We will discuss the various ways to check the performance of our machine learning or deep learning model and why to use one in place of the other. We will discuss terms like:

1. Confusion matrix
2. Accuracy
3. Precision
4. Recall
5. Specificity
6. F1 score
7. Precision-Recall or PR curve
8. **ROC** (**R**eceiver **O**perating **C**haracteristics) curve
9. PR vs ROC curve.

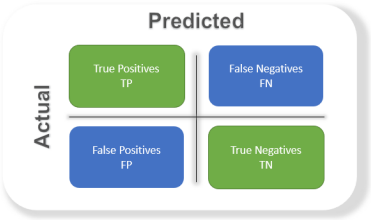
For simplicity, we will mostly discuss things in terms of a binary classification problem where let's say we'll have to find if an image is of a cat or a dog. Or a patient is having cancer (positive) or is found healthy (negative). Some common terms to be clear with are:

True positives (TP): Predicted positive and are actually positive.
False positives (FP): Predicted positive and are actually negative.
True negatives (TN): Predicted negative and are actually negative.
False negatives (FN): Predicted negative and are actually positive.

So let's get started!

Confusion matrix

It's just a representation of the above parameters in a matrix format. Better visualization is always good :)



Accuracy

The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Take for example a cancer detection model. The chances of actually having cancer are very low. Let's say out of 100, 90 of the patients don't have cancer and the remaining 10 actually have it. We don't want to miss on a patient who is having cancer but goes undetected (false negative). Detecting everyone as not having cancer gives an accuracy of 90% straight. The model did nothing here but just gave cancer free for all the 100 predictions.
We surely need better alternatives.

Precision

Percentage of positive instances out of the ***total predicted positive*** instances. Here denominator is the model prediction done as positive from the whole given dataset. Take it as to find out '*how much the model is right when it says it is right'.*

$$\frac{TP}{TP + FP}$$

Recall/Sensitivity/True Positive Rate

Percentage of positive instances out of the ***total actual positive*** instances. Therefore denominator (*TP + FN)* here is the *actual* number of positive instances present in the dataset. Take it as to find out '*how much extra right ones, the model missed when it showed the right ones'.*

$$\frac{TP}{TP + FN}$$

Specificity

Percentage of negative instances out of the ***total actual negative*** instances. Therefore denominator (*TN + FP)* here is the *actual* number of negative instances present in the dataset. It is similar to recall but the shift is on the negative instances. *Like finding out how many healthy patients were not having cancer and were told they don't have cancer*. Kind of a measure to see how separate the classes are.

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

F1 score

It is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better. See that due to the product in the numerator if one goes low, the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).

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

One drawback is that both precision and recall are given equal importance due to which according to our application we may need one higher than the other and F1 score may not be the exact metric for it. Therefore either weighted-F1 score or seeing the PR or ROC curve can help.

PR curve

It is the curve between precision and recall for various threshold values. In the figure below we have 6 predictors showing their respective precision-recall curve for various threshold values. The top right part of the graph is the ideal space where we get high precision and recall. Based on our application we can choose the predictor and the threshold value. PR AUC is just the area under the curve. The higher its numerical value the better.

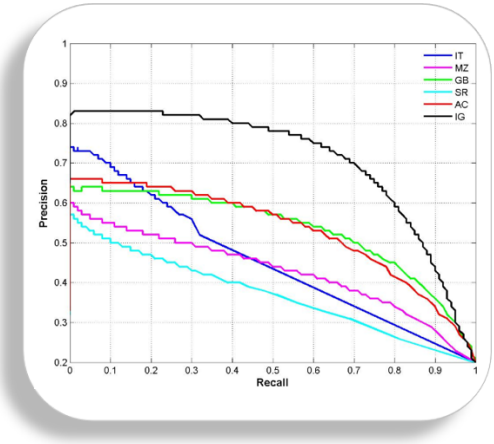


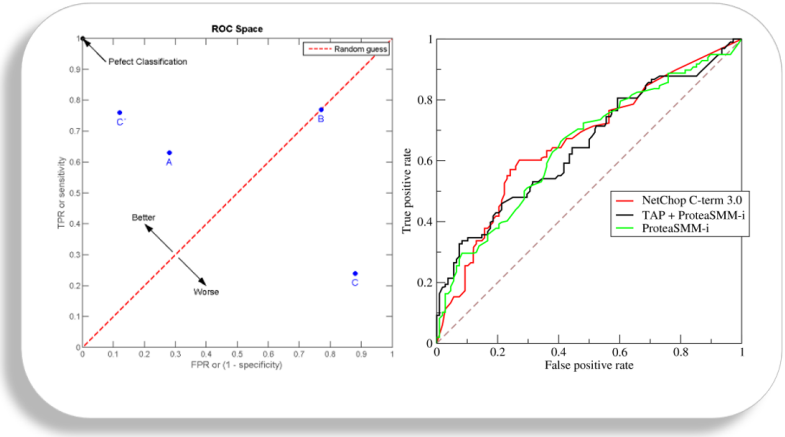
image link

ROC curve

ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. As you can see in the first figure, we have four categories and we want the threshold value that leads us closer to the top left corner. Comparing different predictors (here 3) on a given dataset also becomes easy as you can see in figure 2, one can choose the threshold according to the application at hand. ROC AUC is just the area under the curve, the higher its numerical value the better.

$$\textit{True Positive Rate (TPR) = RECALL} = \frac{TP}{TP+FN}$$

$$\textit{False Positive Rate (FPR) = 1} - Specificity = \frac{FP}{TN+FP}$$

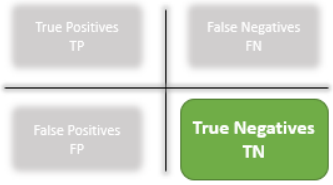


wiki link

PR vs ROC curve

Both the metrics are widely used to judge a models performance.

Which one to use PR or ROC?



The answer lies in TRUE NEGATIVES.

Due to the absence of TN in the precision-recall equation, they are useful in imbalanced classes. In the case of class imbalance when there is a majority of the negative class. The metric doesn't take much into consideration the high number of TRUE NEGATIVES of the negative class which is in majority, giving better resistance to the imbalance. This is important when the detection of the positive class is very important.

Like to detect cancer patients, which has a high class imbalance because very few have it out of all the diagnosed. We certainly don't want to miss on a person having cancer and going undetected (recall) and be sure the detected one is having it (precision).

Due to the consideration of TN or the negative class in the ROC equation, it is useful when both the classes are important to us. Like the detection of cats and dog. The importance of true negatives makes sure that both the classes are given importance, like the output of a CNN model in determining the image is of a cat or a dog.

Conclusion

The evaluation metric to use depends heavily on the task at hand. For a long time, accuracy was the only measure I used, which is really a vague option. I hope this post would have been useful for you. That's all from my side. Feel free to suggest corrections and improvements.