**Logistic Regression**

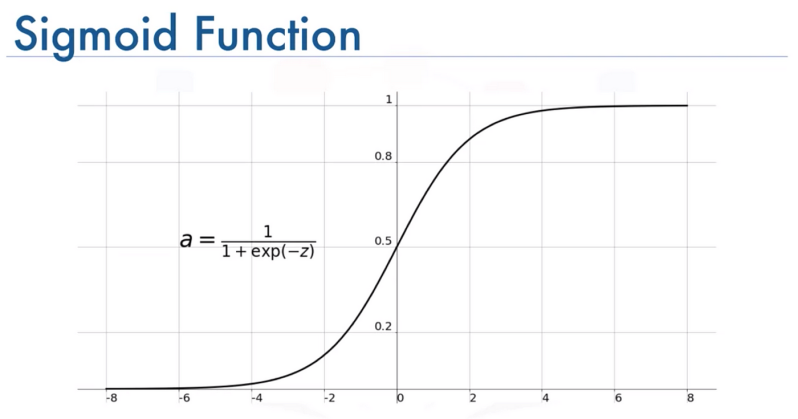
Logistic Regression is an algorithm for classification tasks. The goal here is to predict the probability.

Logistic regression is a supervised ML algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithm its name is logistic regression. It is referred to as regression because logistic regression gives a continuous value of P(Y=1) for a given input X, which is latter converted to Y=0 r Y=1 based on threshold value.

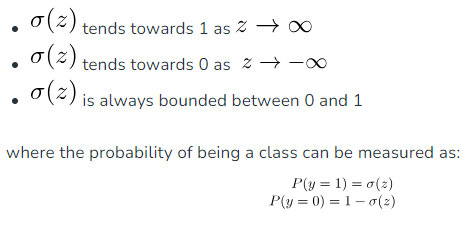
The [difference between linear regression and logistic regression](https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/) is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not. It is used for predicting the categorical dependent variable using a given set of independent variables.

**Sigmoid Function**

Now we use the [sigmoid function](https://www.geeksforgeeks.org/derivative-of-the-sigmoid-function/) where the input will be z and we find the probability between 0 and 1. i.e predicted y.

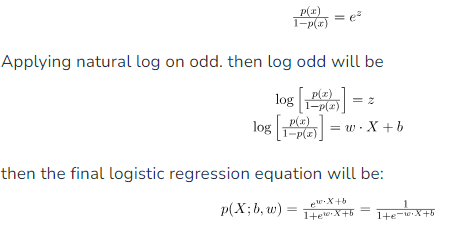


As shown above, the figure sigmoid function converts the continuous variable data into the [probability](https://www.geeksforgeeks.org/probability-gq/) i.e. between 0 and 1.



**Logistic Regression Equation**

The odd is the ratio of something occurring to something not occurring. It is different from probability as the probability is the ratio of something occurring to everything that could possibly occur. So odd will be



**Type of Logistic Regression:**

On the basis of the categories, Logistic Regression can be classified into three types:

1. Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
3. Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

**Difference between logistic regression and linear regression-**

|  |  |  |
| --- | --- | --- |
| No. | Logistic Regression | Linear Regression |
| 1 | Logistic regression is used to predict the categorical dependent variable using a given set of independent variables. | Linear regression is used to predict the continuous dependent variable using a given set of independent variables. |
| 2 | It is used for solving classification problems. | |  | | --- | | Linear regression is used for solving Regression problem. | |
| 3 | In this we predict values of categorical variables. | |  | | --- | | In this we predict the value of continuous variables. | |
| 4 | In this we find S-Curve. | In this we find best fit line. |
| 5 | Maximum likelihood estimation method is used for Estimation of accuracy. | |  | | --- | | Least square estimation method is used for estimation of accuracy. | |
| 6 | Output is must be categorical value such as 0 or 1, Yes or no, etc. | Output is must be categorical value such as 0 or 1, Yes or no, etc. |
| 7 | It not required linear relationship. | It required linear relationship between dependent and independent variables. |
| 8 | There should not be collinearity between independent variable. | |  | | --- | | There may be collinearity between the independent variables. | |

