Model Evaluation & Selection

Tue, 24 May 2022 at 12:35

#accuracy #recall #percision #tpr #fpr

Model Evaluation

> *"Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring." domino*

Concerned with testing the classifier's accuracy in terms of predictions and inference.

Terms

`Under the assumption of two classes only`

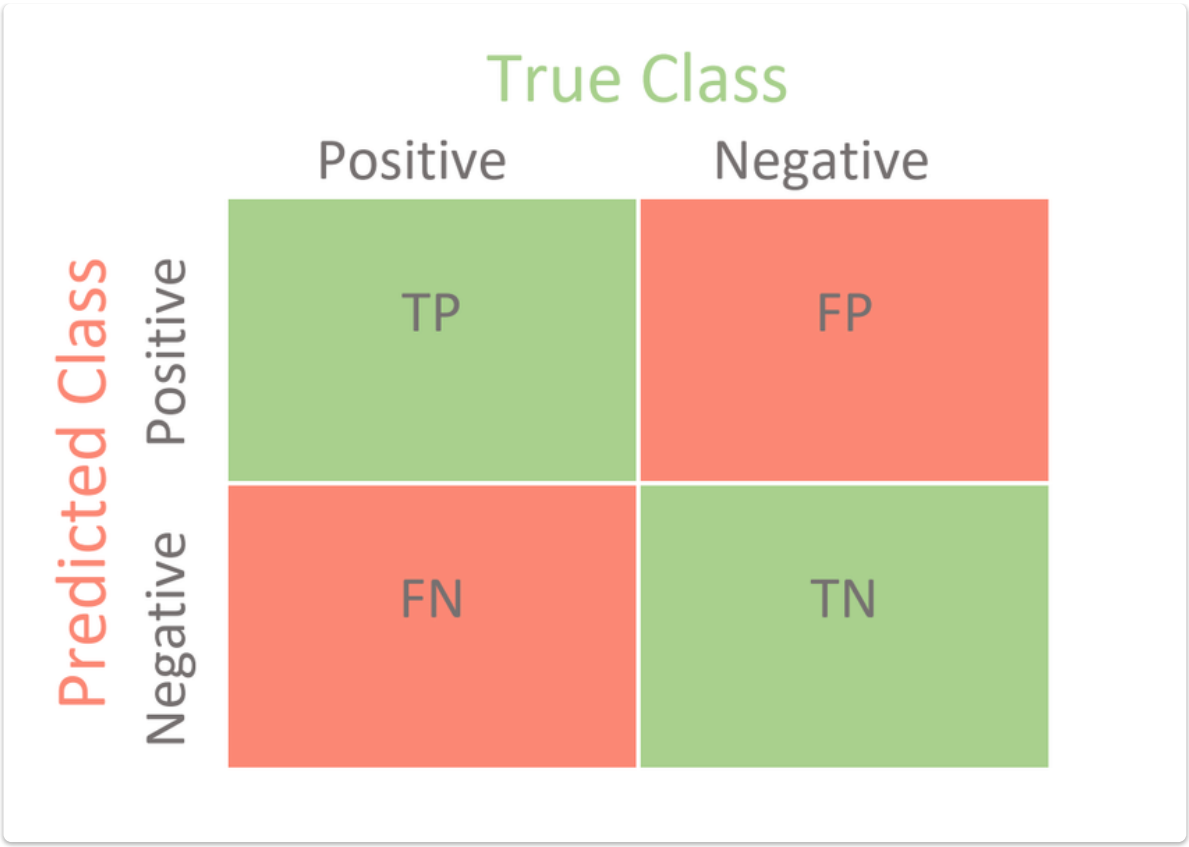
1. *Positive Samples (P)* → the tuples of the positive class
2. *Negative Samples (N)* → the tuples of the other class `negative`
3. *True Positives (TP)* → correctly classified positive tuples
4. *True Negatives (TN)* → correctly classified negative tuples
5. *False Positives (FP)* → incorrectly classified negative tuples
6. *False Negatives (FN)* → incorrectly classified positive tuples

Metrics

CONFUSION MATRIX

Analyses how well your model can recognize tuples of different classes
The goal is for `FP` and `FN` to be 0

