Short & Sweet (Episode 2)

Logistic Regression

Logistic Regression is a **Machine Learning Algorithm** which is actually not an algorithm to predict the continuous value, it is classification problem which is inspired from the concepts of **Linear Regression**. This algorithm is beneficial for classification task to classify real world use cases like,

- Fraud Detection (Email is fraud or not)
- Student passed in examination or not and many more.

Benefits of Logistic Regression are -

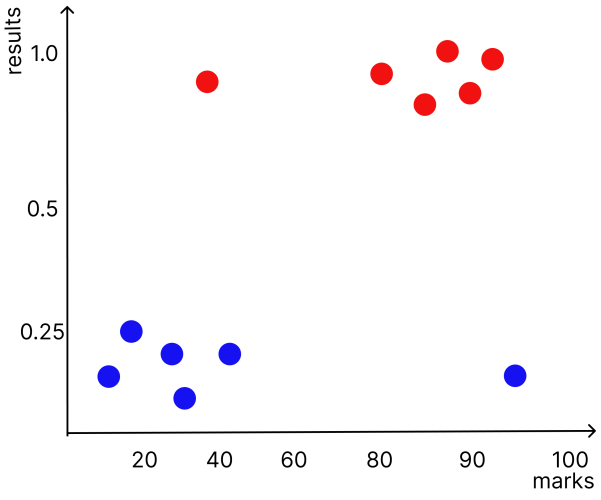
- It is easier to implement and efficient to train.
- It is accurate to simple datasets, and it performs well when data is **linearly separable**.
- It is less inclined to **overfitting,** but it can be overfit in high dimensional datasets.

Note - It is mostly used for Binary classification (having at most two categories)

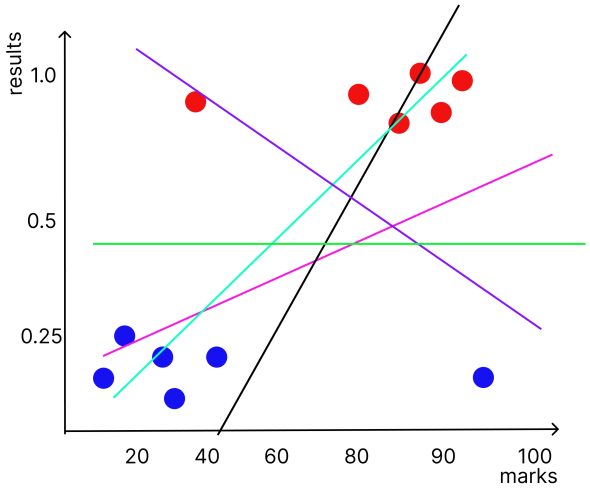
How It Works

As I've said, this algorithm is inspired from **Linear Regression,** Let's see the datapoints (based on classification) and try to fit a best line.

- Assume a dataset of student having marks and result that they are pass or not plotted as below -
  - Here, we can see students with marks less than 50 has less probability to pass in an exam whereas students who scored more than 60 has high probability to pass.
  - **1.0** has means student has high probability of passing and below **0.5** has less probability to pass.



- Now, let's try to fit some lines and observe is there any line that can be the best option.
  - We can see below, any straight line can't be the best fit, and it also can't detect the **outliers** (some exceptional points different from regular pattern).
  - We need to make some change (or apply some function) to line so that it can classify the point instead of regression task.



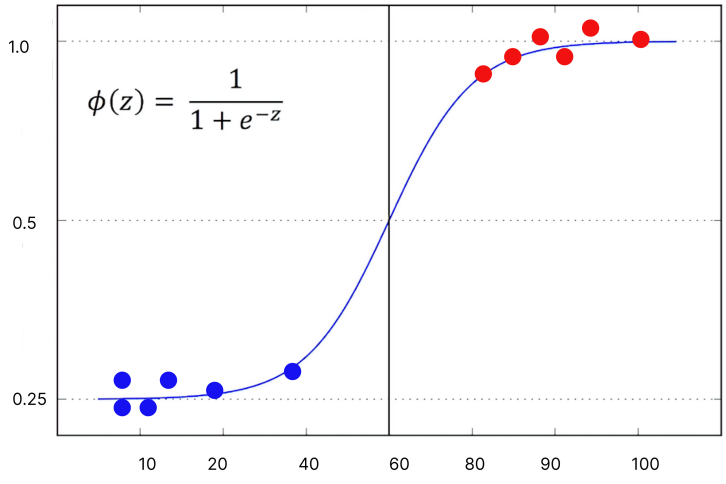
- We'll try to make line little bit curvy using **sigmoid function** to our equation of line.
- The **sigmoid function**, also known as the **logistic function**, is a crucial component used to model the probability that a given input belongs to a particular category or class. It's an S-shaped curve that maps any real-valued number to a value between 0 and 1. The formula for the sigmoid function is:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

  - Here ***z*** is the equation of line with optimal parameter i.e. **weights** and **bias** using the previously discussed linear regression technique. It is defined as -

$$z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

- On applying sigmoid function, we'll get the below resultant line -



- Basically, we are trimming our regressed output in terms of probability between 0 and 1 using sigmoid function so that it can classifies two targets efficiently.
- Some terms about which you should also know -
  - **Maximum Likelihood** - Maximum likelihood is a statistical method used to estimate the parameters of a probability distribution by finding the values that maximize the likelihood of observed data occurring under the assumed model. It aims to identify the most probable values that make the observed data most likely to have been generated from the given model.
  - **Logit Function** - It is a mathematical transformation used in statistics, particularly in logistic regression, to model probabilities. It transforms the probability, which ranges between 0 and 1, to a log-odds scale that ranges from negative infinity to positive infinity.
  - **Log odds** - Log odds, short for logarithm of odds, refer to the logarithm of the ratio of the probability of success to the probability of failure in a binary event. It is a way to represent the likelihood or probability of an event occurring in a logarithmic form.

    The formula for calculating log odds is:

$$\text{Log Odds} = \log\left(\frac{p}{1-p}\right)$$

Assumptions of Logistic Regression

- The response variable is binary.
- The observations are independent.
- The model is correctly specified.
- The outcome variable and all predictors are measured without error.

- Each predictor is related linearly to the odds ratio.
- There is no requirement for a linear relationship between the dependent and independent variables.