Module-2

What is Cross-Validation?

Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets,

What is cross-validation used for?

The main purpose of cross validation is to prevent overfitting, which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. By evaluating the model on multiple validation sets, cross validation provides a more realistic estimate of the model's generalization performance, i.e., its ability to perform well on new, unseen data.

Types of Cross-Validation

There are several types of cross validation techniques, including k-fold cross validation, leave-one-out cross validation, and Holdout validation, Stratified Cross-Validation.

1. Holdout Validation

In Holdout Validation, we perform training on the 50% of the given dataset and rest 50% is used for the testing purpose. It's a simple and quick way to evaluate a model. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e. higher bias.

2. LOOCV (Leave One Out Cross Validation)

In this method, we perform training on the whole dataset but leaves only one data-point of the available dataset and then iterates for each data-point. In LOOCV, the model is trained on n−1samples and tested on the one omitted sample, repeating this process for each data point in the dataset. It has some advantages as well as disadvantages also.

An advantage of using this method is that we make use of all data points and hence it is low bias.
The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point.

3. **Stratified Cross-Validation**

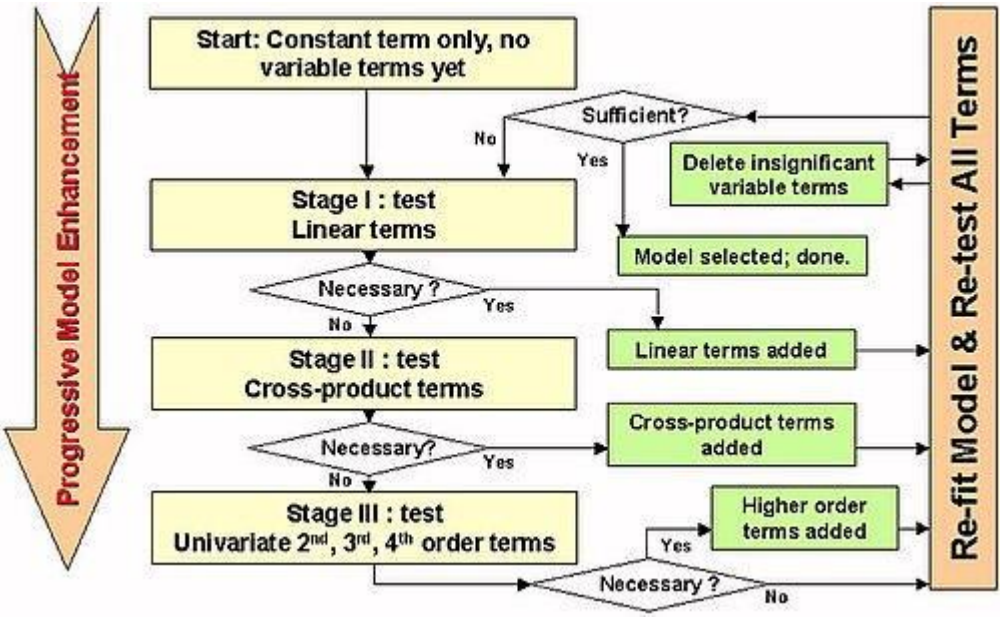
It is a technique used in machine learning to ensure that each fold of the cross-validation process maintains the same class distribution as the entire dataset. This is particularly important when dealing with imbalanced datasets, where certain classes may be underrepresented. In this method,
The dataset is divided into k folds while maintaining the proportion of classes in each fold.
During each iteration, one-fold is used for testing, and the remaining folds are used for training.
The process is repeated k times, with each fold serving as the test set exactly once.

4. K-Fold Cross Validation

In K-Fold Cross Validation, we split the dataset into k number of subsets (known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

Stepwise Regression:

Stepwise regression is a method of fitting a regression model by iteratively adding or removing variables. It is used to build a model that is accurate



There are two main types of stepwise regression:

Forward Selection – In forward selection, the algorithm starts with an empty model and iteratively adds variables to the model until no further improvement is made.

Backward Elimination – In backward elimination, the algorithm starts with a model that includes all variables and iteratively removes variables until no further improvement is made.

Use of Stepwise Regression?

The primary use of stepwise regression is to build a regression model that is accurate and parsimonious. In other words, it is used to find the smallest number of variables that can explain the data. Stepwise regression is a popular method for model selection because it can automatically select the most important variables for the model and build a parsimonious model. This can save time and effort for the data scientist or analyst, who does not have to manually select the variables for the model.

Q5) Explain Regression and types of Regression.

Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated.

