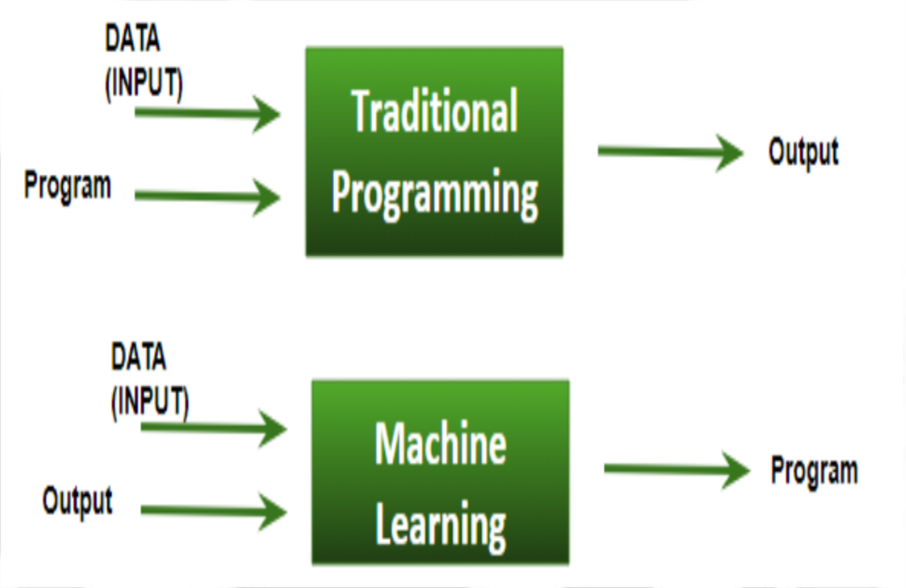
Chapter 3: Machine Learning Algorithms

Machine Learning

Machine learning (ML) is the part of Artificial Intelligence which allows the software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. ML algorithm uses historical data as input to predict new output values.

In other words, we can say it as a field of study that gives the computers the capability to learn without being explicitly programmed.

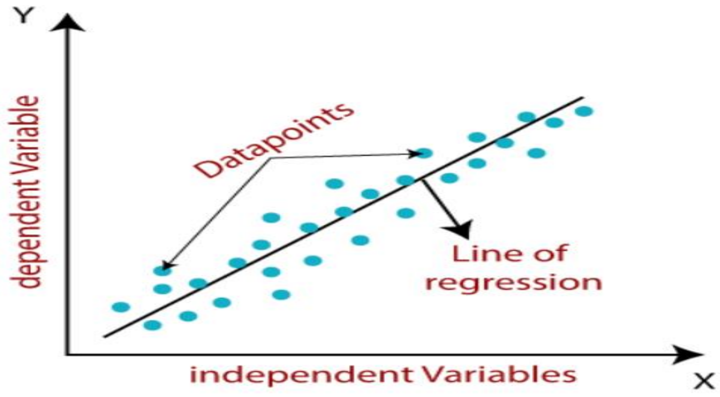


1. Linear Regression Algorithm

Linear regression is one of the most popular used ML algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (x) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.



Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1 x + \varepsilon$$

Here,
Y= Dependent Variable (Target Variable)
X= Independent Variable (Predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation. Linear regression can be further be divided into two types of the algorithm:

- ***Simple Linear Regression*:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- ***Multiple Linear regressions*:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Cost function

The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line. Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing. We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$\text{MSE}= 1\frac{1}{N}\sum_{i=1}^{n}(y_i \;- (a_1x_i + a_0))^2$$

N=Total number of observations
Yi = Actual value
(a1xi+a0) = Predicted value.

Residuals

The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent

Gradient descent is used to minimize the MSE by calculating the gradient of the cost function. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization. It can be achieved by R-squared method.

R-squared is a statistical method that determines the goodness of fit. It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%. The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model. It is also called a coefficient of determination, or coefficient of multiple determinations for multiple regressions.

It can be calculated from the below formula:

$$\text{R-squared}=\frac{\text{Explained variation}}{\text{Total Variation}}$$

For the above linear equation, MSE can be calculated as:

$$MSE \;=\; \frac{1}{n}\,\Sigma\,\underbrace{\left(\,y \;-\; \widehat{y}\,\right)^2}_{\substack{\text{The square of the difference}\\ \text{between actual and}\\ \text{predicted}}}$$