**Assumptions for Logistic Regression**

The assumptions for Logistic regression are as follows:

* Independent observations: Each observation is independent of the other. Meaning there is no correlation between any input variables.
* Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories softmax functions are used.
* Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.
* No outliers: There should be no outliers in the dataset.
* Large sample size: The sample size is sufficiently large

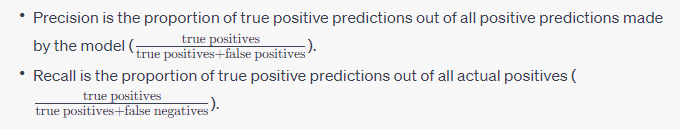
**Model evaluation metrics-**

**F1- Score-** The F1 score is a commonly used metric in binary classification tasks, such as those found in machine learning and data analysis. It is a way to combine both precision and recall into a single value, providing a more balanced assessment of a model's performance.

The F1 score is calculated using the following formula:



where:



The F1 score is the harmonic mean of precision and recall, ensuring that both values have a significant impact on the score. It ranges from 0 to 1, where 1 indicates a perfect balance between precision and recall.

In summary, the F1 score is a useful metric for evaluating the performance of a model in binary classification tasks, especially when both false positives and false negatives are important considerations.

**Precision-** Precision is a performance metric used in binary and multiclass classification tasks. It measures the proportion of true positive predictions (correctly predicted positives) out of the total predicted positives made by the model. Precision is often used in conjunction with recall to evaluate the model's ability to correctly identify positive instances.

The formula to calculate precision is:



- True Positives (TP) are the instances that were correctly predicted as positive by the model.

- False Positives (FP) are the instances that were predicted as positive but were actually negative.

Precision ranges from 0 to 1, where:

- 1 indicates that all predicted positives were correct (no false positives).

- 0 indicates that none of the predicted positives were correct (high false positive rate).

A high precision value is desirable when minimizing false positive predictions is a priority, such as in medical diagnoses or fraud detection. However, it's important to consider precision along with other metrics like recall, accuracy, and the specific context of the problem to get a comprehensive understanding of the model's performance.

**Recall-** Recall, also known as sensitivity or true positive rate, is a performance metric used in binary and multiclass classification tasks. It measures the proportion of true positive predictions (correctly predicted positives) out of the total actual positives in the dataset. Recall is important in scenarios where correctly identifying all actual positives is a priority.

The formula to calculate recall is:



- True Positives (TP) are the instances that were correctly predicted as positive by the model.

- False Negatives (FN) are the instances that were actually positive but were predicted as negative.

Recall ranges from 0 to 1, where:

- 1 indicates that the model correctly identified all actual positives (no false negatives).

- 0 indicates that the model failed to identify any actual positives (high false negative rate).

High recall is desirable when it's crucial to capture as many true positives as possible, such as in medical diagnoses or identifying rare but critical events. However, recall should be interpreted alongside other metrics like precision, accuracy, and the specific context of the problem to obtain a well-rounded assessment of the model's performance. Balancing recall and precision is often essential, and the F1 score is one metric that combines both.

**Accuracy-** Accuracy is a common performance metric used in classification tasks, including binary and multiclass classification. It measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances in the dataset.

The formula to calculate accuracy is:



In a binary classification scenario, the "Number of Correct Predictions" includes both true positives (correctly predicted positives) and true negatives (correctly predicted negatives).

Accuracy ranges from 0 to 1, where:

- 1 indicates all predictions were correct (perfect accuracy).

- 0 indicates no predictions were correct.

Accuracy is a straightforward and easy-to-understand metric that is often used to assess the overall performance of a model. However, it may not be suitable for imbalanced datasets, where one class significantly outweighs the other. In such cases, accuracy can be misleading and should be supplemented with other metrics like precision, recall, F1 score, or area under the ROC curve (AUC-ROC) to provide a more comprehensive evaluation of the model's performance.

**ROC-** ROC, which stands for Receiver Operating Characteristic, is a graphical representation of the performance of a binary classification model. It illustrates the trade-off between the true positive rate (TPR or recall) and the false positive rate (FPR) for different classification thresholds.