• Regression analysis is used to predict or estimate one variable in terms of the other variable.

It is useful in statistical estimation of demand curves, supply curves, production function, cost function etc.

1.Simple and multiple regression

• Simple regression : The regression analysis for studying only two variables at a time is known as simple regression.

• Multiple regression : The regression analysis for studying more than two variables at a time is known as multiple regression.

2. Linear and nonlinear regression

• Linear regression : If the regression curve is straight line, then the regression is said to be linear

. • Nonlinear regression : If the regression curve is not a straight line i.e. not a first-degree equation in the variables X and y, the regression is said to be nonlinear regression.

Q6 What is regression? What is simple linear regression?

Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated.

Regression analysis is used to predict or estimate one variable in terms of the other variable.

It is useful in statistical estimation of demand curves, supply curves, production function, cost function etc.

Simple linear regression is a statistical method used to model the relationship between a single independent variable (predictor variable) and a single dependent variable (response variable). It assumes that the relationship between the variables can be approximated by a straight line.

The equation for simple linear regression can be expressed as:

$y=mx+b$

Where:

y is the dependent variable (response variable),

x is the independent variable (predictor variable),

m is the slope of the line (the change in y for a unit change in x),

b is the y-intercept (the value of y when x is 0).

Fitted Values (Y^)

A **fitted value** (also called a predicted value) is the estimated value of the dependent variable (YYY) based on the regression equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- It represents the point on the regression line corresponding to a given X.
- The fitted value is the model's best guess for Y based on the independent variable X

Example:

Suppose we have a regression equation for predicting exam scores:

$$\hat{Y} = 50 + 5X$$

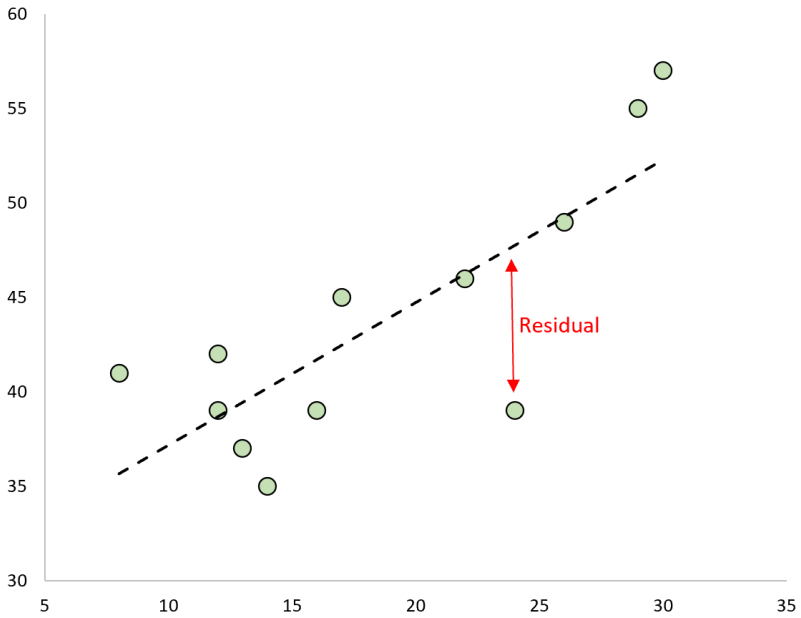
If a student studies for **4 hours**, the predicted exam score is:

$$\hat{Y} = 50 + 5(4) = 70$$

So, **70** is the fitted value for **4 hours** of study.

Residuals (ε)

A **residual** is the difference between the actual observed value (Y) and the fitted (predicted) value (Y ^):
Residual=Y−Y^{Residual} = Y−Y^



- Residuals measure how far the actual values are from the regression line.

- A **small residual** means the model made a good prediction, while a **large residual** indicates poor prediction.

  Example:

