LinearVsLogisticRegression

January 31, 2020

0.1 Linear Vs. Logistic Regression

1 Linear regression

Linear Regression is a form of regression analysis best used to quantify relationships between points in a continuous dataset. It fits a straight line onto a dataset and attempts to minimize the distance between all the values and the best fit line in order to make predictions about possible future points. This implies that the accuracy of the prediction (hyperplane) is directly proportional to the spread of the data and is therefore most useful on data that is relatively uniform with few outliers.

The equation used to describe Linear Regression is,

$$Y = bx + a$$

Where

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

and,

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Linear Regression can be applied to datasets with multiple features where the hyperplane is a subspace dimensionally equivalent to (features -1).

This method is useful in predicting things like housing prices, crypto currency price trends and other continuous data.

2 Logistic regression

Logistic Regression is a form of regression analysis best used to quantify relationships between points in a discrete (categorical) dataset. Predicting discrete variables is useful in predicting binary outcomes. Is someone a potential customer or not, is an email spam or not, is something A or B?

The input variables used to predict the dependant variable however can be either discrete or continuous however making it more useful than Linear Regression. It fits a straight line onto a dataset and attempts to minimize the distance between all the values and the best fit line in order to make predictions about possible future points. This implies that the accuracy of the prediction (hyperplane) is directly proportional to the spread of the data and is therefore most useful on data that is relatively uniform with few outliers.

Logistic Regression uses a sigmoid function,

$$Y = \frac{1}{1 + e^{-x}}$$

Where x = The variable you wish to transform and, e = Euler's constant, 2.718

The sigmoid function produces an S-shaped curve and maps it onto a numerical value between 0 and 1. Logistic regression assigns each data point to a discrete class. It also attempts to minimize the distance between each point and the hyperplane but also divides the data into classes using maximum likelihood estimation (MLE). The logistic hyperplane therefore acts as a classification boundary rather than a prediction line. Another difference is that in logistic regression the independent variable can be plotted along bth axes, and and the output of the dependant variable is determined by the position of the data point in relation to the hyperplane.

This method is useful in predicting things like fraud detection, disease diagnosis, emergency detection, loan default detection or spam identification.