Artificial Intelligence Masterclass

Logistic Regression & Supervised Learning Techniques

H.M. Samadhi Chathuranga Rathnayake

M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Qualitative Responses

What about the response variable is a categorical (Qualitative) variable?

Ex:-

- Yes/ No
- True/ False
- Good/ Bad
- Pass/ Fail

Dichotomous Responses/ Binary Responses (Responses with two classes)

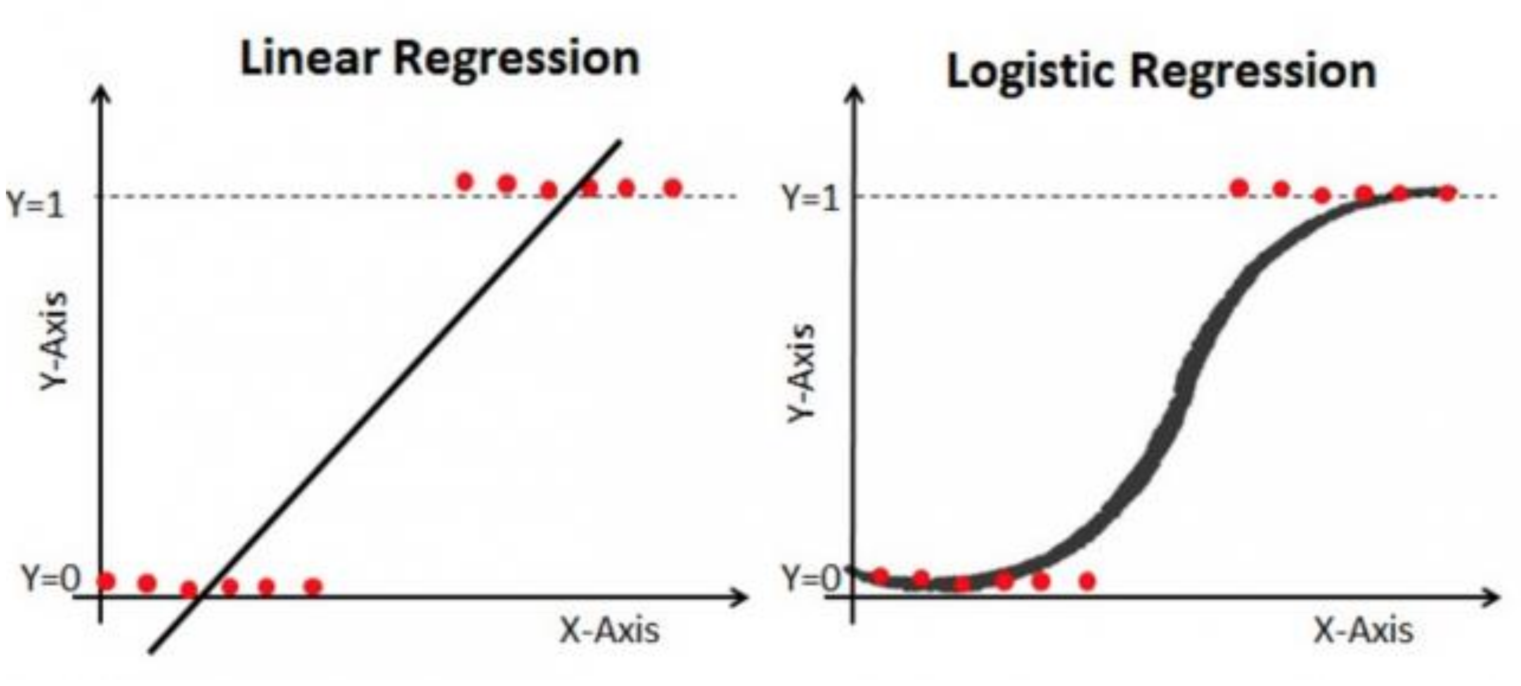
- High/ Medium/ Low
- Class 01/ Class 02/ Class 03/ Class 04
- Red/ Green/ Blue

etc.

