Logistic Regression

INNOMATICS
RESEARCH LABS

Contents

- Logistic Regression
- Types of Logistic Regression
- Logistic Function
- Model building
- Model Evaluation - Classification Metrics
    1. Confusion Matrix
    2. Area Under ROC Curve
    3. Classification Report

Logistic Regression

- Logistic Regression is a Classification algorithm and belongs to the family of **generalized linear model (GLM).**
- Logistic Regression is used when the dependent variable(target) is categorical.
- For example,
  - To predict whether an email is spam (1) or (0).
  - Whether the tumor is malignant (1) or not (0).

Types of Logistic Regression

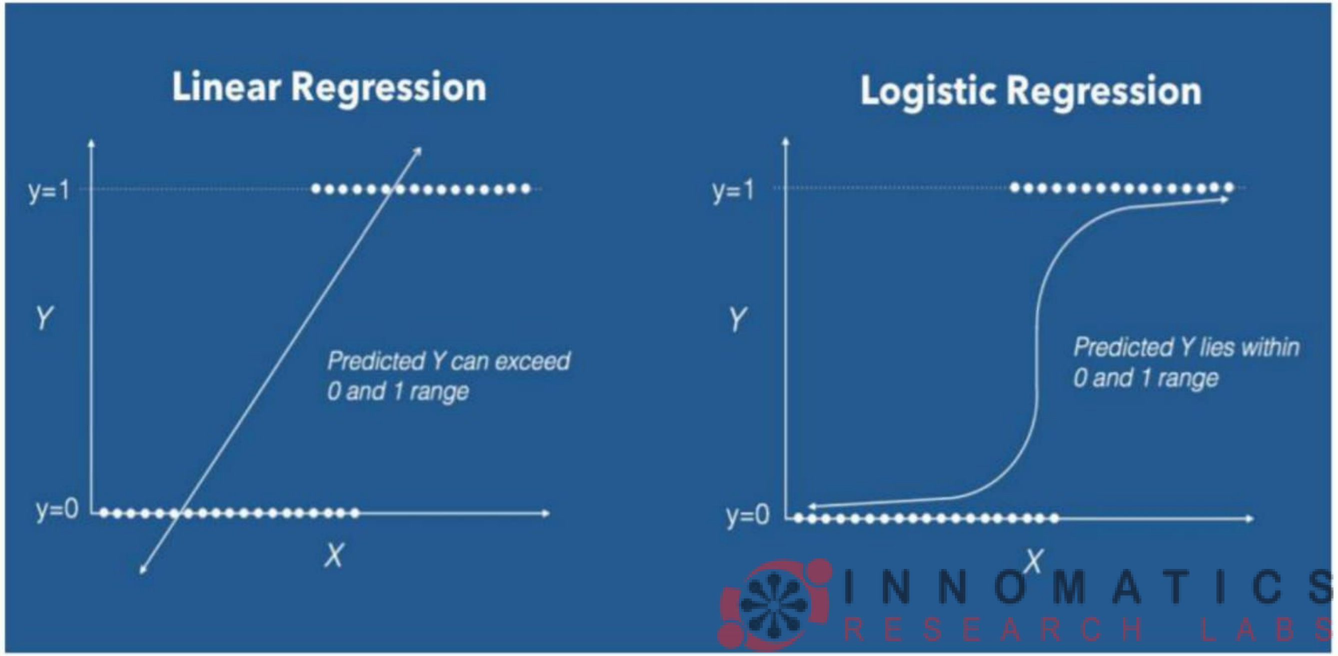
1. Binary Logistic Regression
    - The categorical response has only two 2 possible outcomes. Example: Spam or Not.

