Murphy Campbell
Tech Talent Academy
Data Academy Home Learning 24/11/2021

Machine Learning algorithms:

Linear Regression, K-Means, and Support Vector Machine

Linear Regression

Linear Regression is a statistical method that is used for predictive analysis, it is a supervised learning algorithm. Linear regression makes predictions for continuous/real or numeric variables.

Linear Regression function:

Y = a0 + a1*X1 + error

We have:

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).

error = random error

The objective of the linear regression algorithm is to find the best straight line of fit through the data points that minimizes the error between predicted values and actual values. The coefficient a0, a1 will have to be found that produces the best line of fit, a cost function is utilised. A cost function for linear regression is (MSE) Mean Square Error, which produces the average of error between the predicted values and the actual values.

MSE = 1/N * sum i (Yi - Yhat)** 2, for sum i = 1 to N

Where

N=Total number of observation

Yi = Actual value

Yhat = (a0 + a1*Xi) = Predicted values

Assumptions of Linear Regression:

- Linear relationship between the dependent and independent variables.
- Small or no correlation between the independent variables
- Error term follows the Normal Distribution

Application:

Since the true underlying relationship between human weight and human height is linear, linear regression is great for predicting weight using height as the input variable or, vice versa, for predicting height using weight as the input variable

K-Means

K-Means is an unsupervised learning algorithm that utilizes clustering analysis techniques which enables us to find groups of similar objects that are more related to each other than to objects in other groups.

The algorithm forms k-many clusters (groups), by first randomly initiating the clusters as k-many points in the data space. Each of these points is the mean of a cluster.

An iterative process then occurs, running as follows:

- Each point is assigned to a cluster based on the least (within cluster) sum of squares, which is intuitively the nearest mean.

- The center (centroid) of each cluster becomes the new mean. This causes each of the means to shift. Over enough iterations, the centroids move into positions that minimize a performance metric (the performance metric most commonly used is the "within cluster least sum of squares" measure). Once this measure is minimized, observations are no longer reassigned during iteration; at this point the algorithm has converged on a solution.

Application:

Grouping of documents, music, and movies by different topics

Support-Vector Machine

Support-Vector Machines fall under supervised learning, they can be used for classification or regression tasks. The algorithm constructs a hyperlane or set of hyperplane boundaries in finite-dimensional space to separate the classes of data points, the object of the algorithm is to maximise the separation of the hyperplane boundary distances, thus classifying the data classes correctly.

When the data is non-linear the dimension will have to increase, and with the use of non-linear kernels will allow the algorithm to perform hyperplane decision separation of the data points into hyperplane boundaries that will contain the separated classes.

Application:

Classification of images, text

References:

https://en.wikipedia.org/wiki

https://www.javatpoint.com/machine-learning