- $P = TP + FN$
- $N = TN + FN$



ACCURACY & ERROR RATE

Accuracy

The percentage of test set tuples that are **correctly** classified. Also known as the overall *recognition rate*

$$accuracy = \frac{TP + TN}{P + N}$$

Error Rate

The rate which the model misclassifies the test data

$$error\,rate = 1 - acccuracy = \frac{FP + FN}{N + F}$$

Recall

Also known as *sensitivity* which describes the completeness; what % of all positive tuples did the model classify as positive

$$reacll = \frac{TP}{P}$$

Precision

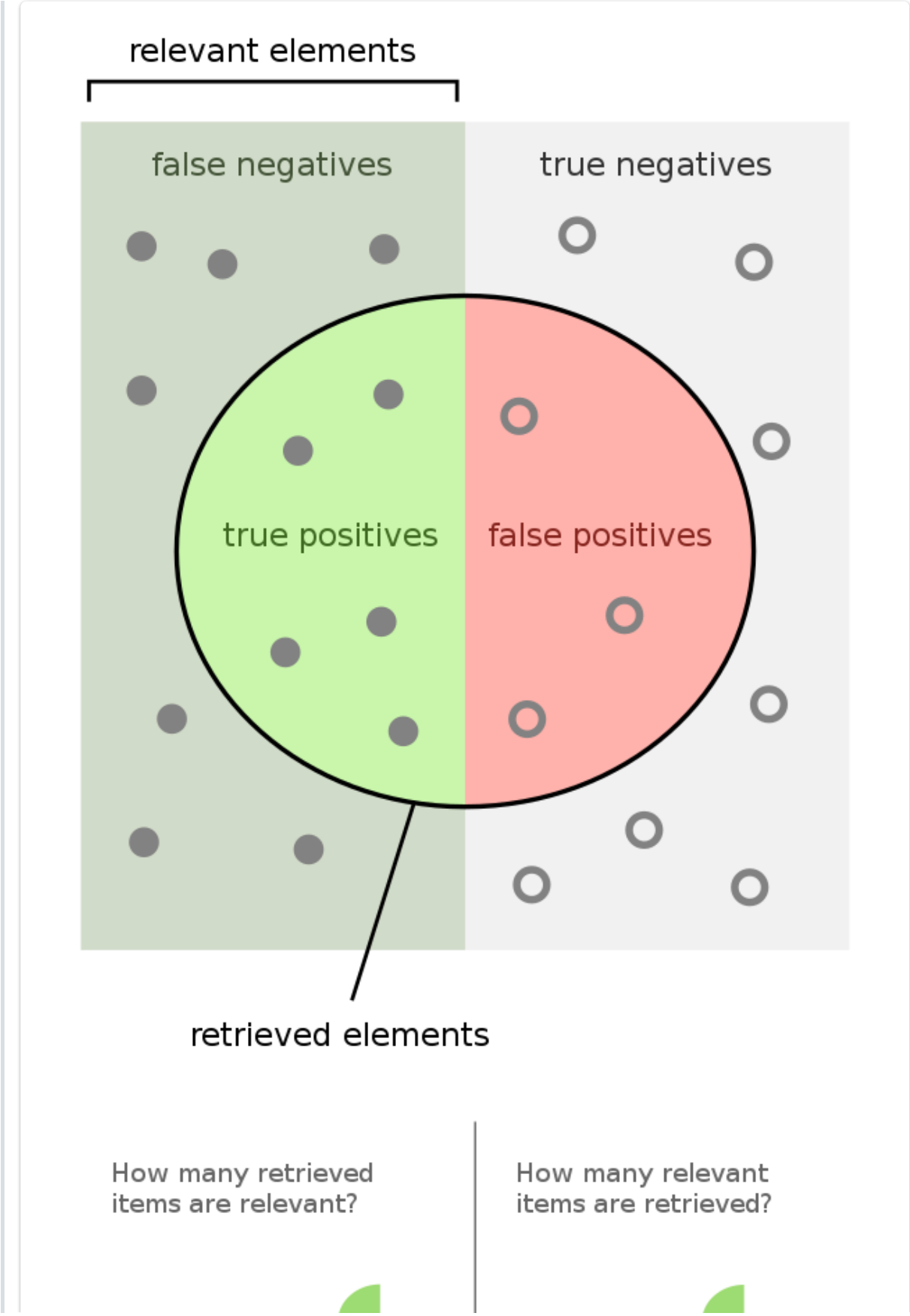
Exactness of what the model predicted as positive is actually positive