In these situations, our goal is to classify observations to these groups. This is called as the **Classification.**

Logistic Regression

Logistic regression is used when we have a dichotomous variable (Two levels categorical variable Ex- Yes/ No) as the response variable. In that case we cannot use linear regression for the predictions.



Here what we can calculate as the output of the model is the probability of belonging to the category.

Logistic Regression

Then we cannot work with this model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p$. We have to convert this mode such that the output is positive as well as the output should be lying between 0 and 1. The we have to deal with the Sigmoid function.

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

So we put above model inside to the sigmoid function and satisfy above discussed criteria.

$$P = S(y) = \frac{1}{1 + e^{-y}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p)}}$$

So,

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p)}}$$

Then we can show that,

$$log\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p$$

Logistic Regression

Here $log\left(\frac{P}{1-P}\right)$ is called as Logit. Parameters of the model will be estimated using Maximum Likelihood Estimation.

When predicting some variable using this model a score will be given through this model and the probability should be found and then with that probability according to some cutoff value, the prediction will be done.

Consider the example,

$$P \geq Some\ cutoff\ \ --\rightarrow Assign\ to\ class\ 01$$

$$P < Some\ cutoff\ \ --\rightarrow Assign\ to\ class\ 02$$

Generally this cutoff values is 0.5 in most cases. But this can be changed according to the situation.

Model Evaluation

The mainly discussed two techniques for evaluating a model in regression can be used for evaluating in the classification too.

- Validation Set Approach
- Cross Validation

Using these approaches we can measure some metrics for evaluating a model. The most popular metrics among these metrics is,

