**Logistic Regression in Machine Learning**

Anish Mahapatra  
anishmahapatra01@gmail.com

Table of Contents

[What is logistic regression? 3](#_Toc95666567)

[How is it useful in the real world? 3](#_Toc95666568)

[The math behind logistic regression 3](#_Toc95666569)

[Assumptions of Logistic Regression 4](#_Toc95666570)

[How to assess a logistic regression model? 5](#_Toc95666571)

[Frequently Asked Questions (FAQs) at interviews 6](#_Toc95666572)

[Q. What is the difference between Linear and Logistic Regression? 6](#_Toc95666573)

[Q. 6](#_Toc95666574)

List of Figures

[Figure 1: Visual representation of Sigmoid / Logit function (Source) 4](#_Toc95666575)

[Figure 2 Measuring performance using the Area under ROC Curve (Source) 5](#_Toc95666576)

**Logistic Regression in Machine Learning**

Welcome to a series of blogs in Machine Learning that you will *actually understand*. When there are so many blogs out there on this machine learning, why would we take the time to write another one?

It’s because all the blogs you’ve read so far are either code intensive, math heavy or have excessive theory. Our attempt through this series of blogs is to help you learn in the most efficient way possible – I will put in just enough theory to help you in technical interviews, just enough code in a way that you can understand it easily and just enough math for you to explain in an interview. Let’s get cracking – this is a series that is worth your time – pay attention and follow along.

# What is logistic regression?

Logistic regression is a popular supervised machine learning models that is used to predict the probability that an observation belongs to one of two possible classes.

Now, this statement comes loaded with new terminology that we need understand. So, let’s do that.

* **Supervised Learning**   
  Machine learning models that learn and predict an output based on example input-output pairs are known as supervised models.
* **Classification**  
  Given an observation, classification is the task of being able to identify which group the observation most likely belongs to. For instance, marking a mail as spam or not is a classification task.

# How is it useful in the real world?

Logistic regression is really popular in the real world and one algorithms that a lot of interviewers will grill you on. Let me give you a few examples where logistic regression can be used:

* Employee Attrition - To determine if an employee will leave the company or not
* Customer Churn – To understand if a customer is going to leave
* Tumor Classification – Analysing radiological images to predict if a tumor is *malignant or benign*
* Loan Default – To determine if a customer will default on a bank loan

To understand how all of this is done, it is critical to understand the mathematics behind this as your interviewer will most likely ask.

# The math behind logistic regression

Logistic regression is based on the logit function. Let’s go ahead and derive the logit function.   
  
Odds in probability is defined as the probability that the event will occur divided by the odds that an event will not occur.

Here, we need a range from (, this can be done by taking a natural logarithm on both sides:

Let’s now take the natural exponential on both sides,

Let’s define . Work through the equation to get on one side and, we get the following:

This is the sigmoid function, also called the logit function.

Now, why is the sigmoid function relevant for us?

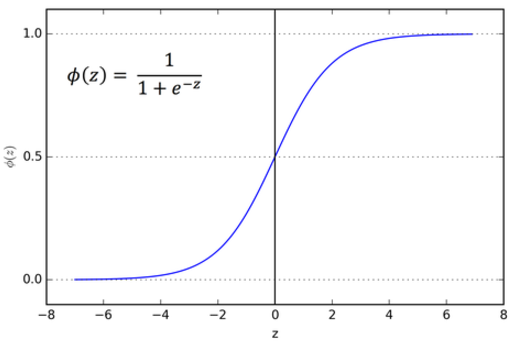


Figure : Visual representation of Sigmoid / Logit function ([Source](https://ai-master.gitbooks.io/logistic-regression/content/sigmoid-function.html))

This is the function that ranges between 0 and 1. It essentially restricts the entire range of probabilities to two values. Based on our initial definition, this is exactly what we are looking for, a function that can help determine if an observation belongs to one of two possible classes.

# Assumptions of Logistic Regression

Let us go through the assumptions of logistic regression.

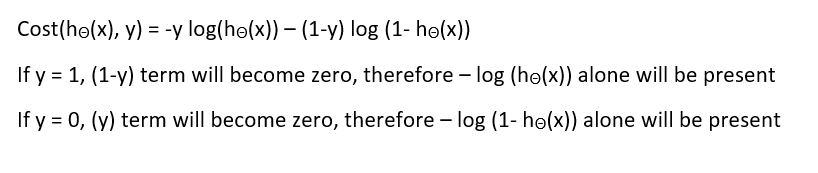
* Logistic regression assumes that the observations are independent
* The independent variables are linearly related to the log of odds for logistic regression
* It assumes that there is **minimal or no multi-collinearity** between the *independent variables* – this is generally validated using Variance Inflation Factor (VIF) that should lie between 0 to 20
* A good prediction is reliant on a large sample size
* Logistic regression assumes that there are no outliers

Cost function of logistic regression

The way we measure the performance of machine learning models is by something called a cost function.

Cost function is defined as the calculation of error between the predicted values and the expected values for a given machine learning model

The cost function for is defined as for logistic regression.



# How to assess a logistic regression model?

Let’s say that we are in the stage where the model is made, now we need to understand if it is doing the job we set out to accomplish. This can be assessed using evaluation metrics. Following are the different ways that we can assess a logistic regression model:

* **Accuracy**   
  Of all the model predictions, what percentage of samples have been correctly classified.
* **ROC-AUC**   
  Area under the Receiver Operating Characteristics (ROC-AUC) is defined as the relationship between True Positive Rate (TPR) and False Positive Rate (FPR). This sounds like a fancy metric, you can read up more on the metric, but to get more intuition on what it means, look at the figure 2 below.