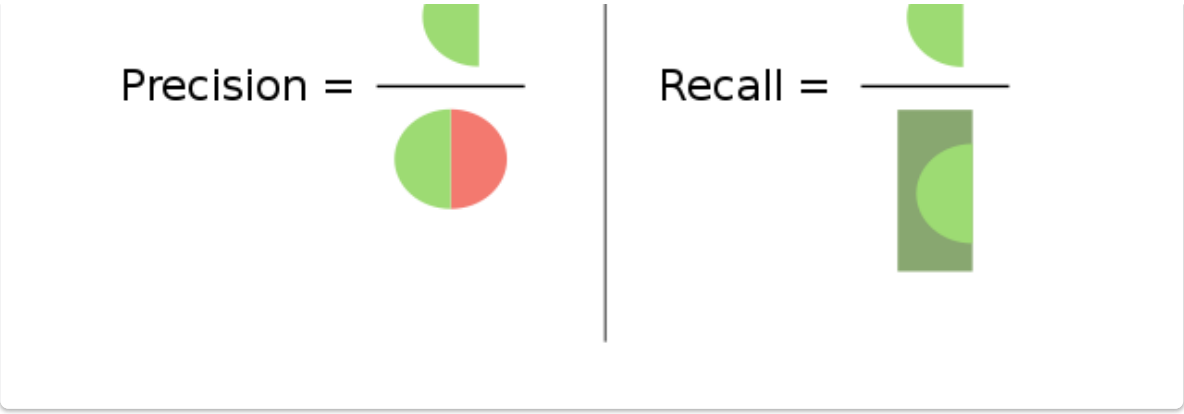
$$percision = \frac{TP}{P'} = \frac{TP}{TP + FP}$$

Specificity

$$specificity = \frac{TN}{N}$$

When a classifier has a good accuracy score but specify or recall is low it means that the model is facing a Class Imbalance Problem
There is an inversed relationship between Recall and Precision





		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

F-MEASURES

F-score

F-measure or F_1 or F-score is the **harmonic mean** of precision and recall. It gives equal weights to both.

- highest value is 1 - perfect precision and recall
- lowest is 0 - either precision or recall is 0

$$F_1 = 2 \times \frac{percision \times recall}{percision + recall}$$

Fbeta

Applies additional weight, valuing precision or recall more than the other. Value of β is chosen such that recall is considered β times more important than precision.

$$F_\beta = (1 + \beta^2) \times \frac{percision \times reacll}{(\beta^2 \times percision) + reacll}$$

Evaluating Classifier Accuracy

HOLDOUT

> *"Holdout Method is **the simplest sort of method to evaluate a classifier**. In this method, the data set (a collection of data items or examples) is separated into two sets, called the Training set and Test set. A classifier performs function of assigning data items in a given collection to a target category or class." geeksforgeeks*

