Lecture 12

Logistic Regression

Some regression algorithms can be used for classification (and vice versa). Logistic Regression (also called Logit Regression) is commonly used to estimate the probability that an instance belongs to a particular class (e.g., what is the probability that this email is spam?). If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labeled "1"), and otherwise it predicts that it does not (i.e., it belongs to the negative class, labeled "0").

Logistic regression mainly is used for binary classifications. You can train your model to predict if someone is cancer or not or your model could be trained to predict if it is cat or not in the photo.

The **logistic regression** technique involves dependent variable which can be represented in the binary (0 or 1, true or false, yes or no) values, means that the outcome could only be in either one form of two. For example, it can be utilized when we need to find the probability of successful or fail event. Here, the same formula is used with the additional sigmoid function, and the value of Y ranges from 0 to 1.

Logistic regression equation :

Linear regression $Y = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$

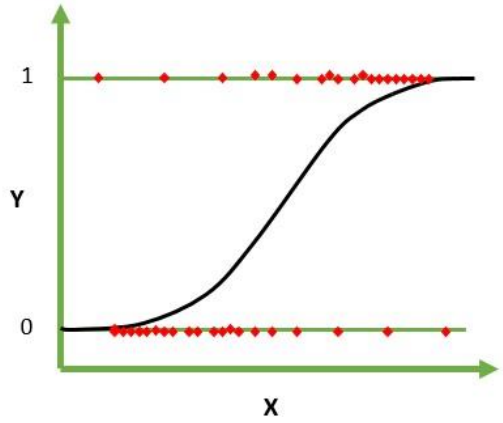
Sigmoid Function $P = \frac{1}{1+e^{-Y}}$

The sigmoid function, also called logistic function gives an 'S' shaped curve that can take any real-valued number and map it into a value between 0 and 1. If the curve goes to positive infinity, y predicted will become 1, and if the curve goes to negative infinity, y predicted will become 0. If the output of the sigmoid function is more than 0.5, we can classify the outcome as 1 or YES, and if it is less than 0.5, we can classify it as 0 or NO. For example: If the output is 0.75, we can say in terms of probability as: There is a 75 percent chance that patient will suffer from cancer.

By putting Y in Sigmoid function, we get the following result.

$$\ln\left(\frac{P}{1-P}\right) = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$$

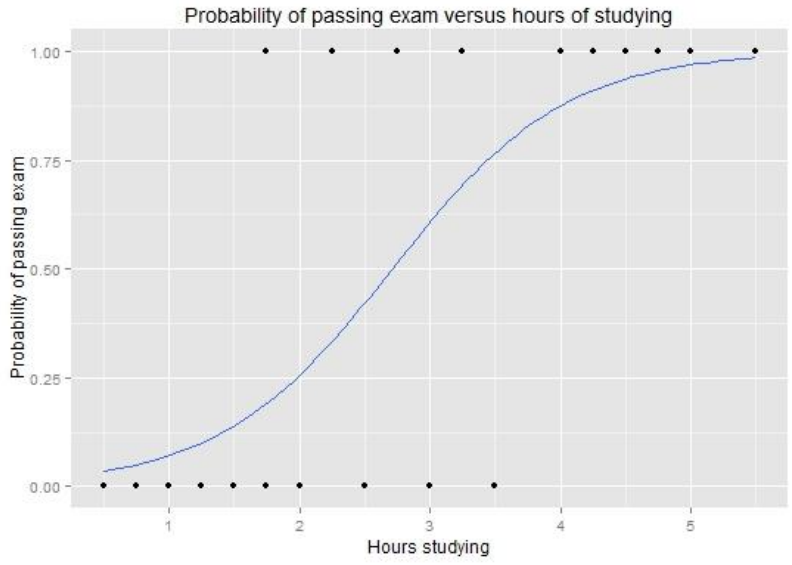
The following graph can be used to show the logistic regression model.



As we are working here with a binomial distribution (dependent variable), the link function is chosen which is most suitable for the distribution. In the above equation, the parameters are chosen to maximize the likelihood of observing the sample values instead of minimizing the sum of squared errors (such as linear regression).

Logistic functions are used in the logistic regression to identify how the probability P of an event is affected by one or more dependent variables.

Example



Types of Logistic Regression

Types of Logistic Regression:

- Binary Logistic Regression: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.
- Multinomial Logistic Regression: The target variable has three or more nominal categories such as predicting the type of Wine.
- Ordinal Logistic Regression: the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.

Advantages

Because of its efficient and straightforward nature, doesn't require high computation power, easy to implement, easily interpretable, used widely by data analyst and scientist. Also, it doesn't require scaling of features. Logistic regression provides a probability score for observations.

Disadvantages

Logistic regression is not able to handle a large number of categorical features/variables. It is vulnerable to overfitting, That is, the models can appear to have more predictive power than they actually do as a result of sampling bias. In the college admissions example, a random sample of applicants might lead a logit model to predict that all students with a GPA of at least 3.7 and a SAT score in the 90th percentile will always be admitted. In reality, however, the college might reject some small percentage of these applicants. A logistic regression would therefore be "overfit," meaning that it overstates the accuracy of its predictions. Also, can't solve the non-linear problem with the logistic regression that is why it requires a transformation of non-linear features. Logistic regression will not perform well with independent variables that are not correlated to the target variable and are very similar or correlated to each other.

Linear Regression Vs. Logistic Regression

Linear regression gives you a continuous output, but logistic regression provides a constant output. An example of the continuous output is house price and stock price. Example's of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn. Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

Key Differences between Linear and Logistic Regression

1. The Linear regression models data using continuous numeric value. As against, logistic regression models the data in the binary values.
2. Linear regression requires to establish the linear relationship among dependent and independent variable whereas it is not necessary for logistic regression.
3. In the linear regression, the independent variable can be correlated with each other. On the contrary, in the logistic regression, the variable must not be correlated with each other.