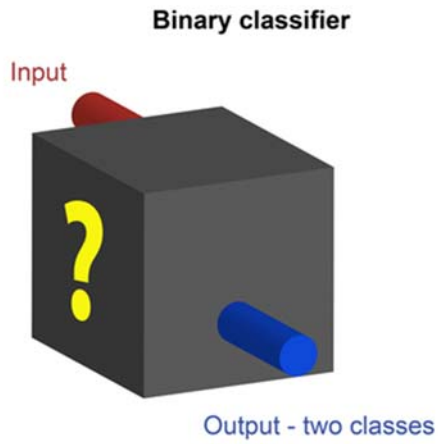
Basic evaluation measures from the confusion matrix

Takaya Saito and Marc Rehmsmeier

We introduce basic performance measures derived from the confusion matrix through this page. The confusion matrix is a two by two table that contains four outcomes produced by a binary classifier. Various measures, such as error-rate, accuracy, specificity, sensitivity, and precision, are derived from the confusion matrix. Moreover, several advanced measures, such as ROC and precision-recall, are based on them.

Test datasets for binary classifier

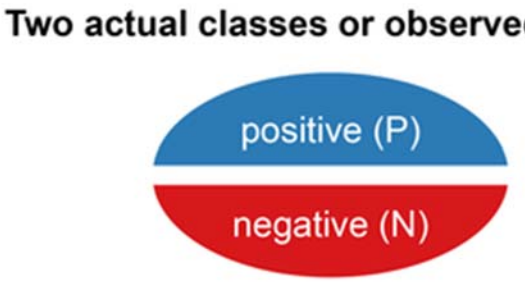
A binary classifier produces output with two class values or labels, such as Yes/No and 1/0, for given input data. The class of interest is usually denoted as “positive” and the other as “negative”.



A binary classifier produces output with two classes for given input data.

Test dataset for evaluation

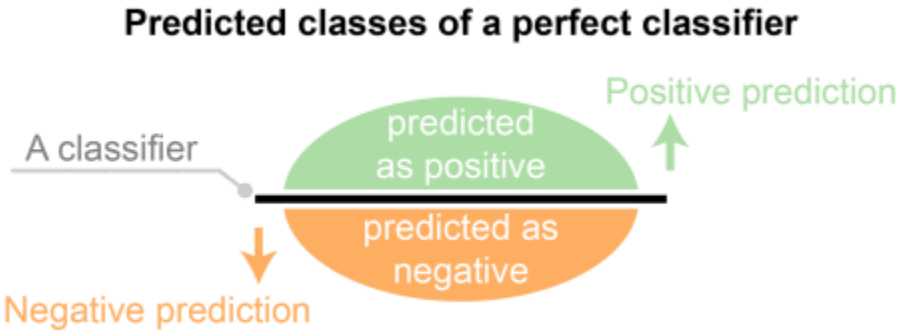
A dataset used for performance evaluation is called a test dataset. It should contain the correct labels (observed labels) for all data instances. These observed labels are used to compare with the predicted labels for performance evaluation after classification.



In binary classification, a test dataset has two labels; positive and negative.

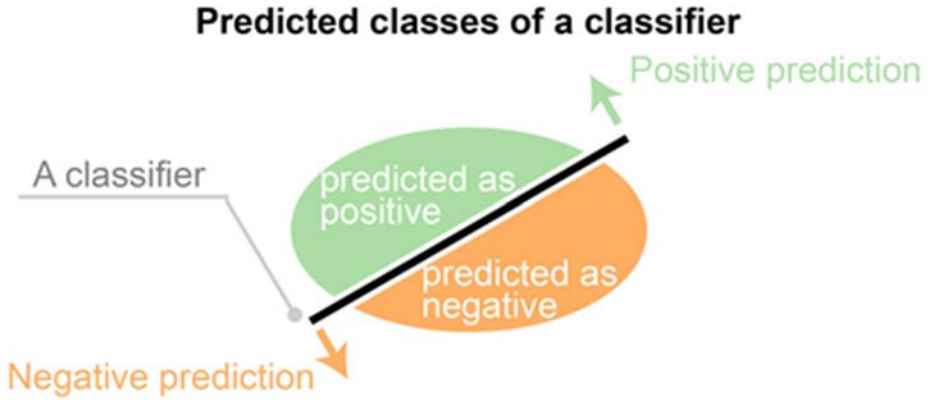
Predictions on test datasets

The predicted labels will be exactly the same if the performance of a binary classifier is perfect, but it is uncommon to be able to develop a perfect binary classifier that is practical for various conditions.



