UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

Course Name : AI & ML

Types of Supervised Learning

Regression:- It is a Supervised Learning task where output is having continuous value.

Classification:- It is a Supervised Learning task where output is having defined labels(discrete value).

Supervised Learning Algorithms:

- Linear Regression
- Nearest Neighbor
- Naive Bayes
- Decision Trees
- Support Vector Machine (SVM)

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

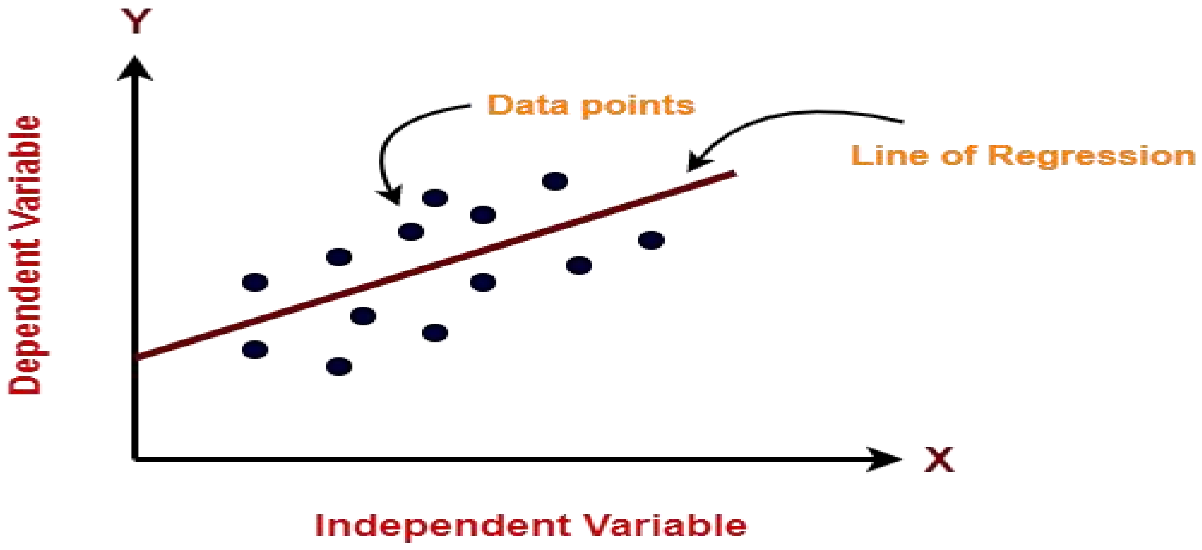
Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

