FINGERTIPS

ML Classification &

Logistics Regression

Classification

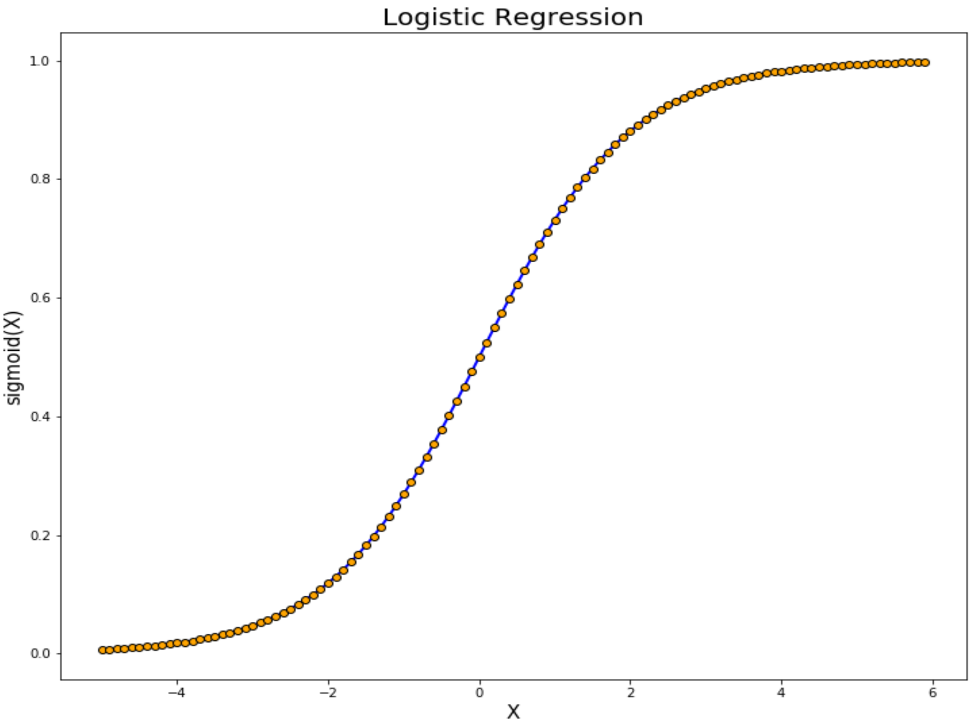
- Classification of machine learning and statistics is a supervised technique of learning in which the computer program learns and produces new discoveries or classifications from the data presented to it.
- Classification is a method that can be used on both structured and unstructured data to categorize a given set of data into classes. The process begins with the class of given data points being projected. Groups are also referred to as target, label or categories.
- The predictive modelling of classification is the phenomenon of approximating the mapping function to discrete output variables from input variables. The primary objective is to determine which class/category the new data would fall under.

Terminologies of Classification in Machine learning

- Classifier is used to map the input data into specific categories.
- Classification model from observed values, a classification model tries to draw some conclusion. A classification model will attempt to predict the significance of one or more effects attributable to one or more inputs. Outcomes are labels that a dataset may be attributed to.
- Feature is a single measurable property of the phenomenon being observed.
- Binary Classification is a type of classification that has only two outcomes. For instance- True or False.
- Multi-Class Classification is the kind of classification that has more than two classes. In multi class classification every sample is assigned to only one target or label.
- Training the Classifier - The fit(X, y) approach is used by each classifier in sci-kit learn to fit the model for training the train X and train label y.
- Predict the Target- give an unlabelled observation x, the predict(X) method would return label y.
- **Evaluate** is basically means the evaluation of the model. i.e – accuracy, classification etc.

Logistic Regression

- It is a classification algorithm in machine learning that uses one or multiple independent variables to extrapolate or determine some outcome.
- The basic idea of logistic regression is to find relationships between features an the probability of particular outcome.
- A logistic regression can be represented as:
  - $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$
- Wherein, the left side is known as logit or log-odds function, and **p(x)/(1-p(x)** is odds.
- Hence, in Logistic Regression, linear combination of inputs are mapped to the log(odds) - the output being equal to 1.
- Following is a graphical representation of logistic regression.