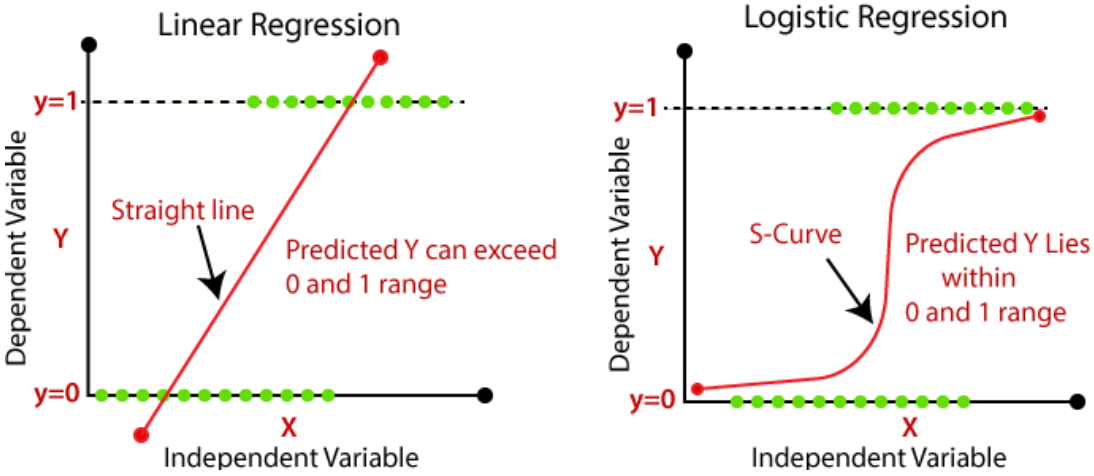
  If a student actually scored **75** after studying for **4 hours**, the residual would be:

  $$\text{Residual} = Y - \hat{Y} = 75 - 70 = 5$$

  Here, the model **underestimated** the score by **5 points**.

- **Residual Analysis**
- Residuals help in diagnosing the quality of a regression model:
  ☑ **If residuals are randomly scattered**, the model is appropriate.
  ❌ **If residuals show patterns**, the model may be missing key relationships.
- Residuals are often visualized using a **residual plot**, where:
- The **X-axis** represents the independent variable or fitted values.
- The **Y-axis** represents residuals (Y−Y^Y - \hat{Y}Y−Y^).
- A random scatter suggests a good fit, while patterns suggest model issues.
-

Q1.What are the similarities and differences between linear regression and Logistic Regression?

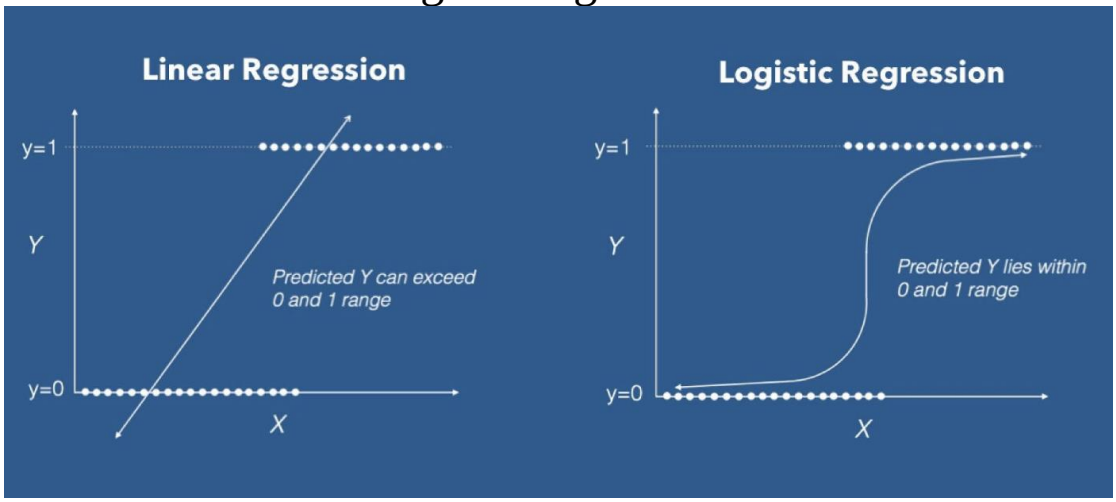


Linear regression is used to predict the continuous dependent variable using a given set of independent variables. Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables. Linear Regression is used for solving Regression problem.

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.

Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

How are linear and logistic regression similar?



Similarities between Logistic and Linear regression:

