Confusion matrix

- In a binary decision problem, a classier labels examples as either positive or negative.
- The decision made by the classier can be represented in a structure known as a confusion matrix or contingency table.
- The confusion matrix has four categories: True positives (TP) are examples correctly labeled as positives.
- False positives (FP) refer to negative examples incorrectly labeled as positive. True negatives (TN) correspond to negatives correctly labeled as negative.
- Finally, false negatives (FN) refer to positive examples incorrectly labeled as negative.

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

(a) Confusion Matrix

Recall	=	$\frac{TP}{TP+FN}$
Precision	=	$\frac{TP}{TP+FP}$
True Positive Rate	=	$\frac{TP}{TP+FN}$
False Positive Rate	=	$\frac{FP}{FP+TN}$

(b) Definitions of metrics

Figure 2. Common machine learning evaluation metrics

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N} \qquad \text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

For the accuracy score, it shows the percentage of the true positive and true negative to all data points. So, it's useful when the data set is balanced.

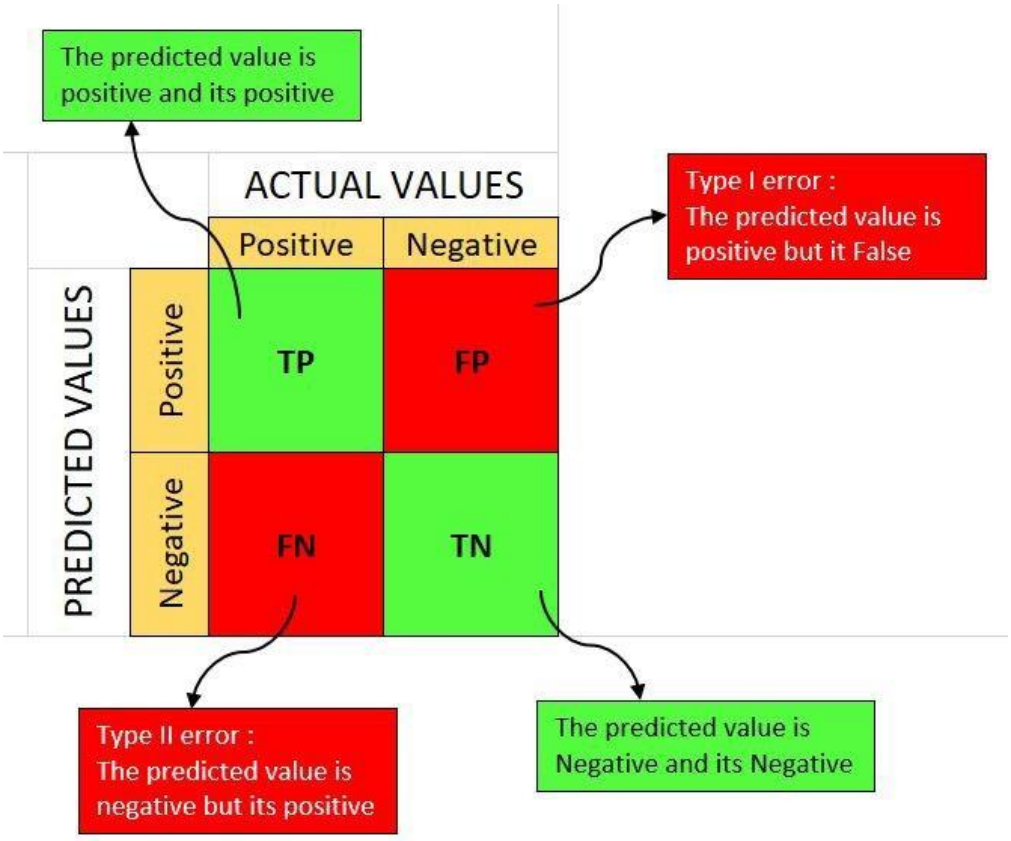
For the f1 score, it calculates the harmonic mean between precision and recall, and both depend on the false positive and false negative. So, it's useful to calculate the f1 score when the data set isn't balanced.

In case of classification problem we should be equipped with different assessment metrics to analyze the classification algorithm. They are:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. Area under ROC curve(AUC)

CONFUSION MATRIX

The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label. Let's take an example of classifying whether a person has a heart disease or not:



In above confusion matrix table, green diagonal shows the result that in actually they are one thing and model also predicted that same thing but red diagonal shows the result that in actual they are one thing and model predicted it another thing.

We can use confusion matrix when we compare different model by looking how well it predicted a true positive(TP) and true negative(TN). If one model predicted a TP and TN very well than other model then we choose this model as our base model.