$$Miss\ Classificatio\ Error\ (MCE)$$

$$Accuracy = 1 - MCE$$

$$F1\ Score$$

Validation Set Approach

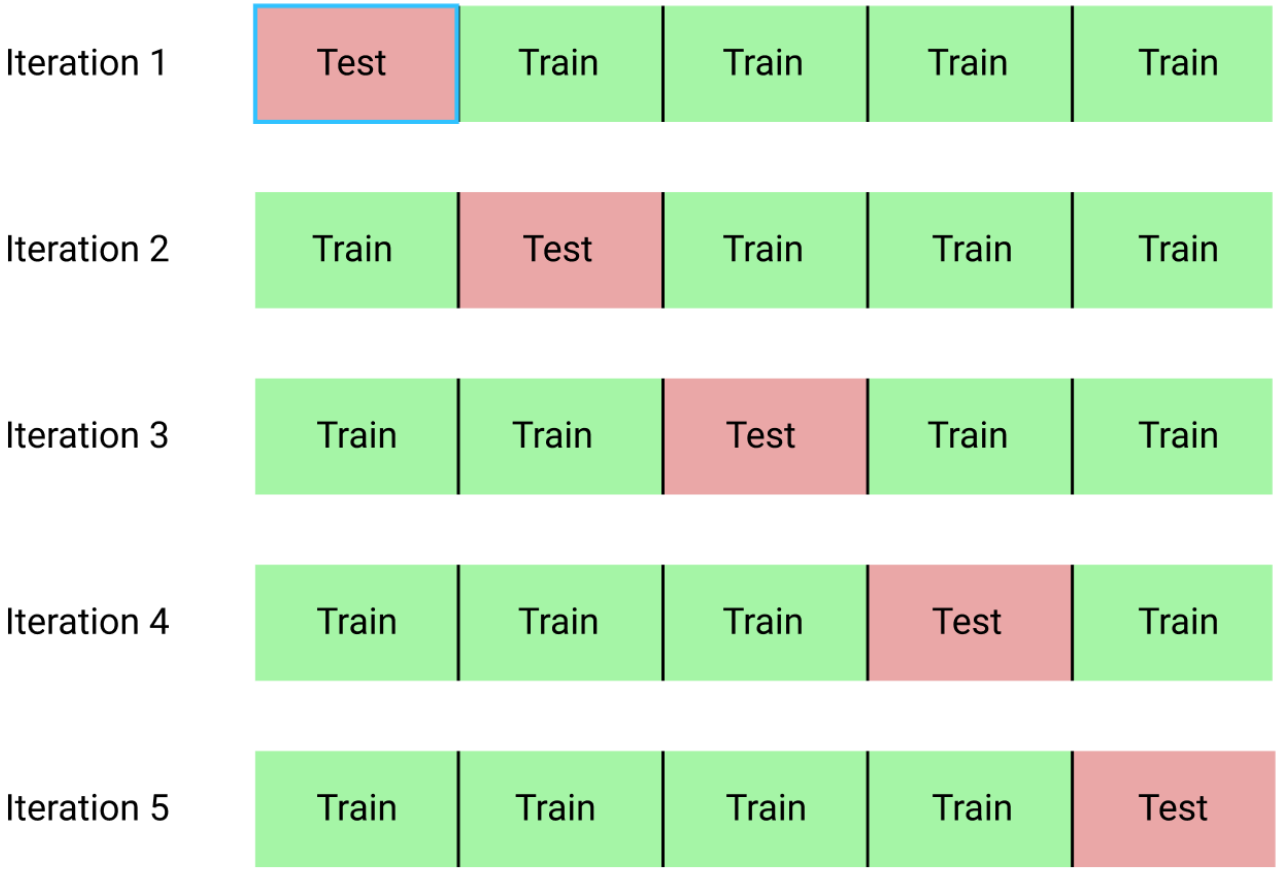
Split the dataset into two sets,

- Training dataset (Generally 80% of the data But it can be changed)
- Testing dataset (Rest of the data)

Train the model using the training dataset and then check the **Accuracy** & the **Miss Classification Error** of the model using the testing dataset. Above mentioned metrics of the model will be calculated and the model will be evaluated.

Cross Validation Approach

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.



Confusion Matrix

For evaluating the above discussed metrics, a special tool is used which is called as Confusion Matrix.

Consider the following example where we have to predict a categorical variable with two levels; Yes & No. Consider that we have predicted 100 testing observations using a trained model. Then the confusion matrix can be obtained as,

	Actual Output		
Predicted Output		Yes	No
	Yes	40	10
	N0	20	30

Confusion Matrix

Correctly classified observations.

	Actual Output	
Predicted Output		Yes	No
	Yes	40	10
	N0	20	30

$$Accuracy = \frac{40 + 30}{100} \times 100\% = 70\%$$

Confusion Matrix

Incorrectly classified observations

	Actual Output	
Predicted Output		Yes	No
	Yes	40	10
	N0	20	30

$$MCE = \frac{10 + 20}{100} \times 100\% = 30\%$$

Now we can see that,

$$Acuracy = 1\ - MCE$$

Confusion Matrix

	Actual Output		
Predicted Output		Yes	No
	Yes	True Positives (TP)	False Positives (FP)
	No	False Negatives (FN)	True Negatives (TN)

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Confusion Matrix

	Actual Output		
Predicted Output		Yes	No
	Yes	True Positives (TP)	False Positives (FP)
	No	False Negatives (FN)	True Negatives (TN)