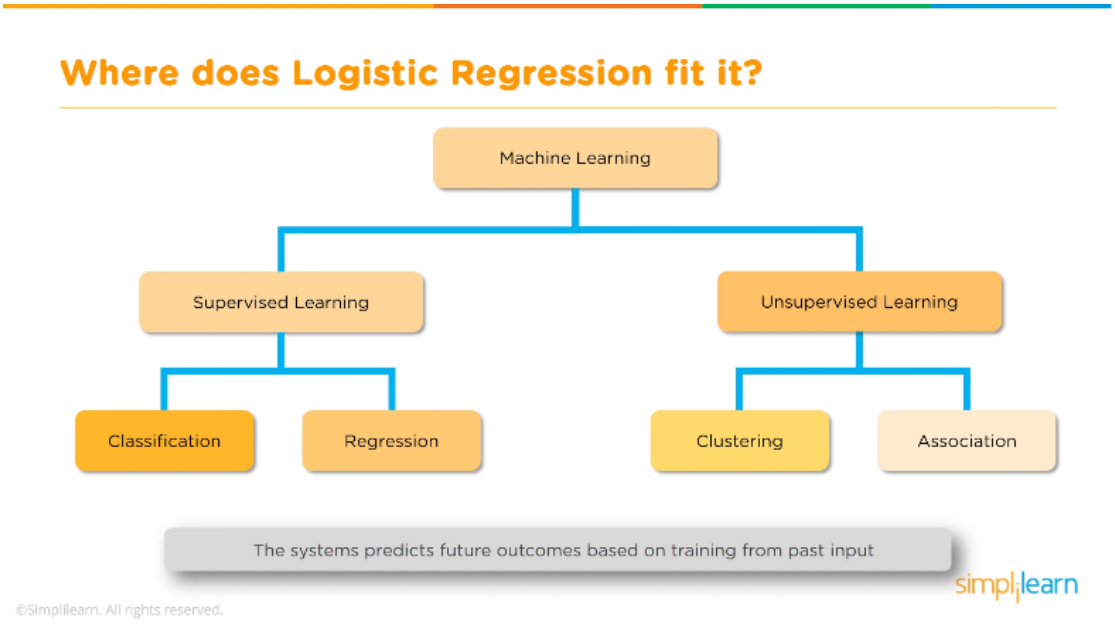
They are both parametric Regressions, and both utilize a linear

equation to arrive at predictions. However, the similarities end there. In Linear regression the result is continuous. In Logistic Regression, there are only a limited number of possible values.

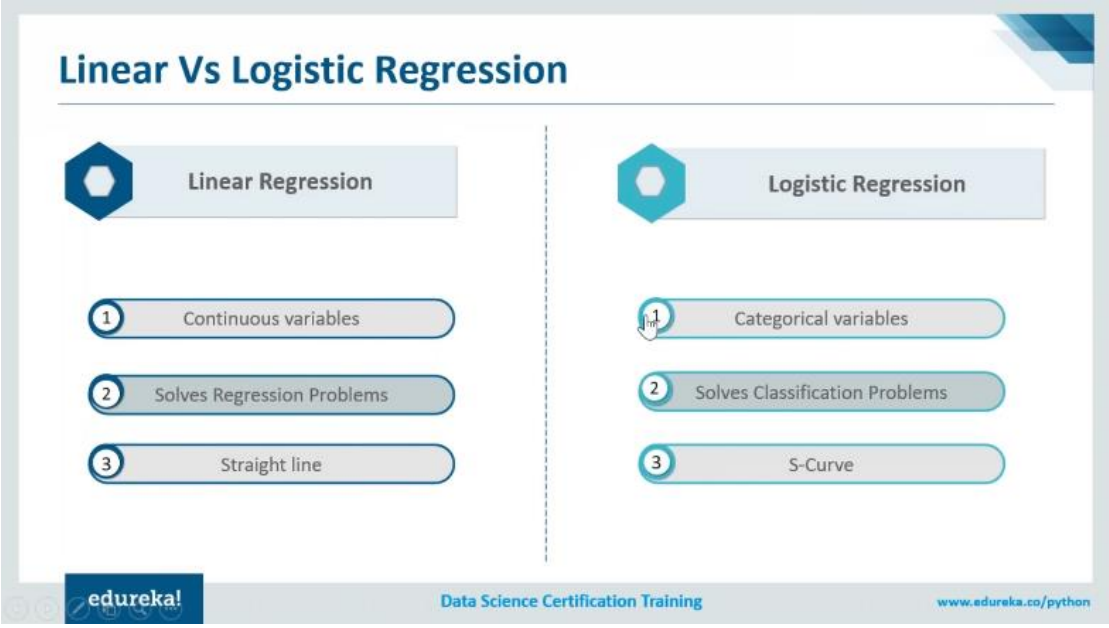
Other hand:

Logistic Regression

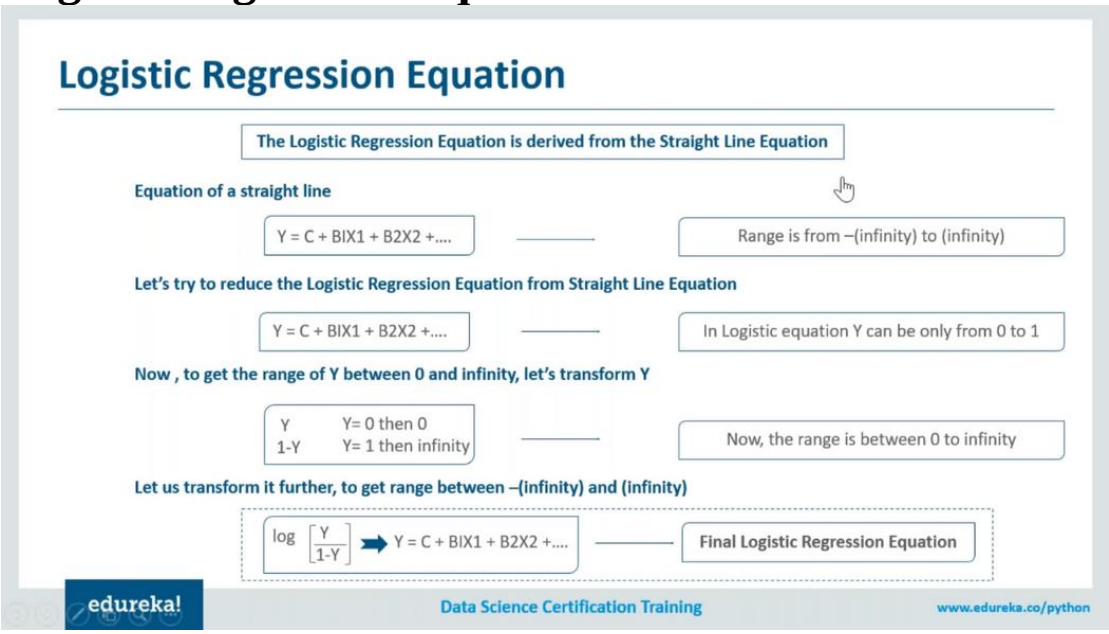
- It is a predictive modeling technique.
- It estimates the relationship between dependent and independent variables.
- It is used when the output is in binary format(discrete or in categorical format) whereas linear regression is used when the output is continuous in range.
- Logistic regression curve converts any value to discrete values.
- There is a threshold that segregates the output as positive or negative(0 or 1).
- It comes under the category of Supervised Learning under classification.



Linear vs Logistic Regression



Logistic Regression Equation



Use-Cases:

1. Weather Prediction(will it rain or not)
2. Classification Problem(bird or not a bird)
3. Determines Illness(patient is ill or not)