The performance of a binary classifier is perfect when it can predict the exactly same labels in a test dataset.

Hence, the predicted labels usually match with part of the observed labels.



The predicted labels of a classifier match with part of the observed labels.

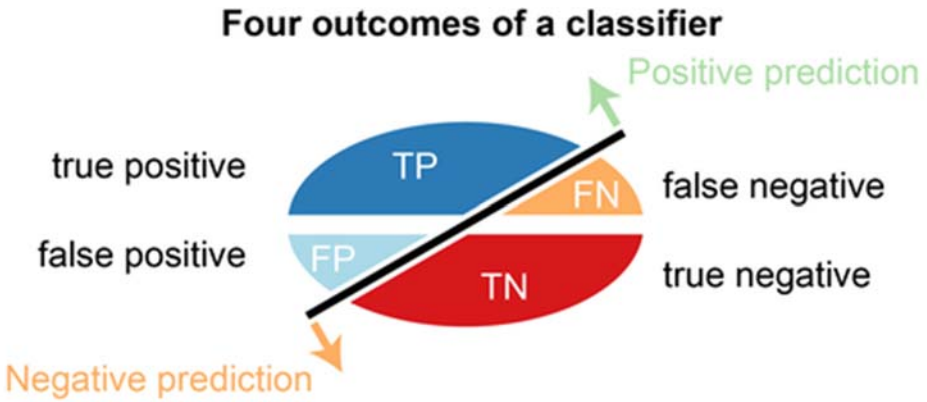
Confusion matrix from the four outcomes

A confusion matrix is formed from the four outcomes produced as a result of binary classification.

Four outcomes of classification

A binary classifier predicts all data instances of a test dataset as either positive or negative. This classification (or prediction) produces four outcomes – true positive, true negative, false positive and false negative.

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction



Classification of a test dataset produces four outcomes – true positive, false positive, true negative, and false negative.

Confusion matrix

A confusion matrix of binary classification is a two by two table formed by counting of the number of the four outcomes of a binary classifier. We usually denote them as TP, FP, TN, and FN instead of "the number of true positives", and so on.

		Predicted	
		Positive	**Negative**
Observed	**Positive**	TP (# of TPs)	FN (# of FNs)
	Negative	FP (# of FPs)	TN (# of TNs)

Basic measures derived from the confusion matrix

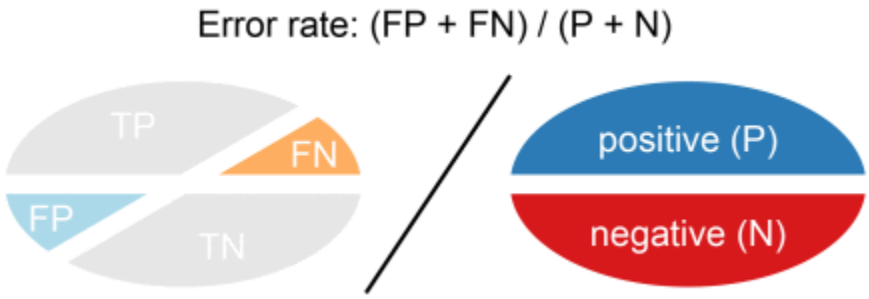
Various measures can be derived from a confusion matrix.

First two basic measures from the confusion matrix

Error rate (ERR) and accuracy (ACC) are the most common and intuitive measures derived from the confusion matrix.

Error rate

Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0.