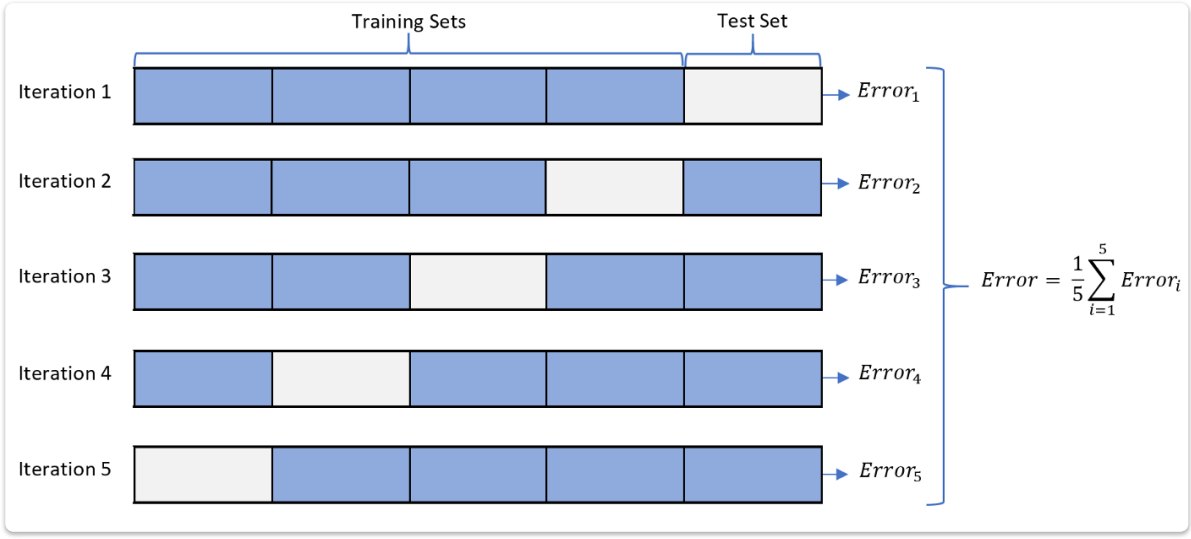
```
# Example in python fr a 70:30 split
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3)
```

RANDOM SUBSAMPLING

A variation of holdout strategy. Repeat holdout k times and the accuracy is the average

CROSS-VALIDATION

k-fold where the data is partitioned into k partitions. iterating the training set each time. Unlike the previous method each sample is used the same number of times for training and once for testing.



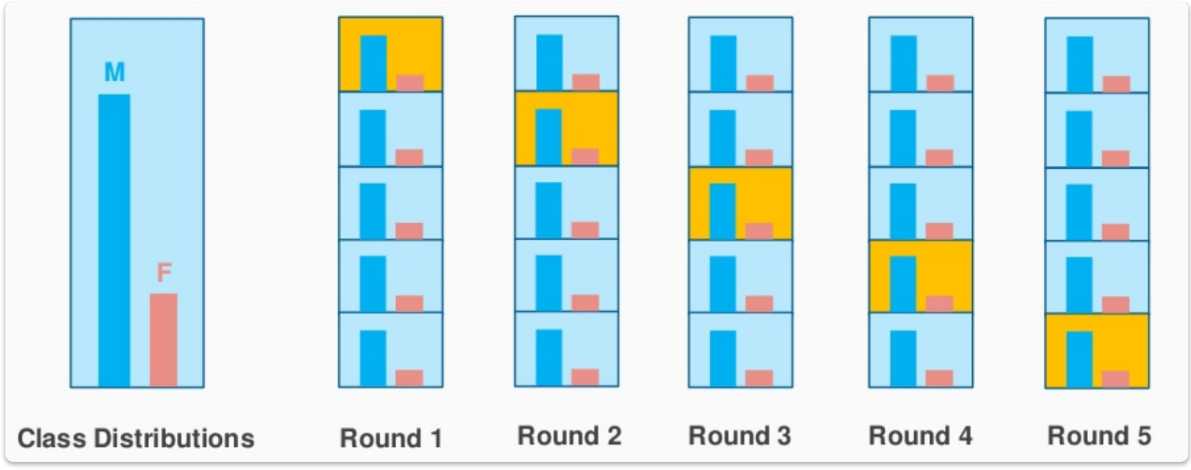
LEAVE ONE OUT

> *"Leave-one-out cross-validation is **a special case of cross-validation where the number of folds equals the number of instances in the data set**."*

k folds where k = # of tuples

STRATIFIED CROSS-VALIDATION

folds are stratified to *class* distribution.