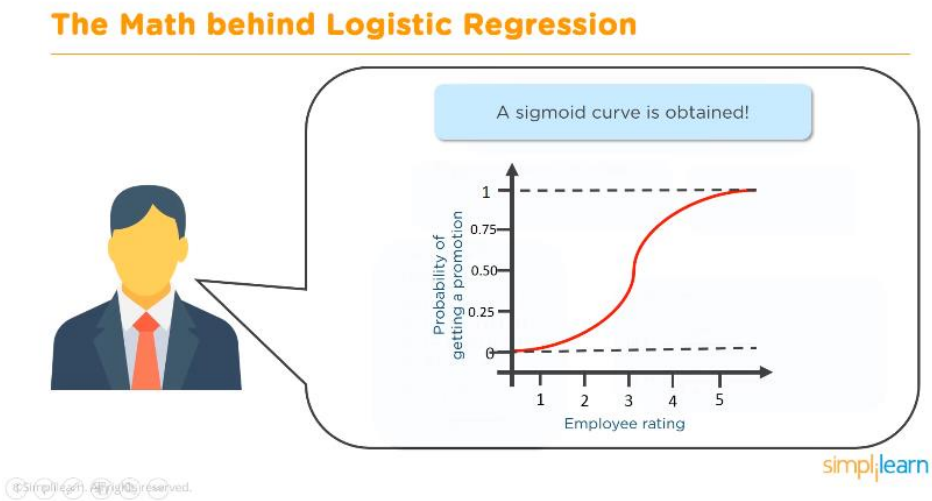
Steps

1. Teaching the model with the dataset

2. Dropping the non-essential components
3. Determining the output and evaluating the model

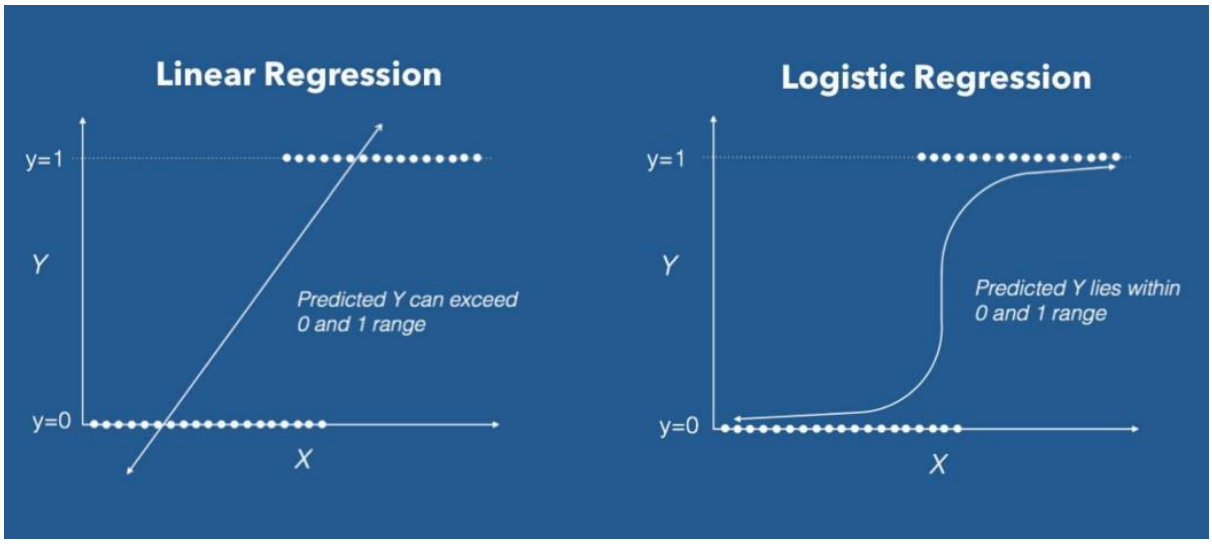
Sigmoid Curve

A threshold has to be set above which the value is 1 and below it, the value is 0.



Logistic regression:

Logistic regression **estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables**. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.



We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

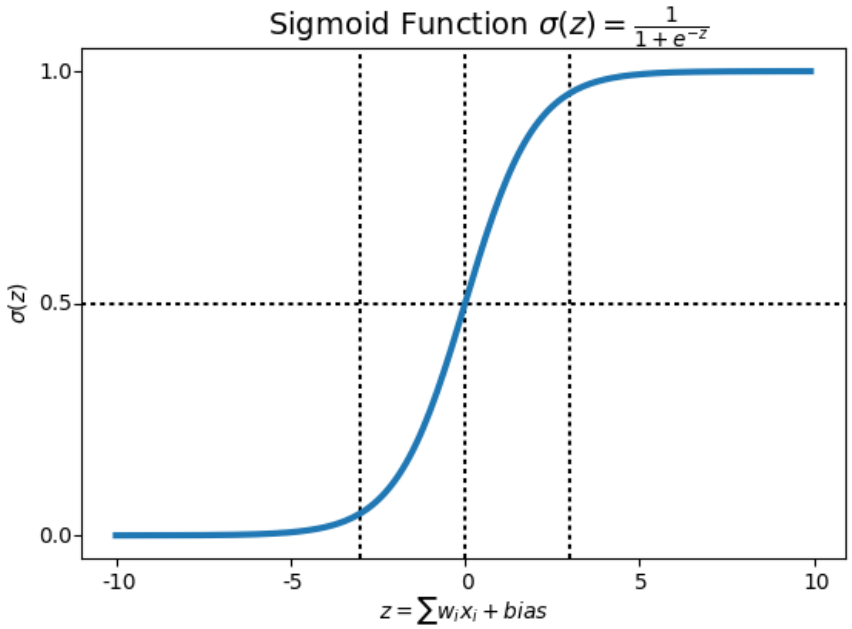
The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

What is the Sigmoid Function?

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



Sigmoid Function Graph

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

Formula of a sigmoid function |

Logit:

What logit means?

log-odds function:

What is a Logit? A Logit function, also known as the log-odds function, is **a function that represents probability values from 0 to 1, and negative infinity to infinity**. The function is an inverse to the sigmoid function that limits values between 0 and 1 across the Y-axis, rather than the X-axis.

Logit:

What logit means?

Odds are basically the ratio of some event happening to some event not happening. It can also be defined as the ratio of the probability of an event happening to the Probability of the event not happening. Odds can be expressed as a Ratio or a Fraction.