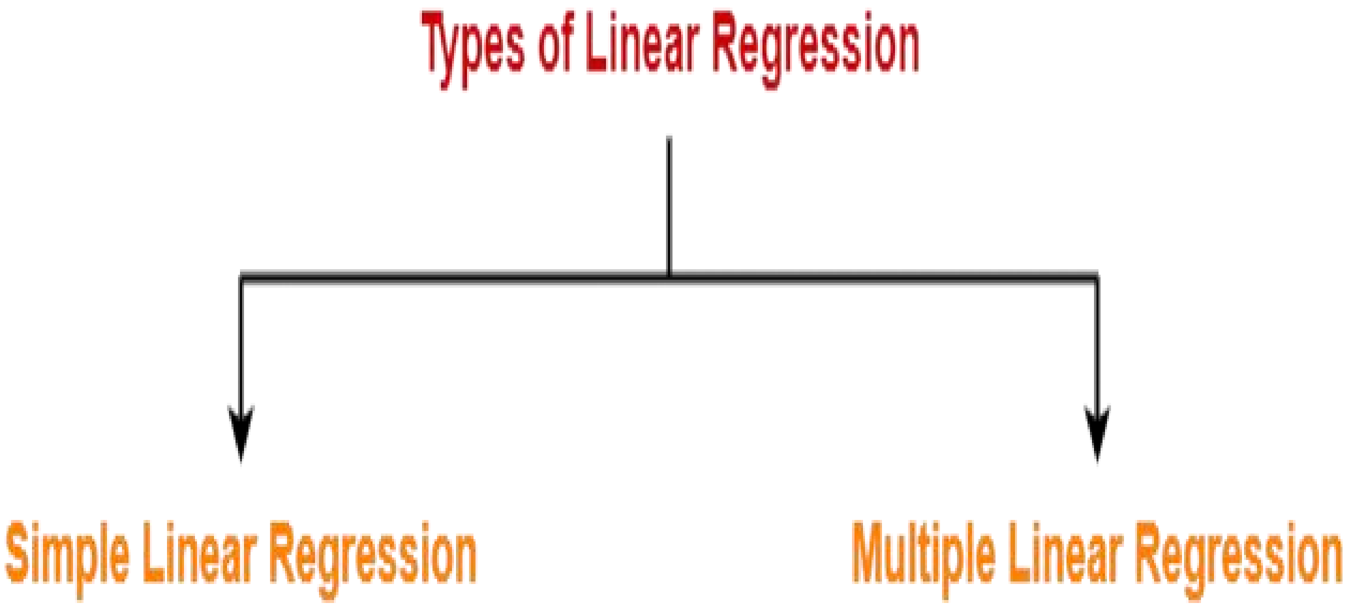
Figure B: REGRESSION

Unsupervised Learning algorithms:

- K-means clustering
- Hierarchal clustering
- Anomaly detection
- Apriori algorithm

Linear regression model represents the linear relationship between a dependent variable and independent variable(s) via a sloped straight line.




Types of Linear Regression
Simple Linear Regression
Multiple Linear Regression

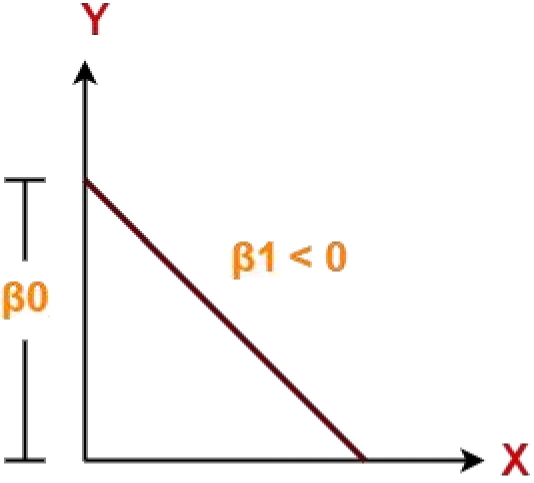
- **Simple Linear Regression-**
- In simple linear regression, the dependent variable depends only on a single independent variable.
-
- For simple linear regression, the form of the model is-
- $\mathbf{Y = \beta_0 + \beta_1 X}$

- Here,
- Y is a dependent variable.
- X is an independent variable.
- β_0 and β_1 are the regression coefficients.
- β_0 is the intercept or the bias that fixes the offset to a line.
- β_1 is the slope or weight that specifies the factor by which X has an impact on Y.

Case-01: $\beta_1 < 0$

It indicates that variable X has negative impact on Y.

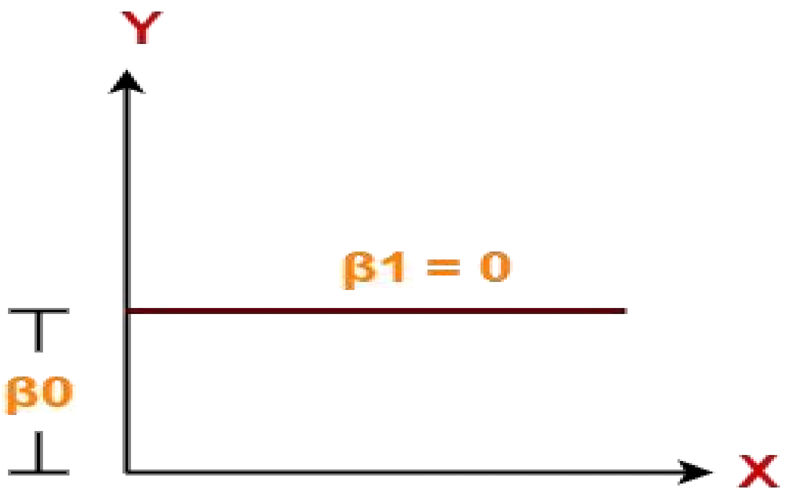
If X increases, Y will decrease and vice-versa.



