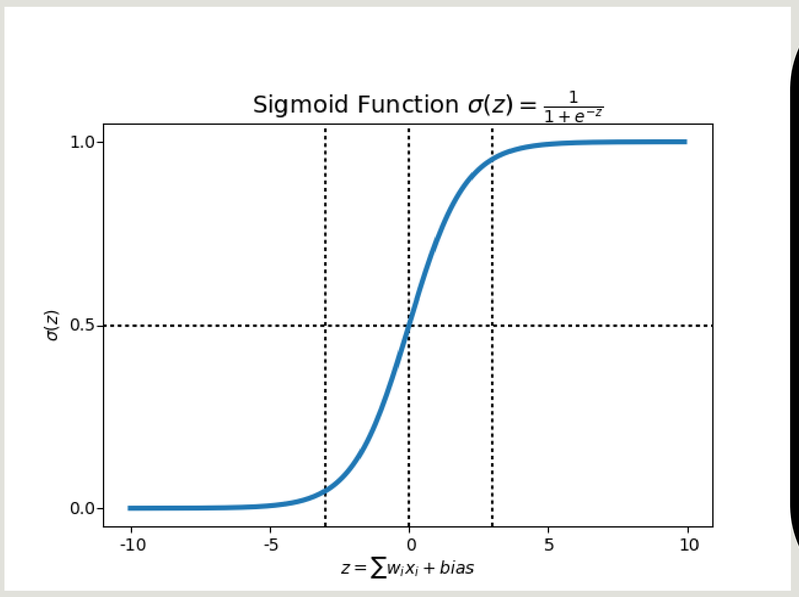
LOGISTIC REGRESSION

What is Logistic regression ?

Logistic regression is used for predicting the categorical dependent variable using a given set of independent variable. Output must be discrete vale i.e. either 0 or 1 but instead of giving value in 0 or 1 it give probabilistic values which is between 0 and 1. It is used to solve classification problems.

SIGMOID FUNCTION

Sigmoid Function acts as an activation function in machine learning which is used to add non-linearity in a machine learning model . The most commonly used sigmoid function in ML works with inputs of any real-value with its output value being between one and zero. We can't go beyond limit, so it forms curve like "S" form. Such as values above the threshold value tends to 1, and a value below the threshold value tends to 0.



Equation for logistic regression:
logit(p) = ln(p/(1-p))
=b0+b1X1+b2X2+b3X3….+bkXk

Where;
p= probability of the occurrence of the feature
x1,x2,..xk= set of input features of x
b1,b2…bk= parameter values to be estimated in the logistic regression formula

How does it work ?

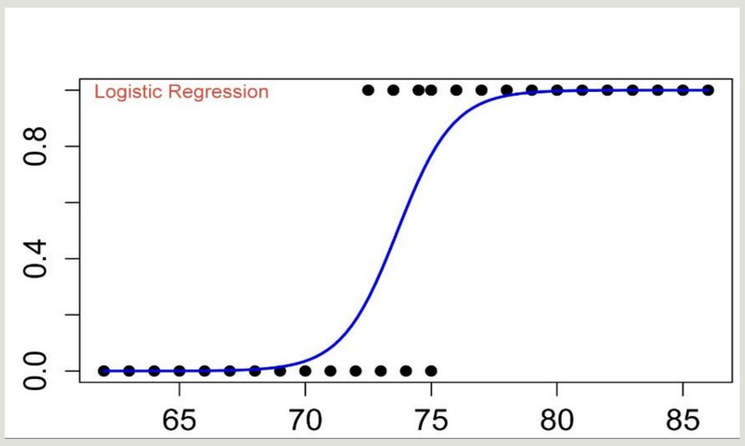
The Sigmoid function (logistic regression model) is used to map the predicted predictions to probabilities. The Sigmoid function represents an 'S' shaped curve when plotted on a map. The graph plots the predicted values between 0 and 1. The values are then plotted towards the margins at the top and the bottom of the Y-axis, with the labels as 0 and 1. Based on these values, the target variable can be classified in either of the classes.

Prediction in logistic regression

Logistic regression is carried out in cases where your response variable can take one of only two forms (i.e. it is binary). There are two general forms your response variable can take:

1. Presence/absence, that is, 0 or 1 (or some other binary form).
2. A success/failure matrix, where you have two frequencies for each observation (the "successes" and the "failures").

The way your data are arranged does not make much difference to how you carry out the predictions but because of the different forms it is useful to see an example of each.



Advantages

- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
- It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).
- It can interpret model coefficients as indicators of feature importance. Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.
- One may consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.

Disadvantages

- It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
- Non-linear problems can't be solved with logistic regression because it has a linear decision surface.
- Logistic Regression requires average or no multicollinearity between independent variables. It is tough to obtain complex relationships using logistic regression.
- In Linear Regression independent and dependent variables are related linearly. But Logistic Regression needs that independent variables are linearly related to the log odds (log(p/(1-p)).

Kirti rathore
0827CS193D09