Introduction to Logistic Regression

NOTE : Implementing these Algorithms is dead easy but understanding them is a bit challenging.

What is a Classification problem ?

Classification is the problem of identifying which set of categories or class does a particular sample belong. In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

Example :-
1) Identify if a mail is spam or not ?
2) Identify is a person has cancer or not ?
3) Identify if animal is a dog or cat ?

In Classification the "Target" variable is always Categorical

Structured classification dataset

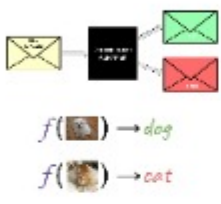
(Example dataset for diabetes test)

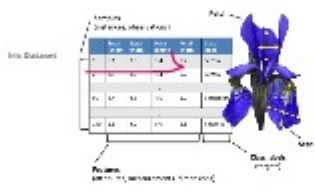
	Age	Glucose	BMI	OUTCOME
1]	34	135	34.3	YES
2]	58	85	25.1	NO
3]	25	97	44.8	NO
4]	29	102	23.1	YES

In Binary ---> 2 Classes
In Multiclass ---> More than 2 classes

Types of Classification problems

- Binary Classification
  (Binary classification refers to predicting one of two classes)
- Multiclass Classification
  (Multi-class classification involves predicting one of more than two classes)





NOTE : For Unstructured data we use Deep Learning.

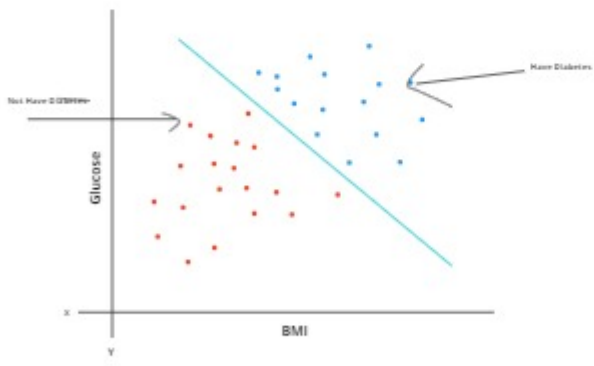
Some common ML model for classification :

- Logistic Regression
- Naive Bayes
- Stochastic Gradient Descent
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Support Vector Machine

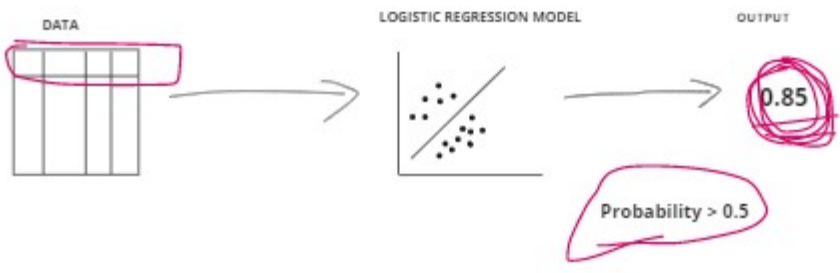
Introduction to Logistic Regression

It is a machine learning model which is used to solve classification problems. By default it only supports "binary" classification and cannot be used for multiclass classification.

NOTE : In Linear Regression we try to find the "Line of Best Fit",but in Logistic Regression we try to find the "Line of Best Seperation". This line perfectly seperates the two classes.



The Logistic Regression model in simple terms, takes the "DATA" and output a probability value. It us upon us to set a THRESHOLD (most common threshold is 0.5

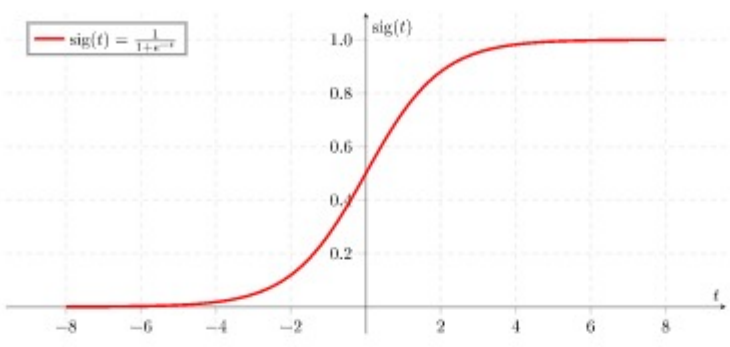


Logistic Regression at its core uses "Sigmoid Function"

$$S(x) = \frac{1}{1 + e^{-x}}$$

This function convert any real value in the range of (0,1) , Higher values are given higher probabilities

The Logistic Regression model uses this function to predict the probabilities



Cost Function in Classification

The most common cost function for regression problem is "MSE". But for classification problems we have different cost functions. The most common cost function for classification is "CROSS ENTROPY"

NOTE : We have different cost function for Binary & Multiclass Classification

CROSS ENTROPY

- Binary Cross Entropy
- Categorical Cross Entropy

Binary Cross Entropy (BCE)

(Loss Function)

$$BCE\ (t,p) = \sum -(t * \log(p) + (1-t) * \log\ (1-p))$$

t = target value
p = predicted value

$$BCE(t,p) = \begin{cases} -\log(1-P) & \text{if } t = 0 \\ -\log(P) & \text{if } t = 1 \end{cases}$$

(Example on how to calculate BCE)

	Age	Glucose	BMI	t (OUTCOME)	P (PREDICTED)	
1]	34	135	34.3	1	0.6	0.22
2]	58	85	25.1	0	0.2	0.09
3]	25	97	44.8	1	0.5	0.3
4]	29	102	23.1	0	0.8	0.6

LOSS = 1.21

DEMO LOGISTIC REGRESSION USING PYTHON