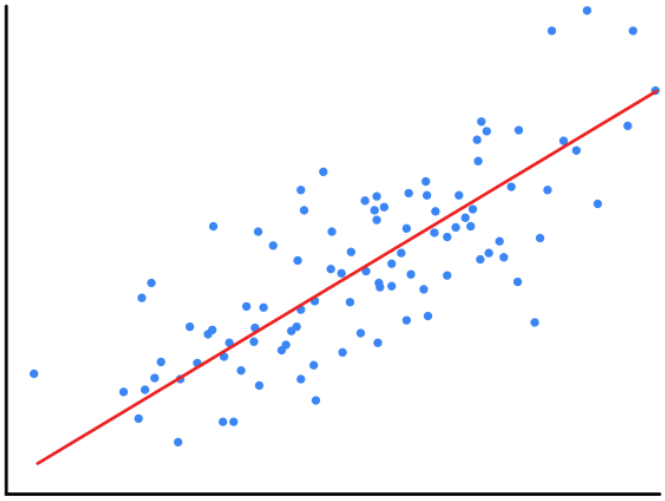
Regression

Machine Learning is an application of artificial intelligence that provides systems the ability to improve from experience. Machine learning algorithms are often categorized as supervised or unsupervised. In supervised learning, machine is trained using data which is well labeled. It means some data are already tagged with correct answers.

Regression and classification are two prominent approaches of learning. In regression, the output variable takes continuous values whereas in classification, the output variable takes class labels (categories / limited discrete values). In regression analysis, curve fitting is the process of specifying model that provides the best fit to the specific curve in dataset. There are many regression techniques such as linear regression, polynomial regression, support vector regression etc.

Linear regression



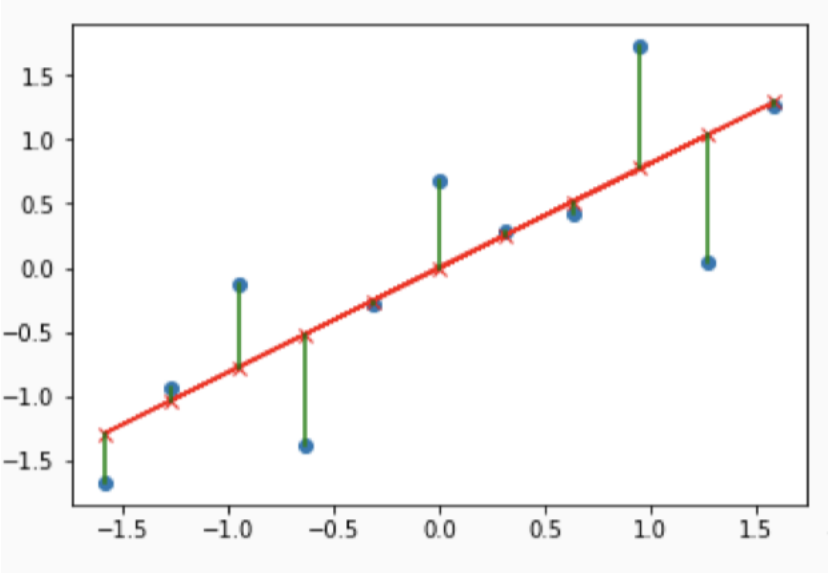
Linear regression is the simplest form of regression where data are modeled using a straight line.

Any straight line can be represented by an equation of the form Y = bX + a, where b and a are constants. The value of b is called the slope constant and determines the direction and degree to which the line is tilted. The value of a is called the Y-intercept and determines the point where the line crosses the Y-axis.

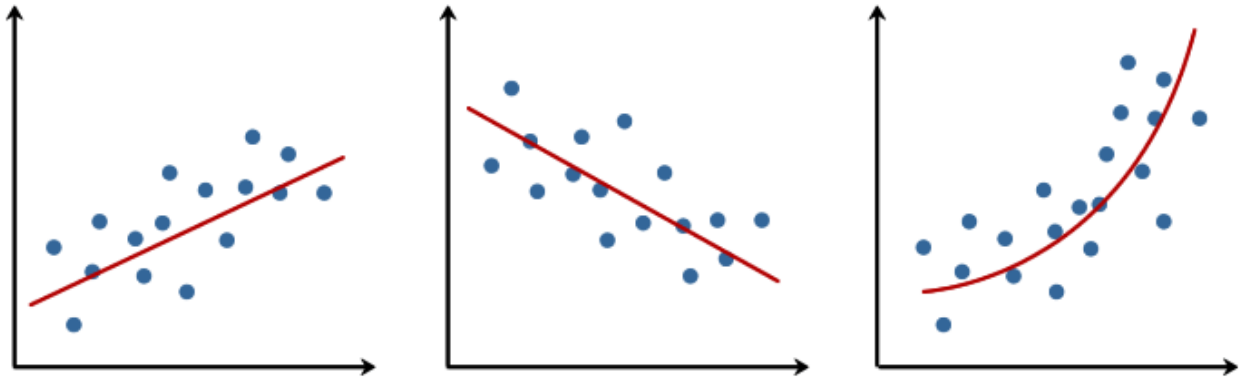
Linear regression assumes

- The relationship between X and Y is linear
- Y is distributed normally at each value of X
- The variance of Y at every value of X is the same (homogeneity of variances)
- The observations are independent

How well a set of data points fits a straight line can be measured by calculating the distance between the data points and the line.



Non-linear Regression



Linear regression is in a line. Non-linear regression is not. There may be a relationship that is non-linear, such as age and income. People tend to have the highest incomes when they are in their 50s. Their income tends to be lower when they are younger or older than their 50s. That is, income goes up and then goes down. There is a clear relationship, but it is not linear. A linear regression analysis would miss the relationship because it looks for how close the data are to a line. Non-linear regression could pick up a relationship that exists - that linear regression would miss. Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression relates the two variables in a nonlinear (curved) relationship.

Evaluation of Regression Models

The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.
- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

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

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

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

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ $-$ *predicted value of* y

$\bar{y}$ $-$ *mean value of* y