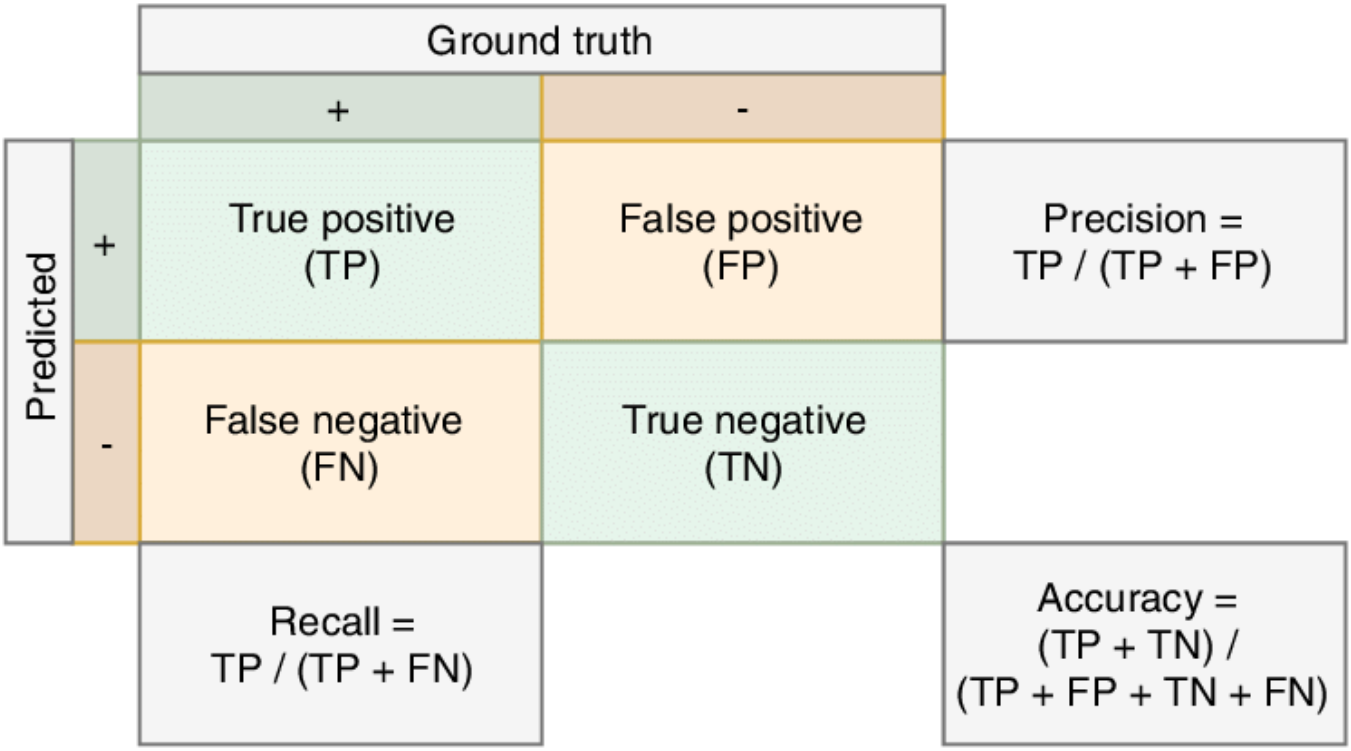
Performance metrics and Model Evaluation

Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

Lets assume we have a binary classification problem. We have some samples belonging to two classes : YES or NO. Also, we have our own classifier which predicts a class for a given input sample.



Classification Accuracy

Classification Accuracy is the ratio of number of correct predictions to the total number of input samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

It works well only if there are equal number of samples belonging to each class.

1. Precision (Specificity, True Negative Rate) : It is the number of correct positive results divided by the number of positive results predicted by the classifier.

1. Recall (Sensitivity, True Positive Rate) : It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

1. False Positive Rate : False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$FalsePositiveRate = \frac{FalsePositive}{TrueNegative + FalsePositive}$$

F1 Score

F1 Score is used to measure a test's accuracy

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as :

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

F1 Score tries to find the balance between precision and recall.

Summary

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\
specificity &= \frac{TN}{TN + FP}
\end{aligned}
$$

		Predicted	
		Cancer = Yes	Cancer = No
Actual	Cancer = Yes	True Positive (TP) = 25	False Negative (FN) = 5
	Cancer = No	False Positive(FP) = 5	True Negative(TN) = 65

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})} = \frac{25}{(25+5)} = \frac{25}{30} = 0.83$$

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} = \frac{25}{(25+5)} = \frac{25}{30} = 0.83$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})}$$