2. Multinomial Logistic Regression
    - Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan).

3. Ordinal Logistic Regression
    - Three or more categories with ordering. Example: Movie rating from 1 to 5.

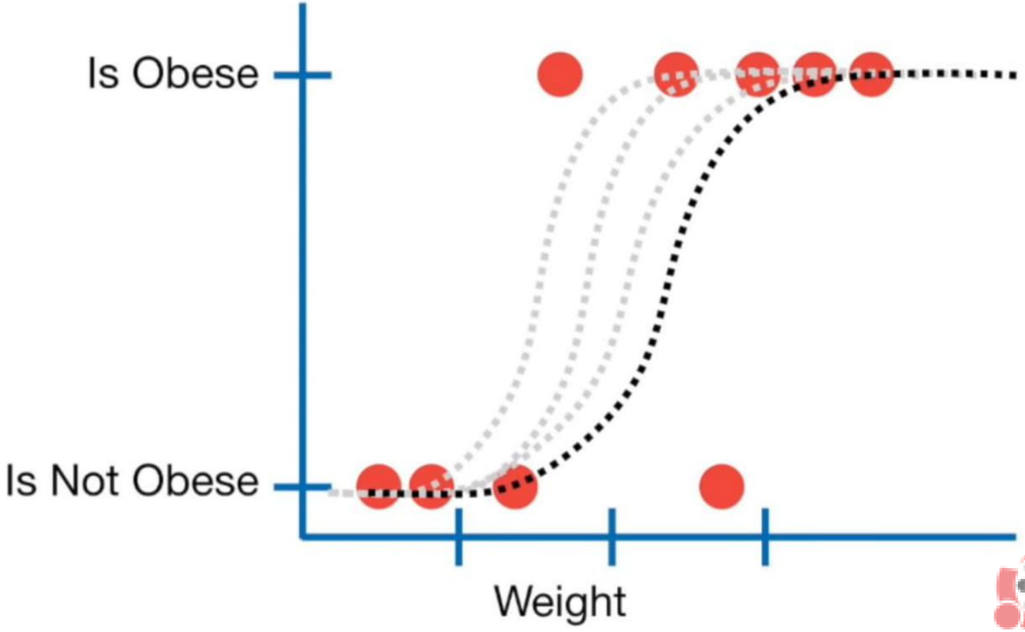
Linear Regression vs Logistic Regression

- Transforms Linear Function to a function that can predict categories with probabilities (0 to 1).
- Assigning observations to classes(Yes(1)/No(0)) depends on threshold probability.



Linear Regression vs Logistic Regression

- Logistic regression doesn't have the concept of "residual" so it can't use OLS method.
- It uses the approach called "Maximum Likelihood".



Logistic Function

- Logistic regression is named for the function used at the core of the method, the <u>logistic function</u>.
- A logistic function or logistic curve is a common "S" shape (**sigmoid** curve), with equation –

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

where

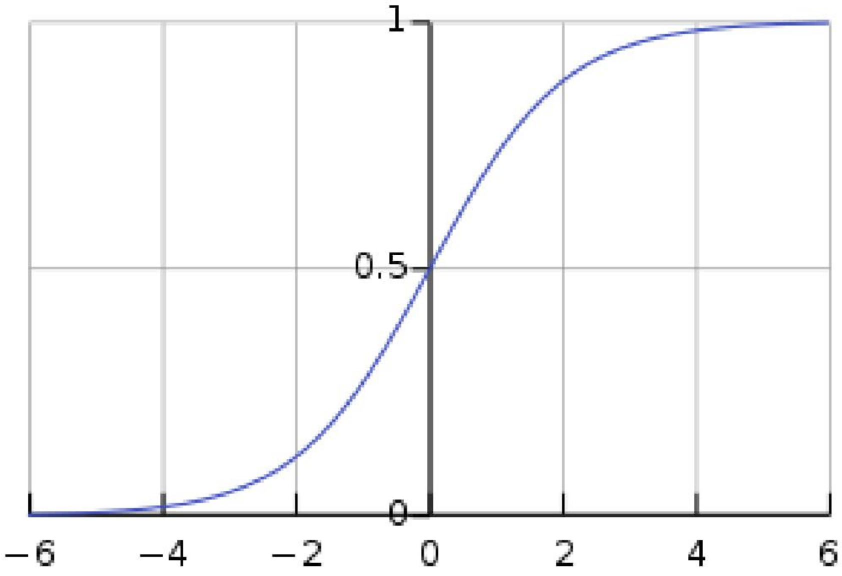
- e = the natural logarithm base (also known as Euler's number),
- x_0 = the x-value of the sigmoid's midpoint,
- L = the curve's maximum value, and
- k = the logistic growth rate or steepness of the curve.

Logistic Function

- Standard logistic sigmoid function i.e.

$$L = 1, k = 1, x_0 = 0$$

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} = \frac{1}{1 + e^{-x}}$$



Logistic Function

$$\text{Odds ratio} = \frac{p}{1-p}$$

Will Home Loan be *approved*?

