Introduction To Logistic Regression: Understanding The Basics

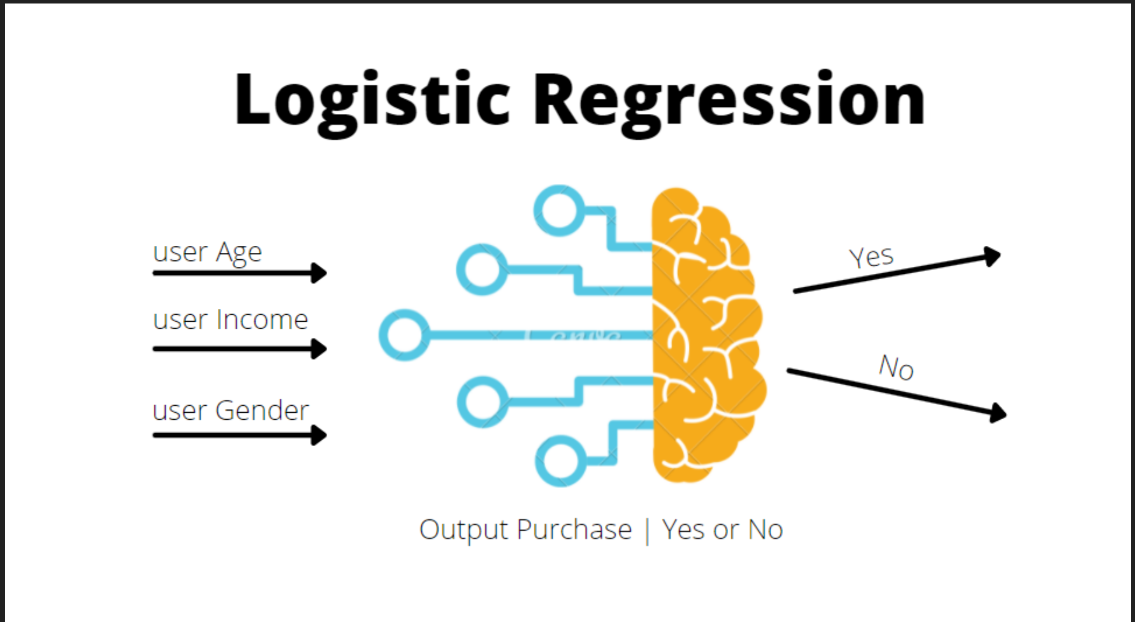
By M.H.Khan

Context

- Introduction To Logistic Regression
- Examples of Logistic Regression
- Transition from Linear To Logistic Regression
- Logistic Regression Equation
- Sigmoid Function For Confining Output
- Linear Transformation or Log Odds Ratio
- Categorization of Output
- Why Logistic Regression
- Conclusion

Introduction To Logistic Regression

Logistic Regression: Logistic Regression is a machine learning technique commonly used for solving classification problems. Despite its name, it doesn't actually perform regression in the traditional sense. Instead, it's a method used to predict categorical outcomes based on input variables.