BOOTSTRAP

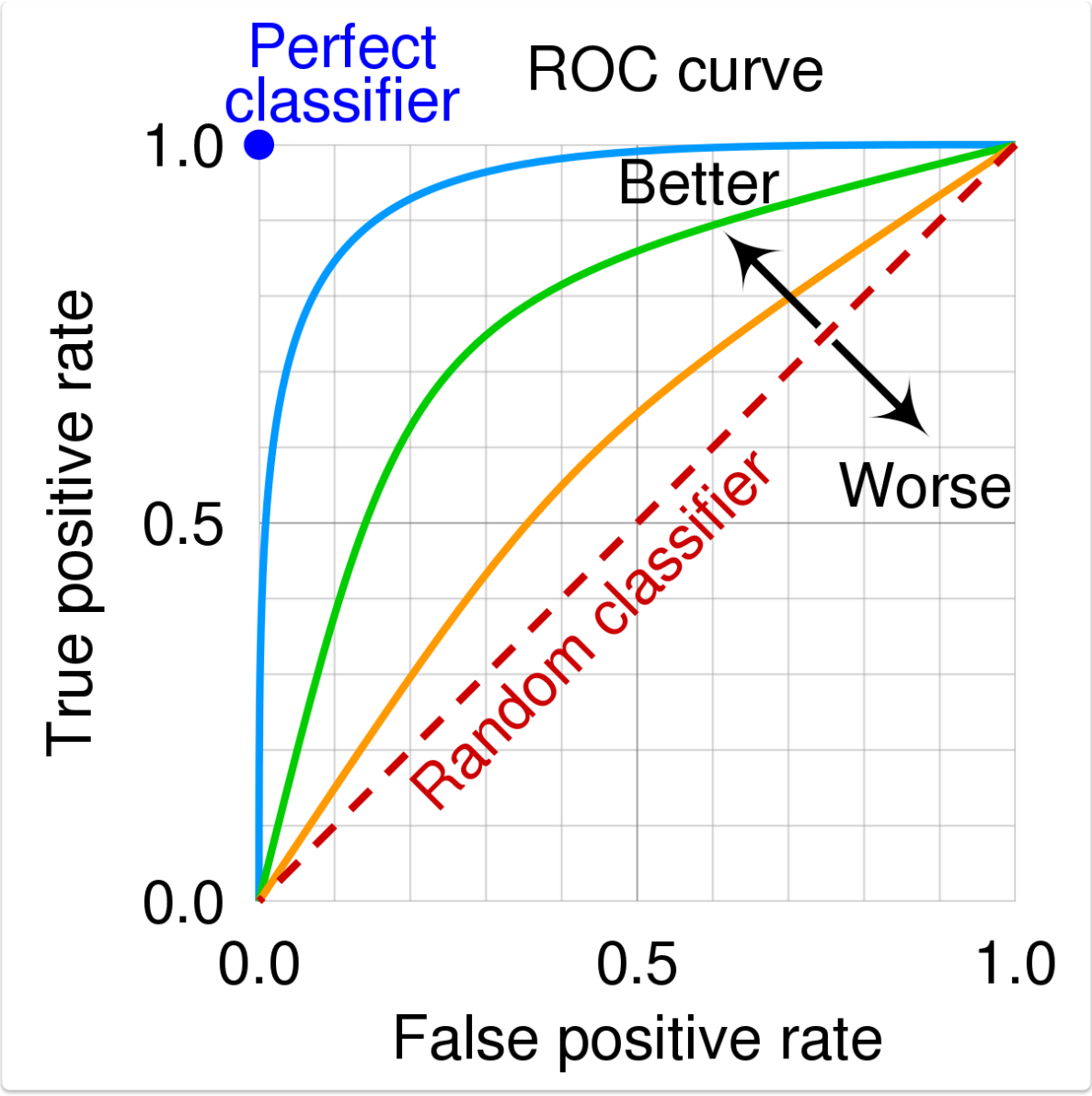
Works well with smaller datasets. Choose the training set with replacement.

Model Selection

Concerned with comparing the classifiers accuracies and choosing the best one

ROC Curves

Shows the tradeoff between the **true positive rate** (TPR) and the **false positive rate** (FPR) used to visually compare different models. The *area under the curve* is **AUC** which is a measure of that model's accuracy.


Perfect classifier
ROC curve
1.0
Better
Worse
0.5
Random classifier
True positive rate
0.0
0.0
0.5
1.0
False positive rate