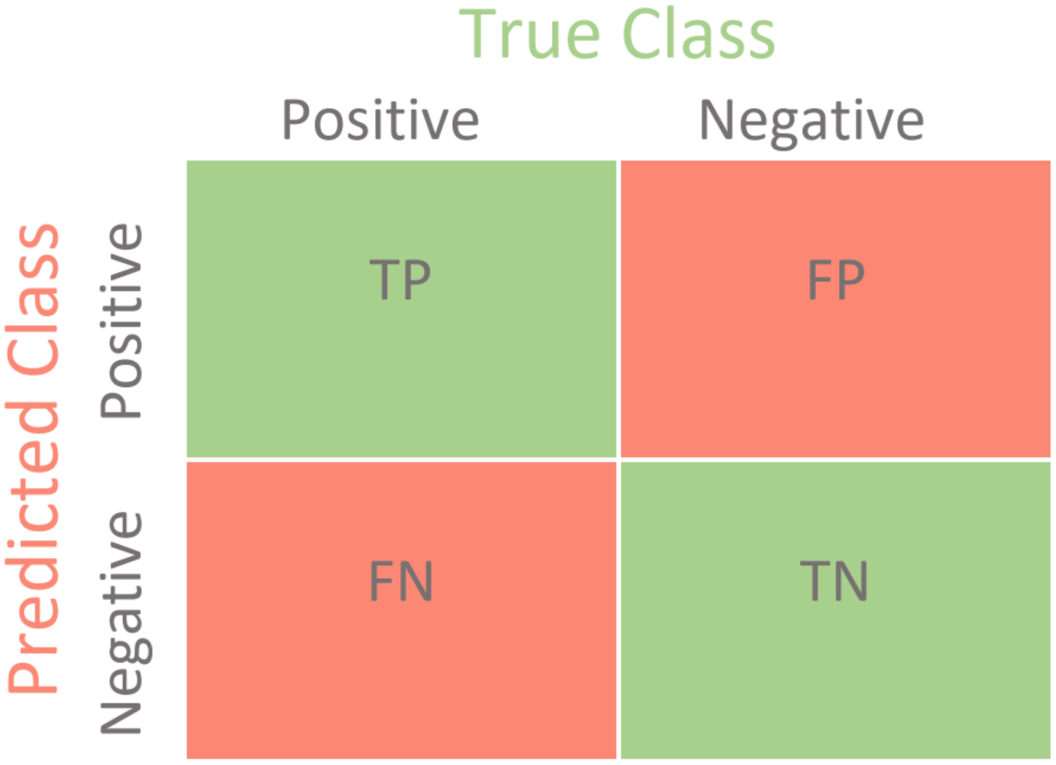


Decision boundary

Confusion Matrix

One way to evaluate the performance of a classifier is to look at the confusion matrix. Usually, the concept is to count the number of times of class A are classified as class B.

In other words, a confusion matrix is a table frequently used to define the output of a classification model (or "classifier") on a collection of test data that are considered to be true values.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Type I Error- predicting a negative case as a negative one.

Type II Error- predicting a positive case as a negative one.

Precision

- Precision is basically a measure of all the accurate predictions out of total positive predictions.
- That is the 'Exactness', the model's capacity to return only relevant events. If your case/problem statement includes eliminating the False Positives. For instance- if you don't want the Forged Notes to be labelled as Genuine by the Model in the current situation, so you need Specificity.

Precision is:

Of all the employees we predicted would stay, what fraction of them actually stayed? In our case the precision was 0.48.

$$\text{precision} = \frac{TP}{TP + FP}$$

- Here, TP = number of true positives
- FP = number of false positives

Recall

- The traditional way to gain a perfect precision is to make a single positive prediction and ensure it is correct. However, this would not be very useful, because the classifier will ignore all but one positive instance.
- Therefore, precision is used another metric named recall, also called sensitivity or the true positive rate (TPR): it is the ratio of positive instances that are accurately detected by the classifier.

recall is:

Of all employees in the project that actually have left, what fraction did we correctly predict? In our case the recall we 0.21.

$$\text{recall} = \frac{TP}{TP + FN}$$

Example:

Assume a project where we are making cancer predictions:

precision is:

Of all patients we predicted have cancer, what fraction of them actually have cancer?

And recall is:

Of all patients in the set that actually have cancer, what fraction did we correctly detect?

Accuracy

- Accuracy: (True Positive + True Negative) / Total Population
- Accuracy is a ratio of correctly predicted observation to the total observations. Accuracy is the most intuitive performance measure.
- True Positive: The number of correct predictions that the occurrence is positive
- True Negative: The number of correct predictions that the occurrence is negative
- In our HR Project the Accuracy was 0.75

F1 score:

- F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.
- Considering out HR dataset, the F1 score was 0.29

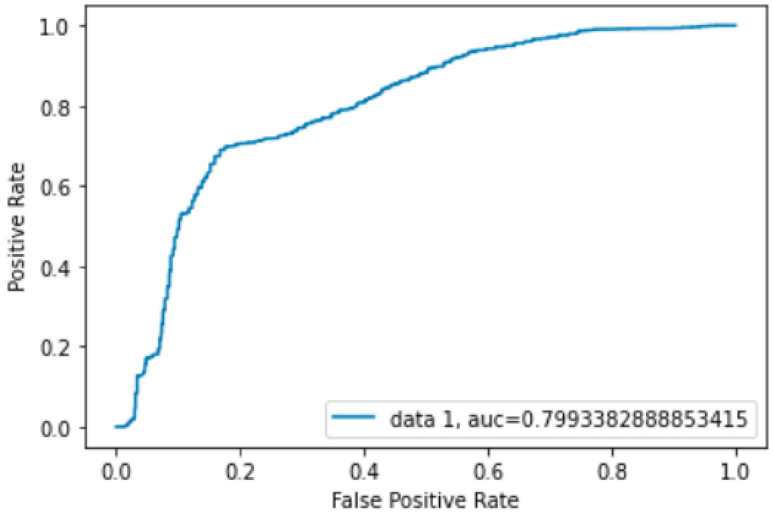
$$\text{F1 Score} = \frac{2 \text{ x } (\text{Precision x Recall})}{\text{Precision + Recall}}$$

ROC CURVE

A Receiver Operating Characteristic (ROC) Curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate.*

A ROC plot shows:

- The relationship between sensitivity and specificity. For example, a decrease in sensitivity results in an increase in specificity.
- Test accuracy; the closer the graph is to the top and left-hand borders, the more accurate the test. Likewise, **the closer the graph to the diagonal, the less accurate the test**. A perfect test would go straight from zero up to the top-left corner and then straight across the horizontal.
- The likelihood ratio; given by the derivative at any particular cut point.
- Given below is the Roc obtained from our HR Project


Positive Rate
data 1, auc=0.7993382888853415
False Positive Rate