Consider a team that played 100 matches and won 25 of them and lost 75 of them. Now we can calculate the Odds and Probabilities as follows,

We can say that the Odds in favor of the team winning are 1:3 or 1/3 or 0.333. Since we have odds in favor of the team winning,

P(A)/P(-A),

where P(A) is the probability of A, and P(-A) the probability of 'not A' (i.e. the complement of A).

Taking the logarithm of the odds ratio gives us the log odds of A, which can be written as

log(A) = log(P(A)/P(-A)).

Since the probability of an event happening, P(-A) is equal to the probability of an event not happening, 1 – P(A), we can write the log odds as

log [p/(1-p)]

- p = the probability of an event happening
- 1 – p = the probability of an event not happening

Using Log Odds

reason to use log odds is that it is usually difficult to model variables with restricted ranges, such as probabilities.

Example:

Consider the example of Smoking and its effects on Lung cancer, if we are to form a two by two table showing the effects of smokers and non-smokers in causing lung cancer, the table would look something like this,

Cancer	Non-Cancer	Totals	
Smoker	100	60	160
Non-Smoker	34	125	159
Totals	134	185	319

Out of the 319 patients,

Out of the 319 patients,

160 are smokers and 159 are non-smokers.

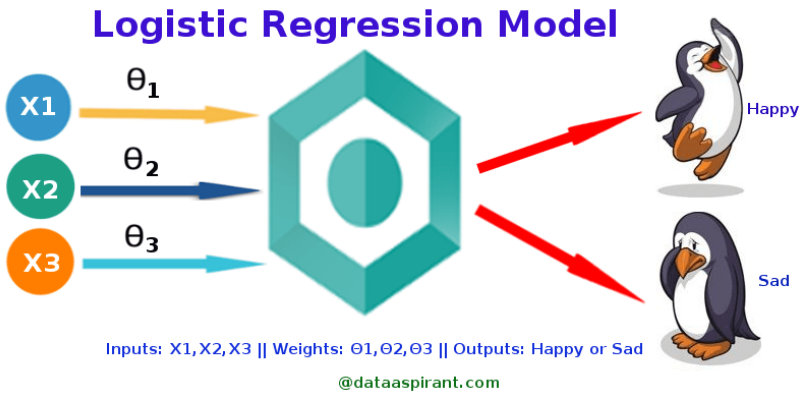
134 have cancer and 185 do not have cancer.

100 smoke and have cancer, 60 smoke and do not have cancer.

34 do not smoke and have cancer and 125 do not smoke and do not have cancer.

From the above information, if you want to calculate the Odds Ration, you just have to cross multiply and take the ratio.

$$\text{Odds ratio} = \frac{100*125}{34*60} = 6.127$$



Types of Logistic Regression

Binary logistic regression

Binary logistic regression is used to predict the probability of a binary outcome, such as yes or no, true or false, or 0 or 1. For example, it could be used to predict whether a customer will churn or not, whether a patient has a disease or not, or whether a loan will be repaid or not.

Multinomial logistic regression

Multinomial logistic regression is used to predict the probability of one of three or more possible outcomes, such as the type of product a customer will buy, the rating a customer will give a product, or the political party a person will vote for.

Ordinal logistic regression

is used to predict the probability of an outcome that falls into a predetermined order, such as the level of customer satisfaction, the severity of a disease, or the stage of cancer. The odds ratio is the probability of success/probability of failure. As an equation, that's

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 Decision Boundary To predict which class a data belongs; a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes. Say, if predicted value ≥ 0.5, then classify email as spam else as not spam. Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary.

Cost Function:

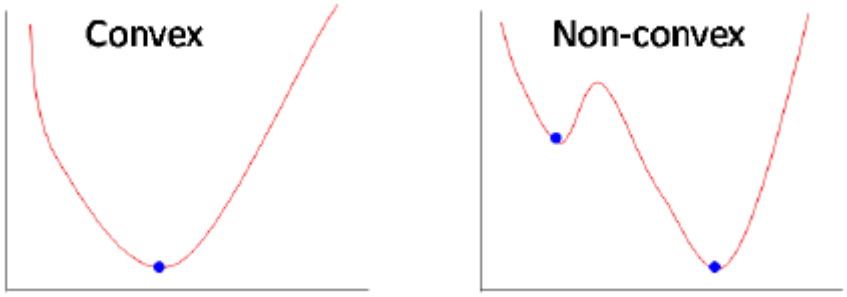
$$\text{Cost}(h_\Theta(x), Y(\text{actual})) = -\log(h_\Theta(x)) \text{ if } y=1$$

$$-\log(1- h_\Theta(x)) \text{ if } y=0$$

Cost Function of Logistic Regression

Why cost function which has been used for linear can not be used for logistic?