Yes(0.78) No(0.22)

(p) **(1-p)**

Applying Odds ratio on Loan Approved example,

Odds of loan being *approved* is $\frac{0.78}{0.22} = 3.54.$

Logistic Function

- $probability = \frac{one\ outcome}{all\ outcomes}$

- $odds = \frac{one\ outcome}{all\ other\ outcomes} = \frac{probability}{1 - probability}$

- For example, drawing an Ace from a deck of cards –
  - Probability = 4/52
  - Odds = 4/48

Logistic Function

$$\text{logit(p)} = \ln \frac{p}{1-p} = ß_0 + ß_1 X$$

Logit or Sigmoid function :
Transformation of Linear Function into *non linear* function that predicts *categories*.

$$\ln\left(\frac{p(y)}{1-p(y)}\right) = ß_0 + ß_1 X$$

$$\frac{p(y)}{1-p(y)} = e^{ß_0+ß_1 X}$$

$$\text{p(y)} = e^{ß_0+ß_1 X} - \text{p(y)}.(e^{ß_0+ß_1 X})$$

$$\text{p(y)}+\text{p(y)}.(e^{ß_0+ß_1 X}) = e^{ß_0+ß_1 X}$$

$$\boxed{\boldsymbol{p(y) = \frac{e^{ß_0+ß_1 X}}{1+e^{ß_0+ß_1 X}}}}$$

This is also called sigmoid function that is used to predict probability to assign observations to categories

Logistic Function

$$\boxed{e^{\beta_0+\beta_1 X}}$$

- Taking **exponent of linear function** always returns **positive** value.
- Thus we can get rid of negative infinity to 0.

0 < $e^{\beta_0+\beta_1 X}$ **< infinity**

$$\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Dividing a number by a number slightly greater than the actual number gives value smaller than actual value.

$$0 <= \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} < 1$$

Model Evaluation - Classification Metrics

Confusion Matrix

- A **confusion matrix** is a table that is often used to describe the **performance of a classification model** on test data for which the true values are known.
- Test data size (n) = 165
  - Predicted - "yes" 110 times"no" 55 times.
  - Actual - 105 patients have the disease,

  and 60 patients do not.

predicting the presence of a disease

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Confusion Matrix

- **True positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **True negatives (TN):** We predicted no, and they don't have the disease.
- **False positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **False negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

Confusion Matrix

- True Positive Rate (TPR) (Sensitivity or Recall)- model correctly predicted the positive class.
  - True positive rate is the y-axis in an ROC curve.

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate (FPR) - model mistakenly predicted the positive class.
  - X-axis on ROC curve.

$$FPR = \frac{FP}{TN + FP}$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Confusion Matrix

- True Negative Rate (TNR) (Specificity)

$$TNR = \frac{TN}{TN + FP} = 1 - FPR$$

- False Negative Rate (FNR)

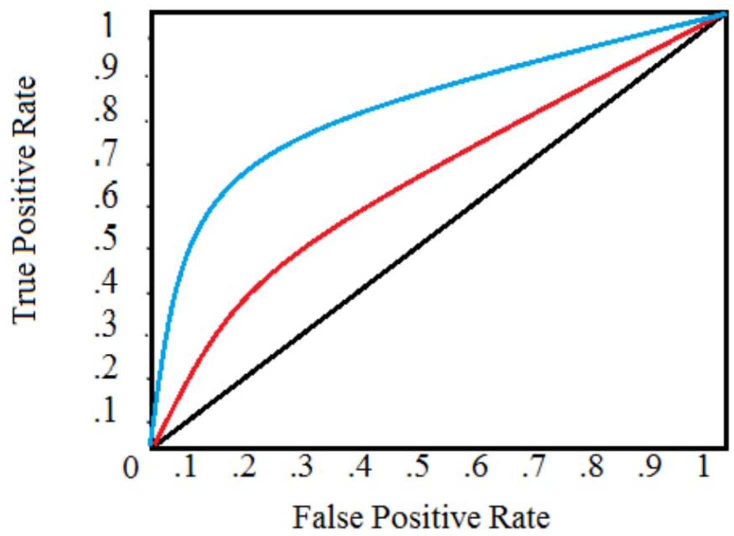
$$FNR = \frac{FN}{FN + TP} = 1 - TPR$$

- $Precision = \frac{TP}{TP+FP}$ & $Accuracy = \frac{TP+TN}{Total}$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

ROC Curve and AUC

- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
  - True Positive Rate
  - False Positive Rate



- AUC stands for 'Area under the ROC Curve'. That is, AUC measures the entire two-dimensional area underneath the entire ROC curve.
- A metric which falls between 0 and 1 with a higher number indicating better classification performance.

Classification Report

- The classification_report() function displays the precision, recall, f1-score and support for each class.
- This is a weighted average of the true positive rate (recall) and precision, where an F1 score reaches its best value at 1 and worst at 0.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$