Motivation: - This blog post will describe nuance of Logistic Regression, there would be a many article you can find out there, but in this blog, I am trying to collect all elements of Logistic Regression in one stop, which will great reference.

I am going to cover three topics.

1. Univariate Logistic Regression
2. Multivariate Logistic Regression
3. Model Evaluation

Univariate Logistic Regression:

Before jumping to Logistic regression lets understand some basic concepts related to this topic.

1. Binary Classification
2. Sigmoid function
3. Likelihood function
4. Odds and Log odds

The most common use of logistic regression model is in binary classification, the classification problem is the one where is the output is categorical variable instead numerical variable, regression output is numerical where else classification output is categorical variable.

For example:

A finance company who lends out the loan to its customer wants to know whether customer will get default or not and the reason is they want to make sure customer doesn’t run away, the money they have given.

Now through EDA (Exploratory data analysis) you identify strong indicators of default and you have gotten some variable such as income, purpose of loan etc. now what classification model enable us to extract similar pattern of data and short of classify into whether they would default or not. So, we have these two categories which we need to find out. Other example like email classification spam or ham.

Let’s take some concreate example of classification of diabetes or non-diabetes

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Fig-1 Fig-2

As we can see in Fig-1 Diabetes is categorical variable, now in fig-2 we code the “No” and “Yes” into 0 and 1. Now we have plotted this dataset blood sugar level at x-axis and diabetes in y-axis.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Fig-3 Fig-4

In fig-2 if you see few points are showing diabetes and non-diabetes, so one way we can decide through decision boundary in fig-4, for example we can say all patent who’s suger level more then 210 the value of diabetes variable is 1 which means they diabetes, and patents who’s suger level is bellow 210 is non-diabetes. So in that case our predication is represented by fig-5 curve, however there is a problem in this boundary based curve (fig-5), clearly in this case fig-5 we are misclassifying two patens (the 2 squire bracket pointer in fig-5).

A close up of text on a white background

Description automatically generatedA close up of text on a white background

Description automatically generated

Fig-5 Fig-6

So is there a decision boundary which help us to get zero misclassification, I doubt, there is no decision boundary for which there is no misclassification, the best way to make the cutoff roughly by 195 in fig-6 so there will be only one misclassification, however even in this case Fig-6 we will have one misclassification.

So what is problem with this decision boundary approach (please add you thought in comment), clearly deciding the class blatantly on the basis of cutoff seems very risky specially near the middle of the graph, this sharp boundary hearting as lot because patients intermediate suger level either belongs to diabetes or non-diabetes.

One way of overcoming this problem where we have very sharp boundary