$$= \frac{25 + 65}{(25 + 5 + 65 + 5)} = \frac{90}{100} = 0.90$$

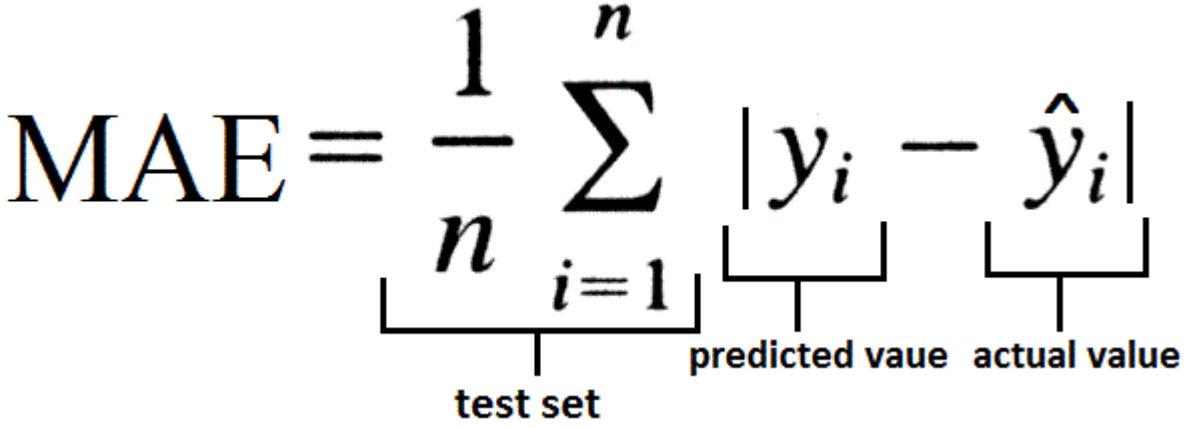
$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

To summarise the differences between the F1-score and the accuracy,

1. Accuracy is used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial
2. Accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes as in the above case.
3. In most real-life classification problems, imbalanced class distribution exists and thus F1-score is a better metric to evaluate our model on.

Mean Absolute Error

Mean Absolute Error (MAE) is the average of the difference between the Original Values and the Predicted Values. However, they don't gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Mathematically, it is represented as :



Mean Squared Error

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the square of the difference between the original values and the predicted values. T As, we take square of the error, the effect of larger errors become more pronounced than smaller error, hence the model can now focus more on the larger errors.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

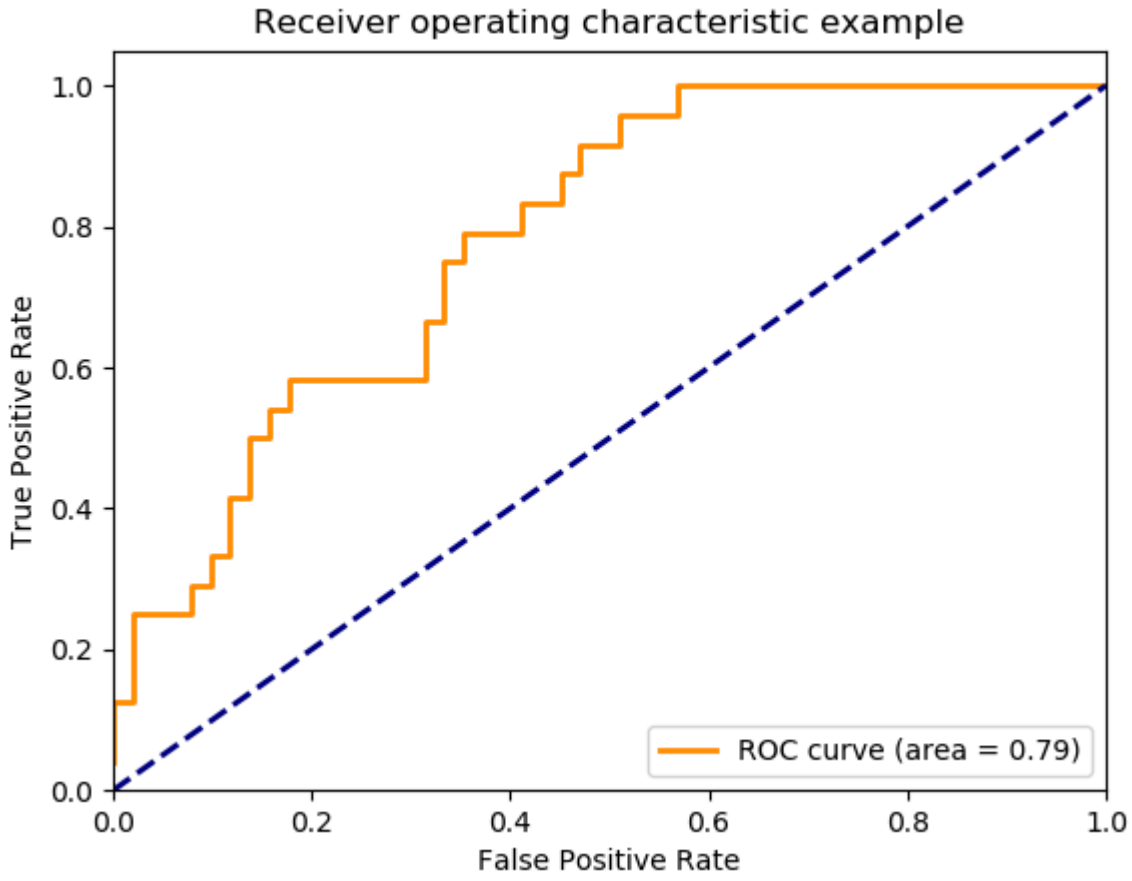
Receiver Operating Characteristics (ROC) Curve

Measuring the area under the ROC curve is also a very useful method for evaluating a model. It shows the performance of a binary classifier as function of its cut-off threshold. It essentially shows the sensitivity against the **false positive rate** for various threshold values.

We write a function which allows use to make predictions based on different probability cutoffs, and then obtain the accuracy, sensitivity, and specificity for these classifiers.

Area Under Curve

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.



In [ ]: