

Metrics



Metrics

True Positives (TP) - The correctly predicted positive values. The value of the actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that the patient survived and predicted class says the same.

True Negatives (TN) - The correctly predicted negative values. The value of the actual class is no and value of predicted class is also no. E.g. the actual class says the patient did not survive and predicted class says the same.

False Positives (FP) – The actual class is no and predicted class is yes. E.g. if actual class says the patient did not survive but predicted class tells you that the patient survived.

False Negatives (FN) – The actual class is yes but predicted class is no. E.g. if actual class value indicates that the patient survived and predicted class says the patient dies.

We wish to minimize False Positives and False Negatives.

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Confusion Matrix

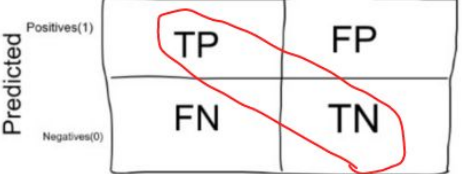
Accuracy

Accuracy is a good measure when the target variable classes in the data are nearly balanced. Accuracy should NEVER be used as a measure when the target variable classes in the data are a majority of one class.

In cancer detection, with 100 people, assume only 5 people have cancer. Our model is that no one has cancer. We have classified the 95 non-cancer patients correctly and the 5 cancer patients as incorrectly (as Non-cancerous). Even though the model is *terrible* at predicting cancer, The accuracy is 95%.

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN



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

Accuracy

Precision

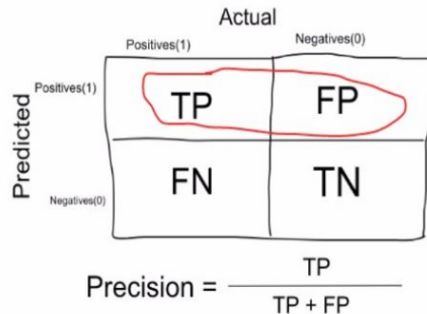
There are two types of error we wish to avoid in statistics.

Type I Error: False positive (rejection of a true null hypothesis)

Type II Error: False negative (non-rejection of a false null hypothesis)

Precision assists in this, giving us a broader picture. It evaluates how precise a model is in predicting positive labels. Use precision if you want to be more confident of your true positives. As an example, in spam emails, we prefer some spam emails in your inbox rather than some regular emails in your spam box. That is to say, use precision when the cost of a false positive is high.

As another example, assume only 5 out of 100 people have cancer and we predict that everyone has cancer. While we would be very *accurate* (all cancer cases were identified correctly), we would have a precision rate of 5%.



Precision

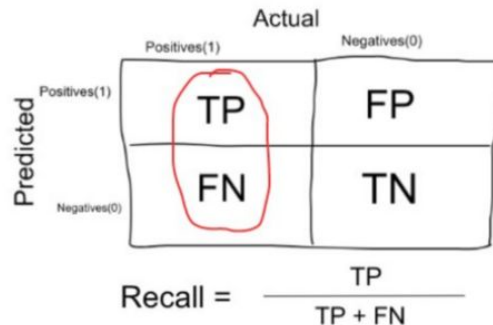
Recall or Sensitivity

Sensitivity of a classifier is the ratio between how much were correctly identified as positive to how much were actually positive. We would use this when classification of positives are a high priority, such as in security checks at an airport.

Use recall if false positives is far better than false negatives, in other words, if the occurrence of false negatives is unacceptable/intolerable, i.e. you would rather have false alarms over false negatives, like in the security example.

In fraud detection, if a fraudulent transaction (actual positive) is predicted as non-fraudulent (predicted negative) the cost could be quite high. Therefore, recall would be a good metric to use.

Similarly, we would rather get *some* healthy people labeled diabetic over leaving a diabetic person labeled healthy. However at an extreme, we would have to be careful. For example, if a model predicted that everyone had a disease, the model would have a perfect recall but it would have a lot of false positives and be telling people they were sick when they were not.



Specificity

Specificity of a classifier is the ratio between how much were correctly classified as negative to how much was actually negative. It is used where classification of negatives is a high priority, for instance, diagnosing a health condition before treatment.

We use specificity if we want to cover all true negatives, i.e. we don't want any false alarms (false positives). Assume we are running a drug test in which all people who test positive will immediately go to jail, you don't want anyone drug-free people going to jail. False positives here are intolerable.

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

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

F1-Score

The F1-Score considers both precision and recall. This score takes both false positives and false negatives into account. It is the harmonic mean of the precision and recall. The F1-score gives a larger weight to lower numbers.

It is best if there is some sort of balance between precision and recall. The F1-Score isn't so high if one measure is improved at the expense of the other.

If the cost of false positives and false negatives are different then F1 is your savior. F1 is best if you have an uneven class distribution.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

ROC/AUC

The Receiver Operating Characteristic (ROC) curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

The “steepness” of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

ROC curves are typically used in binary classification to study the output of a classifier. In order to extend ROC curve and ROC area to multi-label classification, it is necessary to binarize the output. One ROC curve can be drawn per label, but one can also draw a ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging). (via scikit-learn)

The ROC curve lets us see the precision vs recall balance.

The AUC (area under the curve) is a metric to calculate the overall performance of a classification model based on area under the ROC curve.