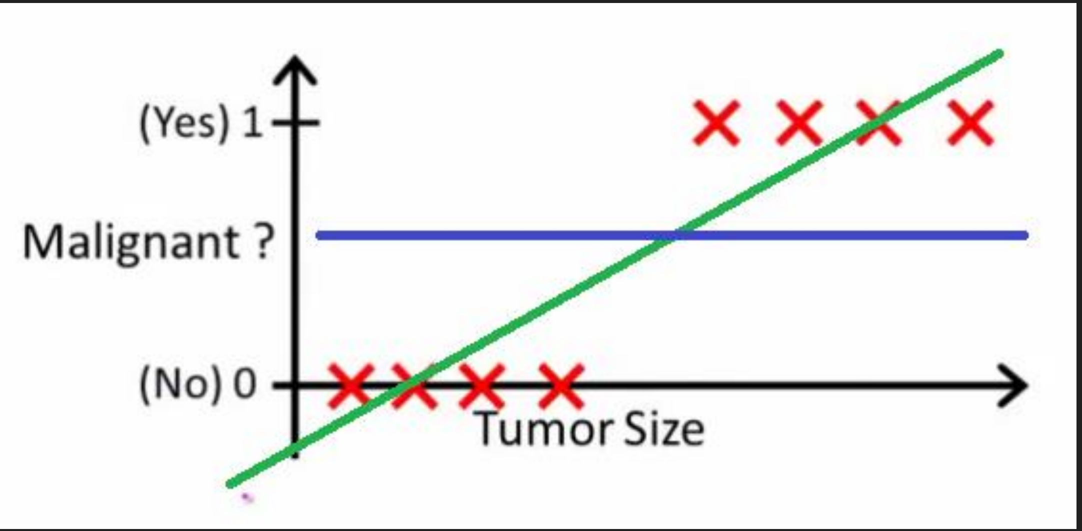
Examples of Logistic Regression

Logistic Regression finds its application in scenarios where the outcome needs to be categorized. Here are some instances:

1. **Student Performance:** Determining whether a student passes or fails an exam based on factors like study hours and attendance.
2. **Weather Prediction:** Predicting whether it will rain today or not based on atmospheric conditions.
3. **Color Classification:** Categorizing objects based on their color, like classifying a ball as red or blue.
4. **Size Categorization:** Deciding if an object is big or small based on its dimensions.

Transition from Linear to Logistic Regression

- **Limitation of Linear Regression:** While Linear Regression is excellent for predicting continuous numerical values, it falls short when dealing with categorical outcomes.
- **Linear Regression Equation:** The equation $y = \beta 0 + \beta 1 * x1$ represents the relationship between the input ($x1$) and the output (y) in linear regression.
- **Adapting to Categorical Data:** To handle categorical data, we need to transform the linear output into a probability range that fits between 0 and 1.

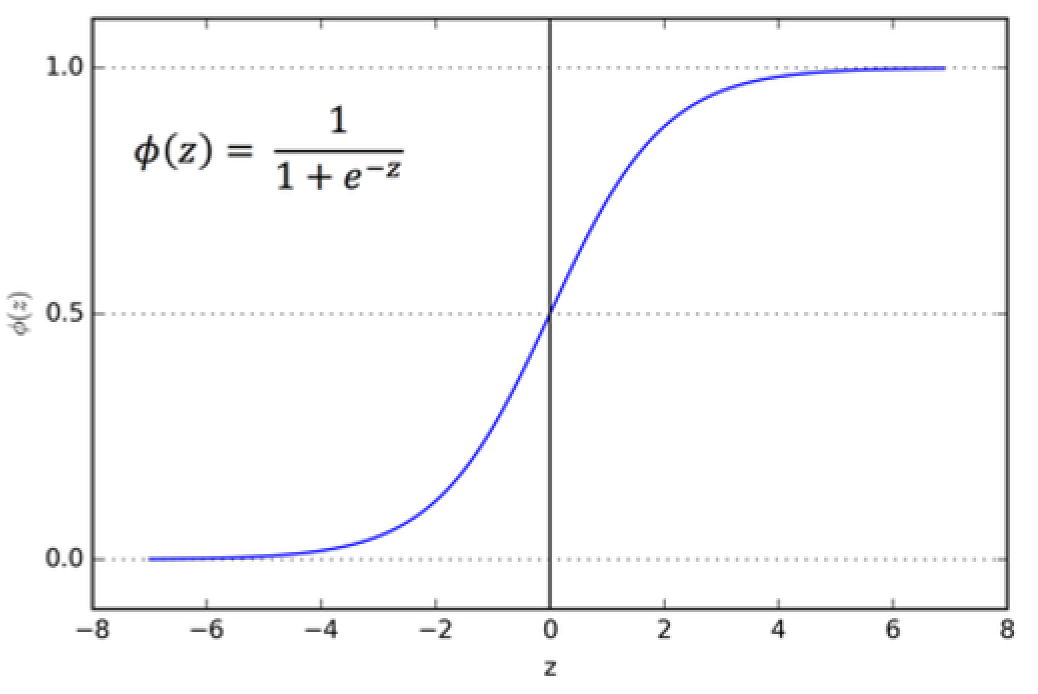


Logistic Regression Equation

- **The Logistic Equation:** Logistic Regression introduces a new equation: $P(y) = \beta 0 + \beta 1 * x1 + ... + \beta n * xn$. Here, $P(y)$ represents the probability of the binary outcome.
- **Constraining Output:** The goal is to ensure that the output probability remains within the bounds of [0, 1], which aligns with the concept of probability.