TPR (True Positive Rate) / Recall /Sensitivity

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

Specificity

$$\text{Specificity} = \frac{TN}{TN + FP}$$

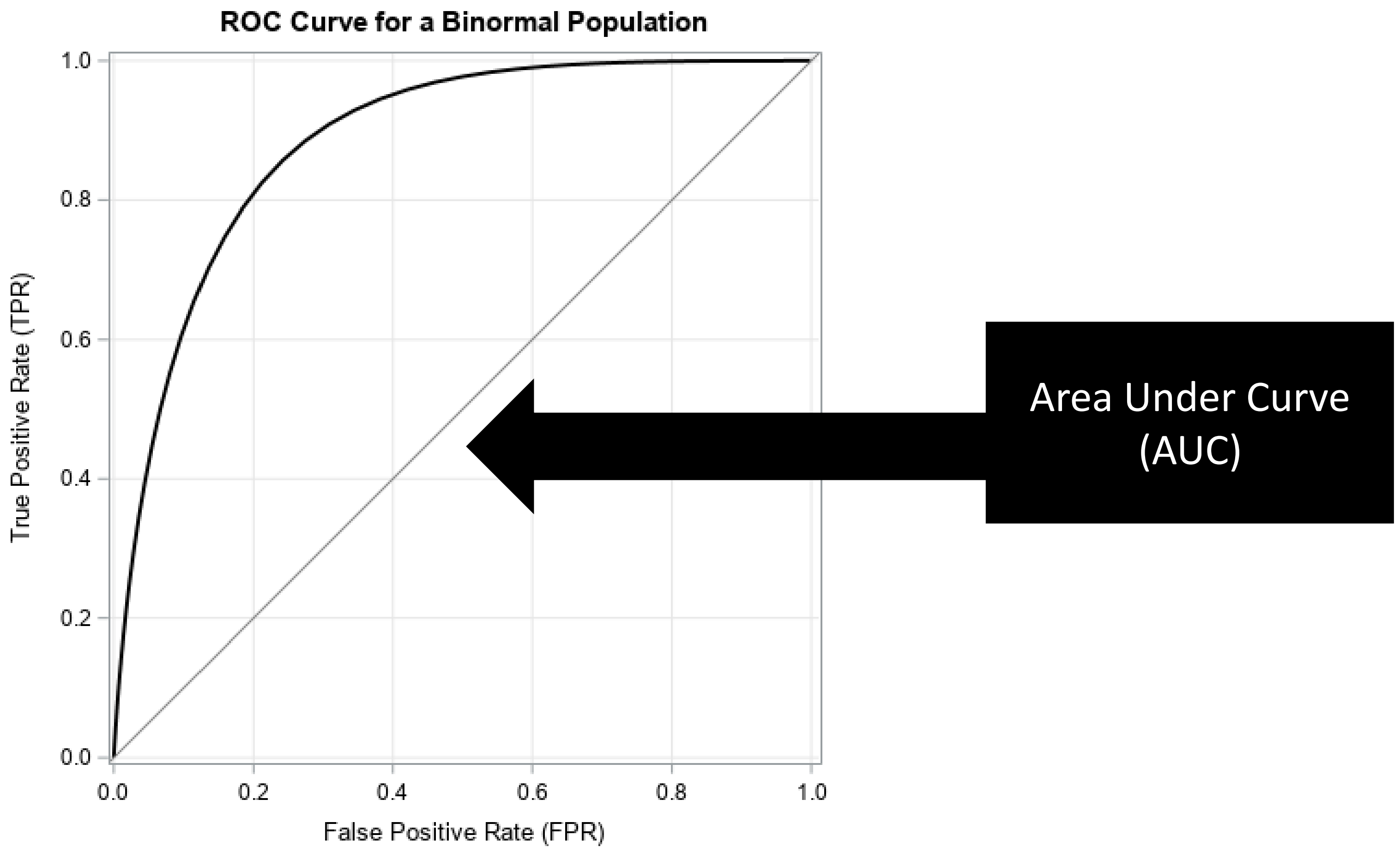
FPR

$$\text{FPR} = 1 - \text{Specificity}$$

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

Receiver Operating Characteristics (ROC Curve)

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis when the cutoff value for classification is changed.



Area Under Curve (AUC)

With the area under the ROC curve, an idea can be got about the accuracy of a model.

AUC Range	Classification Accuracy
0.9-1.0	Excellent
0.8-0.9	Good
0.7-0.8	Fair
0.6-0.7	Bad
0.5-0.6	Very Bad