Case-02: $\beta_1 = 0$

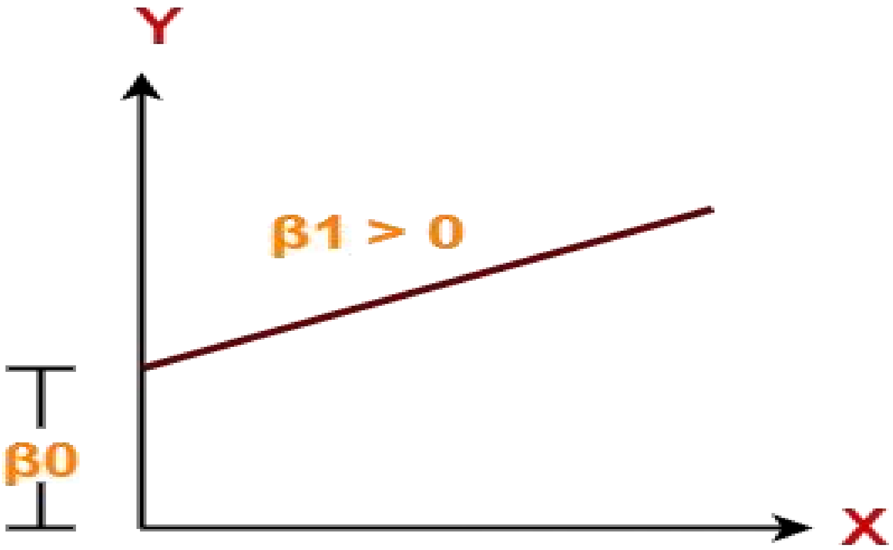
It indicates that variable X has no impact on Y.

If X changes, there will be no change in Y.



Case-03: $\beta_1 > 0$

It indicates that variable X has positive impact on Y.
If X increases, Y will increase and vice-versa.



Consider the following data and construct a function for linear regression.

Internal exam marks: 15,23,18,23,24,22,22,19,19,16,24,11,24,16,23

External exam marks: 49,63,58,60,58,61,60,63,60,52,62,30,59,49,68

Solution

- Consider internal marks as x. Then calculate mean $\bar{x}$. Consider external marks as y. Then calculate $\bar{y}$.
- Need to develop an equation as :

$$\bar{y} = a + b\bar{x}$$

- Where, $b = \frac{Cov(x,y)}{Var(x)}$
- Then calculate $a = \bar{y} - b\bar{x}$
- Equation will be $\bar{y} = 19.05 + 1.89\bar{x}$

- $Y = 4.17 + 0.166\ x$
- X: 3, 9, 5, 3
- Y: 8, 6, 4, 2
- Find error (MSE)

Metrics for model evaluation in regression:-

3 popular methods are :

1. R-squared
2. Mean Square Error (MSE)/ Root Mean Square Error (RMSE)
3. Mean Absolute Error(MAE)
4. Sum of Squares (SSE)

R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Mean Square Error is an absolute measure of the goodness for the fit. MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points.
Root Mean Square Error(RMSE) is the square root of MSE.

