Workflow:

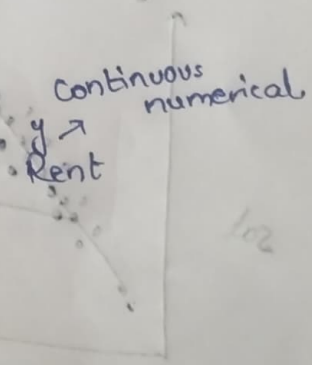
1. Problem Statement.
2. Data Collection (filebase, database)
3. Data cleaning / Transformation / Preprocessing.
   - Data types
   - Null values
   - Duplicated values.
   - Column names
   - ...

   → Feature engineering.
   * adding a column
   * updating a column
   * deleting a column
   * converting categorical to numerical
4. Feature Engineering.
5. Model Validation (only for supervised learning)
6. Model Selection
7. Training the Model
8. Model Evaluation.
9. Deployment & Monitoring

Machine Learning Algorithms:

1. Linear Regression:

When our prediction point / value is continuous numerical in that case we have to use regression algorithms.



Independent variables : x (features) Input

Dependent variables: y (target) output.

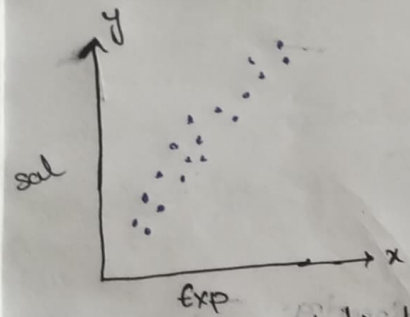
* Note: Always y should be dependent on x & x shouldn't depend on each other.

*If we want to use simple linear regression, the data has to pass some assumptions.

1. Always x & y should be colinear to each other. It may be +ve collinear or -ve collinear but not non-colinear.

2. The data contains only one feature & one target.

Exp	Sal
2	30,000
4	75,000
3.5	60,000
1	25,000



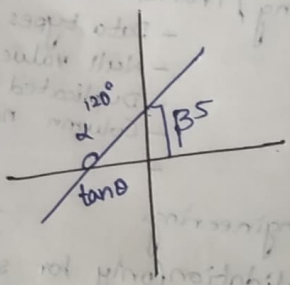
What is math behind simple linear regression? (SLR)

It is linear line equation

$$y = mx + c$$

where, m = slope

c = constant



SLR equation is

$$\hat{y} = \alpha x + \beta$$

where, α - coefficient

β - Intercept

α - $\tan\theta$

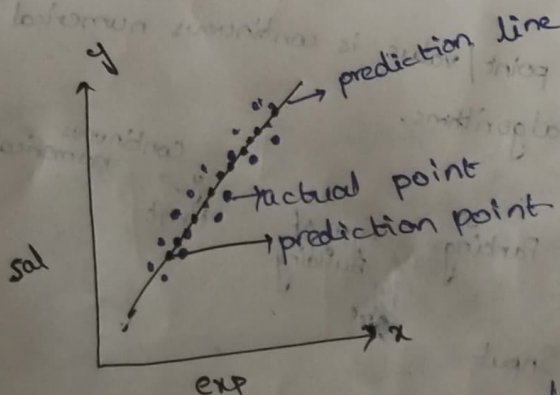
Ex:

$\hat{y} = \alpha x + \beta$

$\hat{y} = \tan 120^\circ x + 5$

$\hat{y} = -\sqrt{3}x + 5$

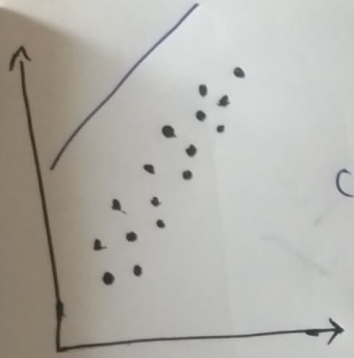
$\hat{y} = -1.73x + 5$

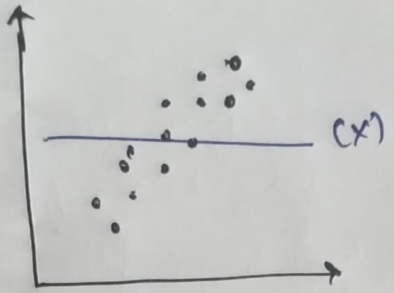


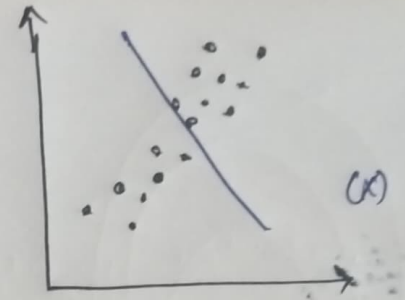
blue dots - actual points

black dots - prediction points

* The prediction line always should be near to the actual points.







To find best fit line

- The average distance b/w actual point & predicted line should be minimum then it is called as Best fit line.

- Always predicted line is dependant on α & β values. So α & β are called as hyper parameters.

hyper parameters - non constant values unlike π. which is 3.14 all the time.