Linear regression uses mean squared error as its cost function. If this is used for logistic regression, then it will be a non-convex function of parameters (theta). Gradient descent will converge into global minimum only if the function is convex.



· In logistic regression, the function is convex, and gradient descent can reliably find the global minimum.
· In non-convex functions (e.g., deep learning models), gradient descent may only find a local minimum or saddle point, depending on the initialization and optimization path. A saddle point is a point on the surface of a function where the gradient (slope) is zero, but it is neither a maximum nor a minimum.

GLM:

The logistic regression model is an example of a broad class of models known as generalized linear models (GLM). For example, GLMs also include linear regression, ANOVA(is a statistical method used to compare the means of three or more groups to determine if there is a statistically significant difference between them.), poisson regression, (Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables.)etc.
There are three components to a GLM:

- ***Random Component*** – refers to the probability distribution of the response variable (Y); e.g. binomial distribution for Y in the binary logistic regression.
- ***Systematic Component*** - refers to the explanatory variables (X_1, X_2, ... X_k) as a combination of linear predictors; e.g. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ as we have seen in logistic regression.
- ***Link Function,*** η *or* $g(\mu)$ - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g. $\boldsymbol{\eta = logit(\pi)}$ **for logistic regression.**

Some of the features of GLMs include:

Flexibility: GLMs can model a wide range of relationships between the response and predictor variables, including linear, logistic, Poisson, and exponential relationships**.**

Model interpretability: GLMs provide a clear interpretation of the relationship between the response and predictor variables, as well as the effect of each predictor on the response.

Robustness: GLMs can be robust to outliers and other anomalies in the data, as they allow for non-normal distributions of the response variable.

Scalability: GLMs can be used for large datasets and complex models, as they have efficient algorithms for model fitting and prediction.

***Ease of* use:** GLMs are relatively easy to understand and use, especially compared to more complex models such as neural networks or decision trees.

Hypothesis testing: GLMs allow for hypothesis testing and statistical inference, which can be useful in many applications where it's important to understand the significance of relationships between variables.

GLMs can be used to construct the models for regression and classification problems by using the type of distribution which best describes the data or labels given for training the model.
(The term data specified here refers to the output data or the labels of the dataset).
Binary classification data **– Bernoulli distribution**
Real valued data **– Gaussian distribution**
Count-data **– Poisson distribution.**
Systematic Component (Linear Predictor)

- The model assumes a linear combination of independent (predictor) variables: $\eta = X\beta$

  where:

  - X = Design matrix of independent variables
  - β = Coefficients (parameters) of the model
  - η = Linear predictor (unconstrained value before transformation)

 where g is the link function and μ is the mean of the response variable.

$$g(\mu)=\eta$$

 Common link functions:

- **Identity Link:** $g(\mu) = \mu$ (Used in linear regression)
- **Logit Link:** $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ (Used in logistic regression)
- **Log Link:** $g(\mu) = \log(\mu)$ (Used in Poisson regression)

Advantages of GLMs

✅ Handles different types of response variables
✅ Provides flexibility with different link functions
✅ Works well with categorical and continuous predictors
✅ Uses Maximum Likelihood Estimation (MLE) for parameter estimation

Accessing the models:

1. Training the Model
To access a logistic regression model, it must first be trained on a dataset. This involves:
Input data: A dataset with features (X) and labels (y).
Training process: Fitting the model to learn the relationship between the features and the target.
2. Accessing the Model
Once trained, the logistic regression model provides several attributes and methods:
Model Coefficients and Intercept
model.coef_: The weights (coefficients) associated with the features.
model.intercept_: The bias (intercept) term.
Making Predictions
model.predict(X): Predicts the class labels (0 or 1) for input data.
model.predict_proba(X): Returns the probabilities for each class.
Model Performance
Use metrics such as accuracy, precision, recall, and ROC-AUC to evaluate performance.
Libraries like scikit-learn provide tools for this:
3. Visualizing the Model
You can plot the decision boundary, especially if working with 2D data.
For example, using Matplotlib:
4. Interpreting the Model
Coefficients: Show how changes in each feature influence the odds of the outcome.
Odds ratio: Can be derived from the coefficients:
Odds ratio=e^coefficient
5. Advanced Usage
Regularization: Control overfitting using parameters like C in LogisticRegression.
Multiclass Classification: Logistic regression can be extended for multiclass problems using the "one-vs-rest" or &" multinomial" approach.