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

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

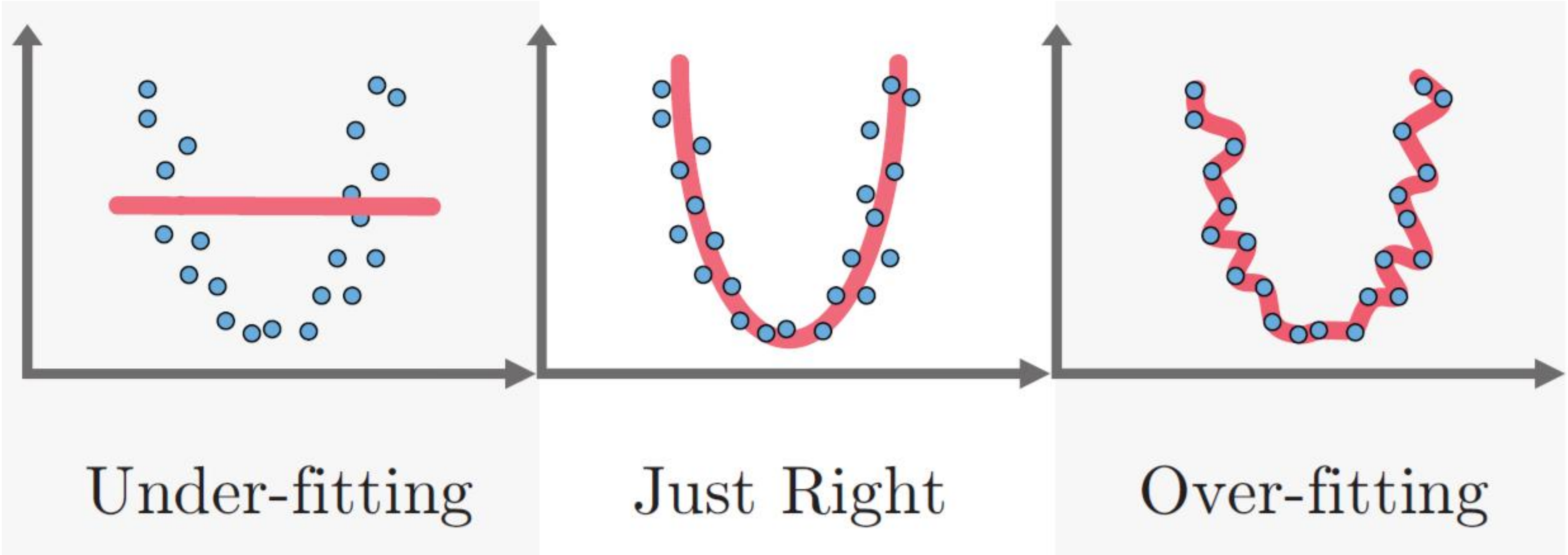
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0.
- The formula for SSE is:-
- SSE = $\sum_{i=1}^{n}(x_i - \bar{x})^2$

Under-fitting, perfectly fitting and Over-fitting Issue



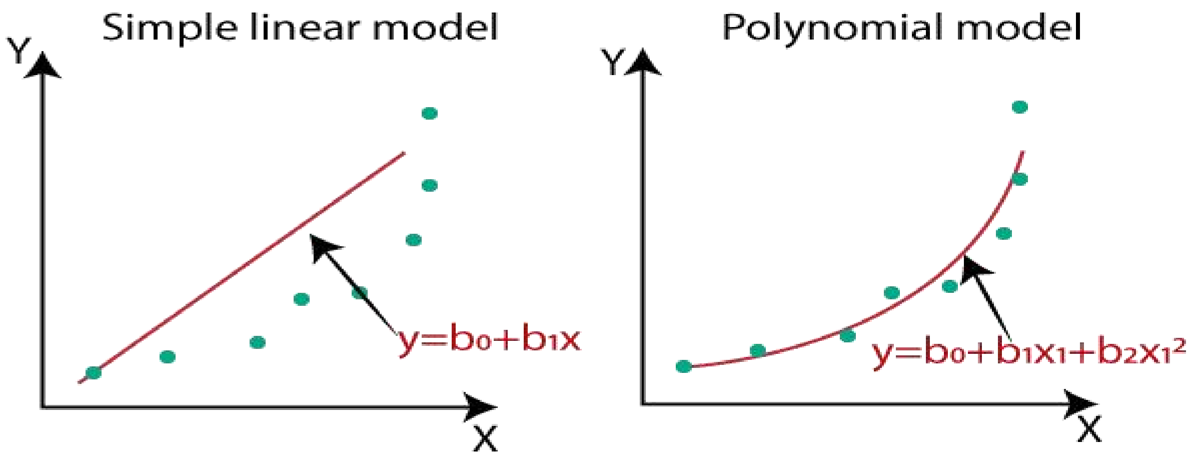
The under-fitted model can be easily seen as it gives very high errors on both training and testing data.

The over-fitted model can perform properly with the present training data. But performance will be poor with a new set of test data.

- **Multiple Linear Regression-**
- In multiple linear regression, the dependent variable depends on more than one independent variables.
- For multiple linear regression, the form of the model is-
- $\mathbf{Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \ldots\ldots + \beta_nX_n}$
- Here,
- Y is a dependent variable.
- $X_1, X_2, \ldots., X_n$ are independent variables.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients.
- β_j ($1<=j<=n$) is the slope or weight that specifies the factor by which X_j has an impact on Y.

Polynomial Regression

If data points are arranged in a non-linear fashion, we need the Polynomial Regression model. Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.



Thank You

Information Sources: NPTEL, Coursera, Analytics Vidya, geeksforgeeks.