The true positive rate (TPR) is another term for recall or sensitivity and is calculated as:



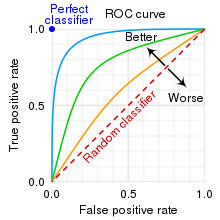
The false positive rate (FPR) is calculated as:

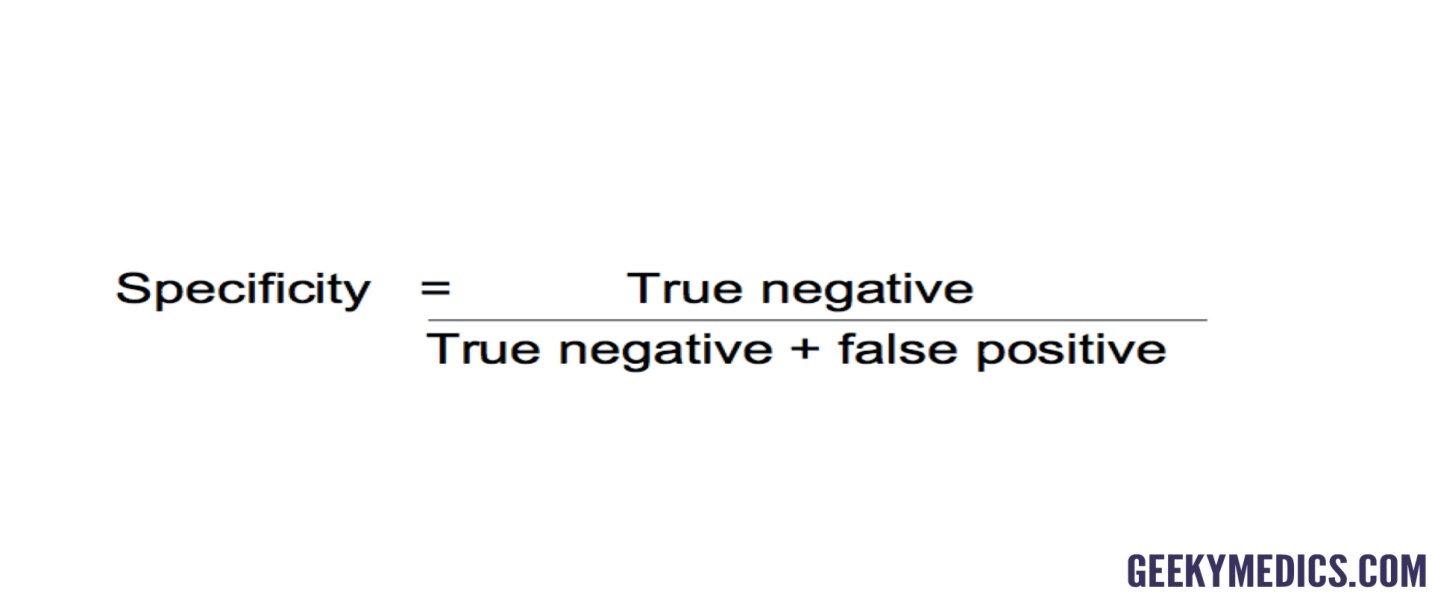


The ROC curve is created by plotting the TPR (sensitivity) on the y-axis against the FPR on the x-axis, while varying the classification threshold. Each point on the ROC curve represents a specific threshold, and moving along the curve indicates using different thresholds to classify the data.

A steeper ROC curve closer to the top-left corner of the plot indicates a better-performing model, as it signifies higher TPR (recall) for a lower FPR. The area under the ROC curve (AUC-ROC) is a single scalar value used to quantify the overall performance of the classification model. AUC-ROC ranges from 0 to 1, where 1 represents a perfect model, and 0.5 indicates a random guessing model.

In summary, the ROC curve and AUC-ROC provide a visual and quantitative assessment of how well a model can distinguish between the two classes in a binary classification task.





FPR= 1- Specificity

If the value of AUC is 0.7 then it means that there is 70% chance that the model will be able to distinguish between +ve and –ve class.