PRECISION

In definition it is define as the actual correct prediction divided by total prediction made by model. In simple language let our model predict that out of 10 patient 7 has heart disease and among that predicted 7 patient, only 3 has actual heart disease so in this case precision is 3/7 = 0.428

$$\text{precision} \stackrel{\text{def}}{=} \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

In our case true positive(TP) is 3 and false positive(FP) 7-3=4

RECALL

In an classification problem with two classes, recall is calculated as the number of true positives divided by the total number of true positives and false negatives. In or case of heart disease patient let's suppose there is 7 actual heart patient but our model predict only 5 has heart disease so in this case recall is 5/7=0.714

How to choose Precision and Recall?

Almost always, in practice, we have to choose between a high precision or a high recall. It's usually impossible to have both. We can achieve either of the two by various means:

- Assigning a higher weighting to the examples of a specific class (the SVM algorithm accepts weightings of classes as input)
- Tuning hyperparameters to maximize precision or recall on the validation set.
- Varying the decision threshold for algorithms that return probabilities of classes; for instance, if we use logistic regression or decision tree, to increase precision (at the cost of a lower recall), we can decide that the prediction will be positive only if the probability returned by the model is higher than 0.9.

Even if precision and recall are defined for the binary classification case, you can always use it to assess a multiclass classification model. To do that, first select a class for which you want to assess these metrics. Then you consider all examples of the selected class as positives and all examples of the remaining classes as negatives.

ACCURACY

It is defined as total correctly classified example divided by the total number of classified examples. Lets express it in terms of confusion matrix:

$$\text{accuracy} \stackrel{\text{def}}{=} \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

This metric is very important when error in predicting all class is equally important. Here False positive is most important to address than False negative. Lets take a example of an email is spam or not spam. In this case if our model classify a email send by boss is spam and don't show it is more harmful than showing small amount of email as spam.

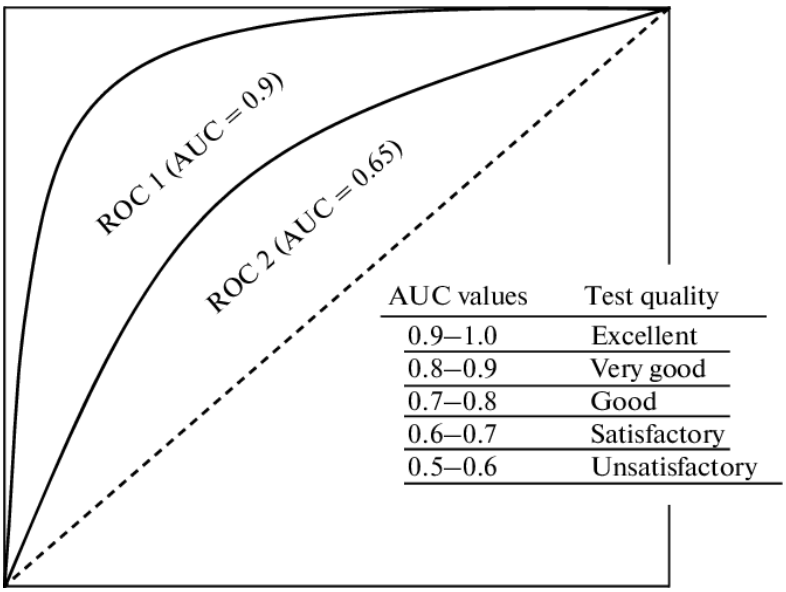
AREA UNDER THE ROC CURVE (AUC)

It can only be used to assess classifiers that return some confidence score (or a probability) of prediction. For example, logistic regression, neural networks, and decision trees (and ensemble models based on decision trees) can be assessed using ROC curves.

ROC curve commonly use the combination of true positive rate(TPR) and false positive rate(FPR) and that is given as:

$$\text{TPR} \stackrel{\text{def}}{=} \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ and } \text{FPR} \stackrel{\text{def}}{=} \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

The higher the area under the ROC curve (AUC), the better the classifier. A classifier with an AUC higher than 0.5 is better than a random classifier. If AUC is lower than 0.5, then something is wrong with your model. A perfect classifier would have an AUC of 1.



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

ROC curve capture more than one aspect of the classification (by taking both false positives and negatives into account) and allow visually and with low effort comparing the performance of different models.

F1 SCORE

F1 score is a weighted average of precision and recall. As we know in precision and in recall there is false positive and false negative so it also consider both of them. F1 score is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)