Logistic Regression

15 May 2023 20:19

Logistic Regression

* It comes under Supervised Learning [both x and y]
* It is used to solve classification problems. [target is categorical or discrete]

Eg: Diabetic data
Cancerous data
Email classification, Covid19, Heart disease.

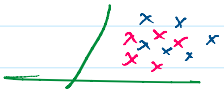
* It is used to predict the probability of categorical dependent variables.

Assumptions of Logistic Regression:

* The target variable is binary ie Yes/No, True/False, 0/1,
* No or little multicollinearity → Less correlation among input variables.
* Outliers should be handled.
* Sample size should be sufficiently large.
* Data is linearly seperable.

Linearly seperable → There should be clear seperation among the classes

Non-Linear → overlapping of classes.



How Logistic Regression works?

Problem statement:
I want to predict whether student will pass or not in exam.

Data

No of hours studied (x)	Result (y)
1	Fail
4	Fail
3	Fail
8	Pass
10	Pass
12	Pass
5	?

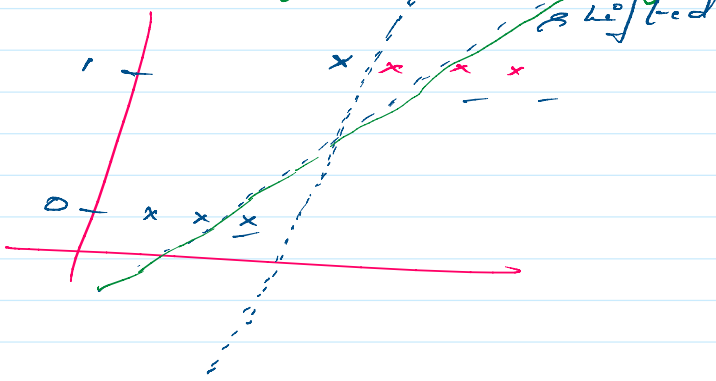
Pass - 1
Fail - 0

What happens if we try to use linear Regression?

* Linear regression will fit a straight line which seperates two classes.

$y = mx + c$

* If we use linear regression to make predictions it will use straight line and gives prediction other 0 and 1



* If data has outliers then line gets shifted, it results in miss classification

Sigmoid function

Sigmoid function is mathematical function that we use to transform continuous values into a probability values which lies within a range 0 to 1.

$$h_\theta(x) = g(z) = \frac{1}{1+e^{-z}} \longrightarrow [0, 1]$$

where $z = mx + c$ or $z = \beta_0 + \beta_1 x$] st line

x	Result	$z = mx + c$	$h_\theta(x)$	Predictions
2	Fail (0)	-50	0.2	Fail
3	Fail (0)	-40	0.4	Fail
5	Fail (0)	-20	0.3	Fail
8	Pass (1)	70	0.6	Pass
10	pass (1)	80	0.8	Pass
12	pass (1)	120	0.9	Pass

Set a Threshold

$h_\theta(x) \geq 0.5 \longrightarrow y = 1$ / Pass

$h_\theta(x) < 0.5 \longrightarrow y = 0$ / Fail

Break

9.50 → 10 pm

How it works?

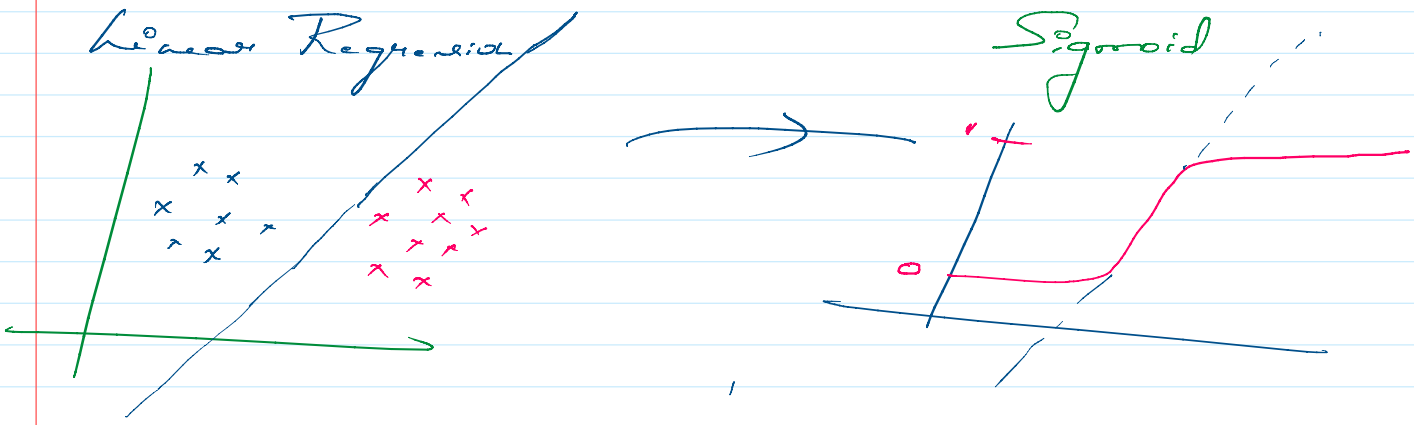
→ It will fit line

→ Make predictions

→ Transform continuous values using sigmoid function

→ Set threshold

→ If threshold is greater than 0.5 classify as pass and if threshold is less than 0.5 classify as Fail



How classification works?

$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + c$

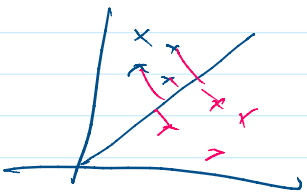
$\rightarrow y = w^T x + c$

where $w^T = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}$ $x = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$

Let us say decision boundary through origin, $C = 0$.

$y = w^T x$

How to find distance b/w point and plane?



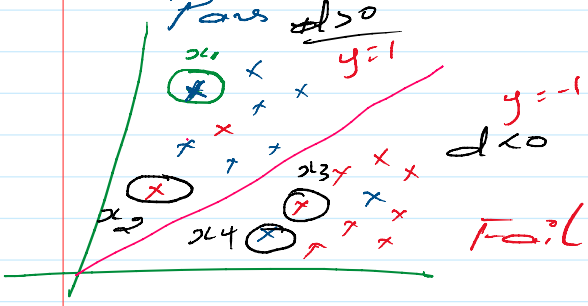
$d = \frac{w^T x}{|w|}$

$|w| = 1$

$d = w^T x$ $\quad y =$

$y \times d = y \times w^T x \geq 0$, then we say the model has

$y \times d = y \times w^T x \geq 0$, then we say the model has done correct classification.



* Any point above the line/plane will have positive distance
* Any point below the line/plane will have negative distance.
* All positive classes are denoted by 1
* All negative classes are denoted by -1

Case 1
x_1 belongs to positive class,
$y = 1$, $w^T x > 0$
$y \times w^T x = (+ve)(+ve) = +ve$
$y \times w^T x > 0$
x_1 is Correctly classification

Case 2
x_2 belongs to negative class
$y = -1$, $w^T x > 0$
$y \times w^T x = (-ve)(+ve) = -ve$
$y \times w^T x < 0$
x_2 is misclassified

Case 3
x_3 belongs to negative class
$y = -1$, $w^T x < 0$
$y \times w^T x = (-ve)(-ve) = +ve > 0$
x_3 is Correct classification

Case 4
x_4 belongs to positive class
$y = 1$, $w^T x < 0$..
$y \times w^T x = (+ve)(-ve) = -ve$
$y \times w^T x < 0$.
x_4 is misclassified