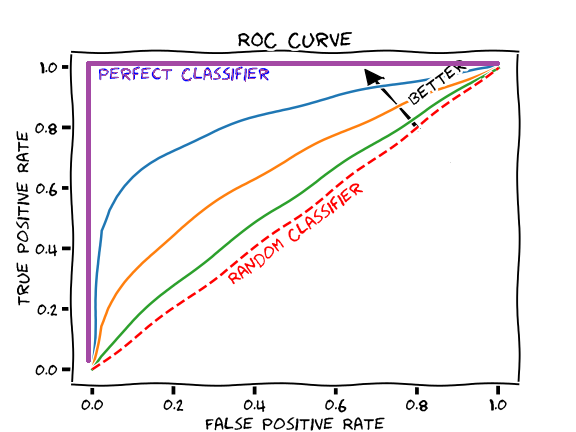
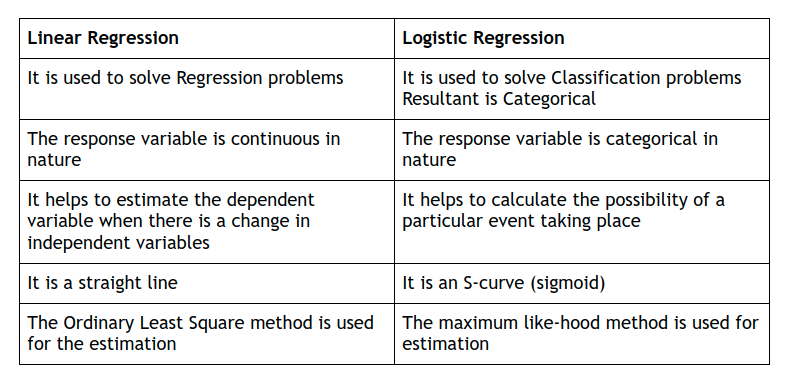


Figure Measuring performance using the Area under ROC Curve ([Source](https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/))

# Frequently Asked Questions (FAQs) at interviews

## Q. What is the difference between Logistic Regression and Linear Regression?

A.

## Q. What is logistic regression?

A. Logistic regression is a supervised machine learning model that is used for classification (generally binary), where the target variable is categorical. The model is based on the sigmoid or logit function, where based on the threshold (generally 0.5), the result is limited between 0 and 1.

Logistic regression is a commonly used machine learning in production as it is easy to use, highly interpretable, easily scalable and can effectively perform real-time predictions.

Some classic example of logistic regression is:

* Is an email spam or not?
* Will a customer churn or not from a company?
* If a student is likely to pass or fail in an exam?
* If a customer is likely to default on a loan

## Q. Does logistic regression have a linear or non-linear decision boundary?

A. A decision boundary is defined as the line or plane that separates the observations into different classes. This is defined as a straight line for the case of logistic regression. This makes logistic regression a good model to use when it is possible to separate the observations using a straight line.

Logistic regression has a linear decision boundary.

## Q. Do outliers impact logistic regression model?

A. Yes, logistic regression is impacted by outliers. The sigmoid function can take care of outliers to an extent, but if there are extreme outliers, it can affect the performance of a logistic regression model. So, it is recommended to handle outliers before training the model.

## Q. What is the range of the logistic regression function?

A. The output of the logistic regression model is logits. The logistic function, defined as , has a range from 0 to 1, with the standard threshold being 0.5

## Q. We say that the logistic regression model can predict categorical values. Does a logistic regression truly handle categorical values? Elaborate.

A. The logistic regression model as such is unable to process categorical values. To enable the algorithm to process categorical values, we perform a process called one-hot encoding, where each level of the categorical values is assigned a “dummy numerical variable”.



Figure : Using categorical values with the help of One-Hot Encoding ([Source](https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding))

## Q. Can logistic regression help solve multiclass classification problems?

A. Logistic regression is designed to solve binary classification problems. However, **it is possible** to solve multiclass classification problems using logistic regression. It is done using the **one-vs-all** Classification.

The idea is to build each individual class a personalized logistic regression model, and then decide if an observation belongs to the class or not. Then, we will run the models on each observation, and whichever observation has a higher probability will belong to that class.

For instance,

Probability (y == Class A) = 0.2   
Probability (y == Class B) = 0.7   
Probability (y == Class C) = 0.1

With the one-vs-all approach, the observation under consideration above would be assigned to Class B.

## Q. A classic question – If logistic regression is used for classification, why is it not called logistic **classification**?

A. Regression has continuous / numerical outputs and classification problems have categorical / discrete outputs. Similar to linear regression, logistic regression also build a model that outputs the probability (between 0 and 1) that an observation belongs to class “1”.

The main difference is that using the sigmoid function, it limits the range between 0 and 1, and that’s why it is used for classification. The name “Logistic” is derived from the logit function that it used for this.

## Q. If logistic regression is like linear regression, why can’t we just use the linear regression to perform binary classification?

A. This is a good question. To recall, the **assumptions of linear regression** are:

* **Linear relationship**: The dependent and independent variable should have a linear relationship
* **Residuals Homoscedasticity**: The residuals should have a constant variance
* **Independence of Residuals**: The residuals should be independent of random variables
* **Normality**: The residuals should be normally distributed

Residual is the difference between the observed value and predicted value.

So, to answer the question, linear regression cannot be used for binary classification for the following reasons:

1. **Variance of error terms**: The error terms for binary classification is not constant
2. **Error term distribution**: The error terms are not linearly distributed in the case of binary classification
3. **Different model outputs**: Linear regression is used to give a continuous and numerical output, whereas logistic regression gives us a discrete and categorical output.