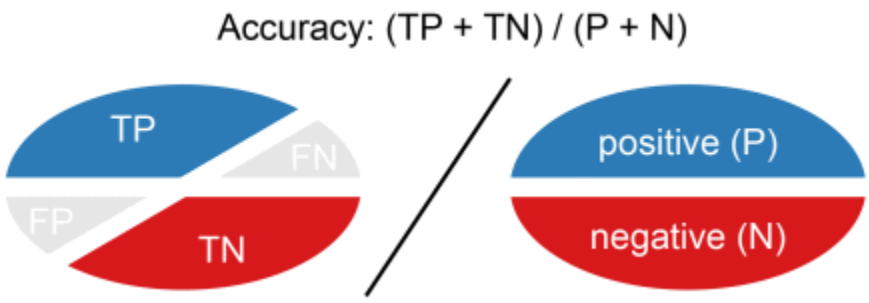


Error rate is calculated as the total number of two incorrect predictions (FN + FP) divided by the total number of a dataset (P + N).

$$\mathrm{ERR} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{P} + \mathrm{N}}$$

Accuracy

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR.



Accuracy is calculated as the total number of two correct predictions (TP + TN) divided by the total number of a dataset (P + N).

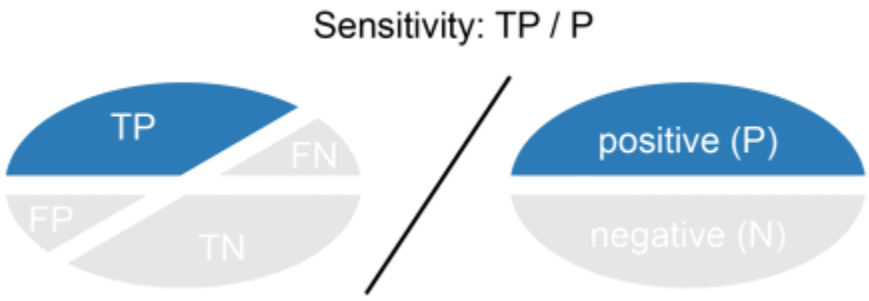
$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

Other basic measures from the confusion matrix

Error costs of positives and negatives are usually different. For instance, one wants to avoid false negatives more than false positives or vice versa. Other basic measures, such as sensitivity and specificity, are more informative than accuracy and error rate in such cases.

Sensitivity (Recall or True positive rate)

Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0.

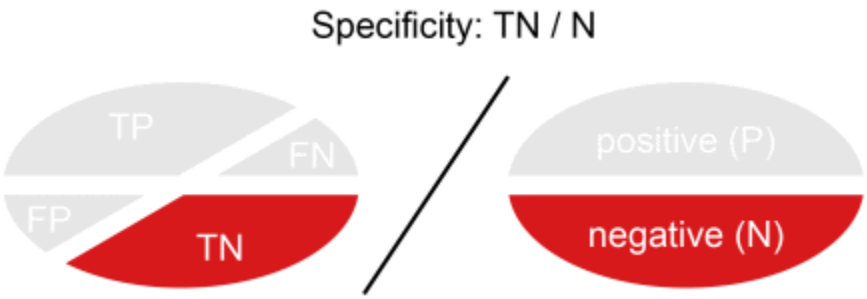


Sensitivity is calculated as the number of correct positive predictions (TP) divided by the total number of positives (P).

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

Specificity (True negative rate)

Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0.

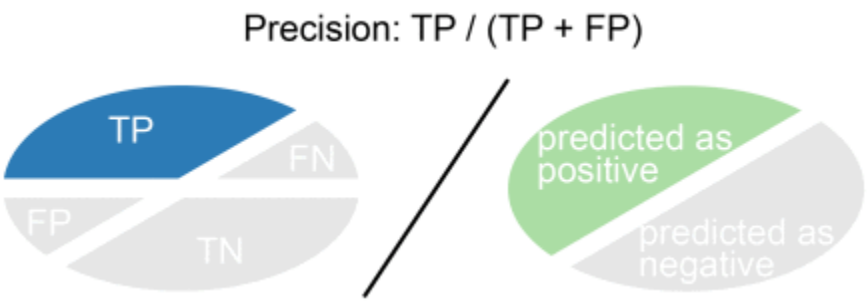


Specificity is calculated as the number of correct negative predictions (TN) divided by the total number of negatives (N).