Sigmoid Function For Confining Output

- **Role of Sigmoid Function:** The sigmoid function, represented as $f(x) = \frac{1}{1+e^{-x}}$, is instrumental in achieving the probability constraint.
- **Applying to Logistic Regression:** In the context of logistic regression, x is replaced with the linear combination β0 + β1 * x1 + ... + βn * xn. The sigmoid function squashes this linear output to the range [0, 1].
- The logistic Regression will be in the form of Non-linear line



Range of Sigmoid Function

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

When z = -Inf,

$$g(z) = \frac{1}{1+e^{inf}} = 0$$

When z = inf,

$$g(z) = \frac{1}{1 + e^{-inf}} = \frac{1}{1} = 1\ (limit)$$

Therefore, $0 < g(z) < 1$

Note: $\exp(-Inf) = 0$

$\exp(Inf) = Inf$

Linear Transformation

P(Y = 1 | x) $= g(a + bx) = \frac{1}{1+ e^{-(a+bx)}}$

Let us write P(Y=1|x) as p(x), for simplicity

$$\Rightarrow \; p(x) = \frac{1}{1 + e^{-(a+bx)}}$$

$$\Rightarrow \frac{1}{p(x)} = 1 + e^{-(a+bx)}$$

, Taking Reciprocal

$$\Rightarrow \frac{1}{p(x)} - 1 = e^{-(a+bx)}$$

$$\Longrightarrow \frac{1 - p(x)}{p(x)} = e^{-(a+bx)}$$

$$\Longrightarrow \frac{p(x)}{1 - p(x)} = e^{(a+bx)}$$

$$\Longrightarrow \underbrace{log\left(\frac{p(x)}{1 - p(x)}\right) = a + bx}_{\text{Linear model}}$$

Log Odds Ratio

Odds ratio:

Odds Ratio is the ratio of two odds. For example, if the odds of getting lung cancer for smokers are 20 and the odds for non-smokers are 1, then the odds ratio is 20 : 1 = 20.
This means that smokers have 20 times higher odds of getting lung cancer than non-smokers

Odds = $\frac{probability\ of\ an\ event\ occuring}{Probability\ of\ an\ event\ not\ occurin}$

Odds ratio for smokers = $\frac{20(smokers)}{1(non{-}smokers)}$ = 20

$$\text{LogOdds} = log\left(\frac{p(x)}{1-p(x)}\right)$$

$$\text{LogOdds} = log(20)$$
$$\text{LogOdds} = 1.30103$$

Logistic Regression Model in Linear Form

$$\Rightarrow log\left(\frac{p(x)}{1-p(x)}\right) = a + bx$$

$$\Rightarrow log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = a + bx$$

Odds of Y=1

Interpreting the coefficient b:

With one unit increase in the value of x the log of odds of Y will increase by an amount equal to b, provided all the other factors remains constant.

Categorization of Output

- After obtaining the transformed output, categorization takes place.
- If the output value is above 0.5, it's classified as "True"; if it's below 0.5, it's classified as "False."
- This threshold-based approach converts the continuous probability into binary outcomes.

Why Logistic Regression?

When we are using the sigmoid function the output value ranges from 0 to1 which is in the continuous form based on the threshold value we decide the category type True/False that's why it is known as Logistic Regression

Conclusion

- Logistic Regression bridges the gap between linear equations and categorical outcomes.
- Leverages the Sigmoid function for probability transformation.
- An essential tool for solving classification problems with clear decision boundaries.