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{N}}$$

Precision (Positive predictive value)

Precision (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0.

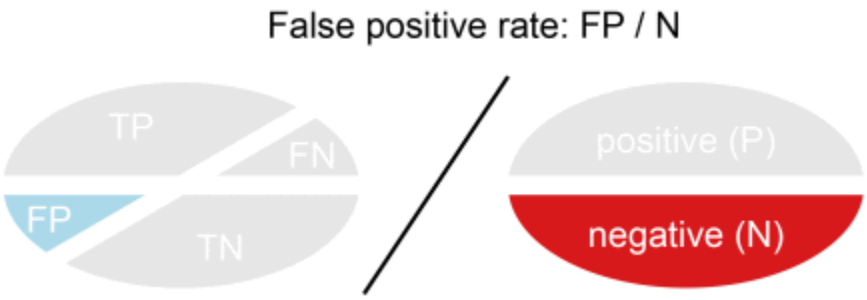


Precision is calculated as the number of correct positive predictions (TP) divided by the total number of positive predictions (TP + FP).

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

False positive rate

False positive rate (FPR) is calculated as the number of incorrect positive predictions divided by the total number of negatives. The best false positive rate is 0.0 whereas the worst is 1.0. It can also be calculated as 1 – specificity.



False positive rate is calculated as the number of incorrect positive predictions (FP) divided by the total number of negatives (N).

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{SP}$$

Correlation coefficient and F-score

Mathews correlation coefficient and F-score can be useful, but they are less frequently used than the other basic measures.

Matthews correlation coefficient

Matthews correlation coefficient (MCC) is a correlation coefficient calculated using all four values in the confusion matrix.

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

F-score

F-score is a harmonic mean of precision and recall.

$$\mathrm{F}_\beta = \frac{(1 + \beta^2)(\mathrm{PREC} \cdot \mathrm{REC})}{(\beta^2 \cdot \mathrm{PREC} + \mathrm{REC})}$$

β is commonly 0.5, 1, or 2.

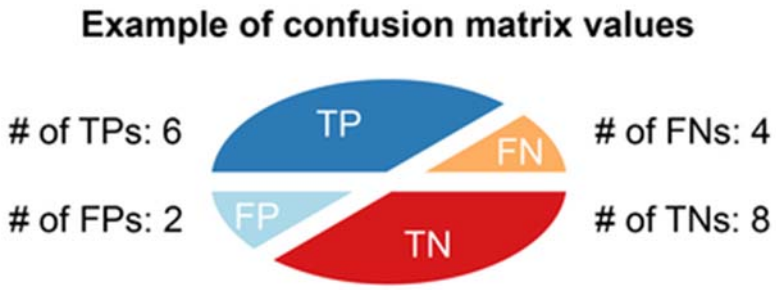
$$\mathrm{F}_{0.5} = \frac{1.25 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{0.25 \cdot \mathrm{PREC} + \mathrm{REC}}$$

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$$

$$F_2 = \frac{5 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{4 \cdot \mathrm{PREC} + \mathrm{REC}}$$

An example of evaluation measure calculations

Let us assume that the outcome of some classification results in 6 TPs, 4 FNs, 8 TNs, and 2 FPs.



This example shows that a binary classifier has produced 6 TPs, 4 FNs, 2 FPs, and 8 TNs.

First, a confusion matrix is formed from the outcomes.

		Predicted	
		Positive	**Negative**
Observed	**Positive**	6	4
	Negative	2	8

Then, the calculations of basic measures are straightforward once the confusion matrix is created.

measure		**calculated value**
Error rate	ERR	6 / 20 = 0.3
Accuracy	ACC	14 / 20 = 0.7
Sensitivity True positive rate Recall	SN TPR REC	6 / 10 = 0.6
Specificity True negative rate	SP TNR	8 / 10 = 0.8
Precision Positive predictive value	PREC PPV	6 / 8 =0.75
False positive rate	FPR	2 / 10 = 0.2