

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Evaluating Machine Learning Models

Terms

► Classification

- process of matching an object instance to a particular class
- process that assigns a label to an object according to some representation of the object's properties

Terms

► Classifier

- device or algorithm that inputs an object representation and outputs a class label

► Reject class

- generic class for objects that cannot be placed in any of the designated known classes

Supervised vs. Unsupervised Learning

► Supervised learning (classification)

- Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
- New data is classified based on the training set

► Unsupervised learning (clustering)

- The class labels of training data is unknown
- Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Testing

▶ Training set

- Sample data for which truth is known
- Used to develop classifier

▶ Independent test data

- Sample data
- Not part of training set
- Sampled from a real population

Evaluating Machine Learning Model Performance

- Various ways to check the performance of our machine learning may include
 - Confusion matrix
 - Accuracy
 - Precision
 - Recall
 - Specificity
 - F1 score
 - **ROC** (Receiver Operating Characteristics) curve

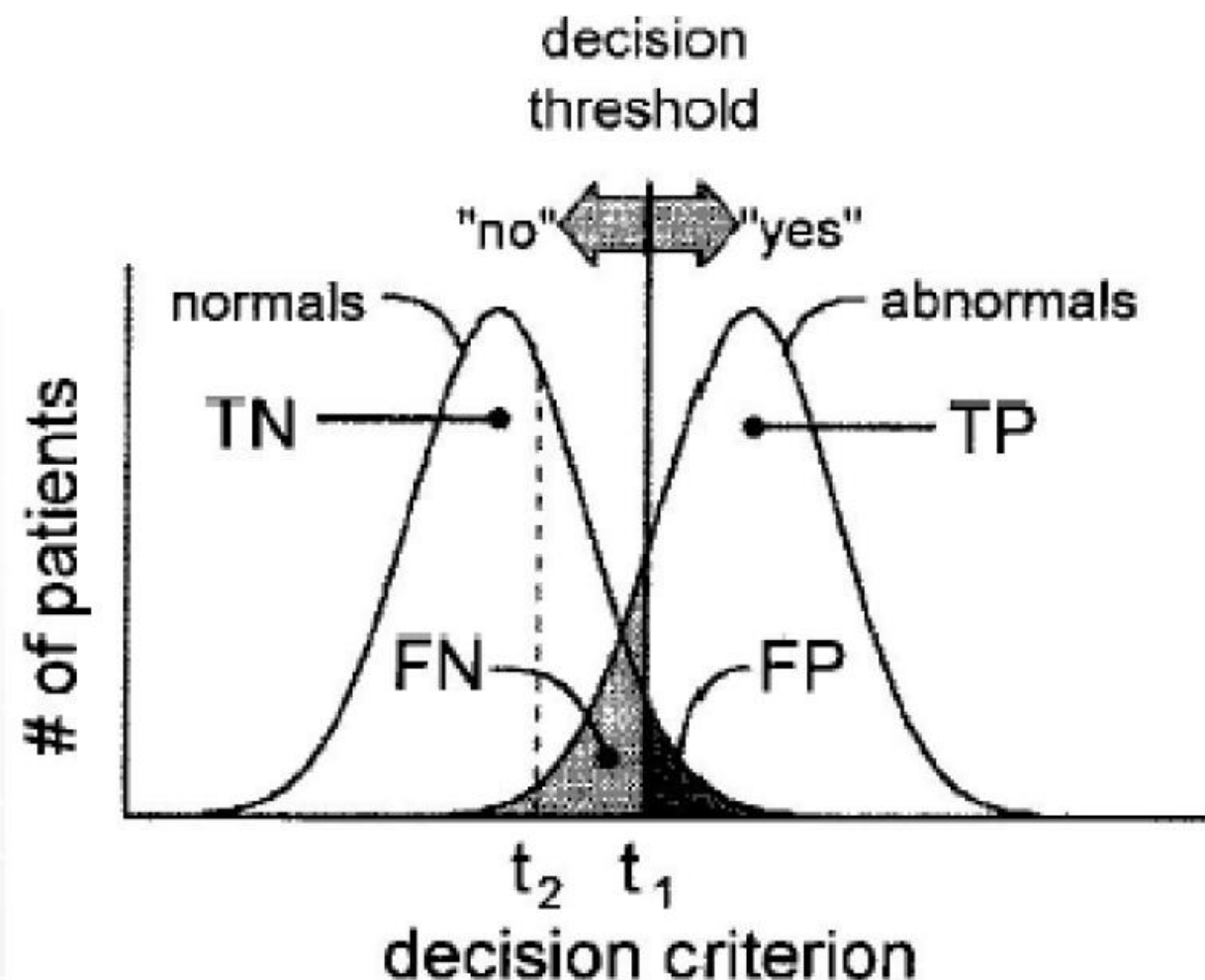
Two-class problems

- ▶ Good or bad?
- ▶ Present or absent?
- ▶ Diseased or not?

truth	classifier	evaluation
+	+	true positive
-	+	false positive
-	-	true negative
+	-	false negative

Two Class Problems

- **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.



THE 2 × 2 DECISION MATRIX

	Actually Abnormal	Actually Normal
Diagnosed as Abnormal	True Positive (TP)	False Positive (FP)
Diagnosed as Normal	False Negative (FN)	True Negative (TN)

Confusion Matrix for Binary Classification









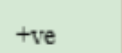
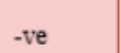
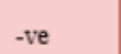

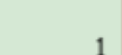


		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	560	60
	NEGATIVE	50	330

Suppose we had a classification dataset with 1000 data points. We fit a classifier on it and get the below confusion matrix:

The different values of the Confusion matrix would be as follows:

- True Positive (TP) = 560; meaning 560 positive class data points were correctly classified by the model
- True Negative (TN) = 330; meaning 330 negative class data points were correctly classified by the model
- False Positive (FP) = 60; meaning 60 negative class data points were incorrectly classified as belonging to the positive class by the model
- False Negative (FN) = 50; meaning 50 positive class data points were incorrectly classified as belonging to the negative class by the model

Confusion Matrix for Multi Class Classification

		ACTUAL VALUES		
				
PREDICTED VALUES		 +ve 1	 -ve 2	 -ve 3
		 -ve 4	 +ve 5	 -ve 6
		 -ve 7	 -ve 8	 +ve 9

Facebook

$$TP = \text{Cell}_1$$

$$FP = \text{Cell}_2 + \text{Cell}_3$$

$$TN = \text{Cell}_5 + \text{Cell}_6 + \text{Cell}_8 + \text{Cell}_9$$

$$FN = \text{Cell}_4 + \text{Cell}_7$$

Instagram

$$TP = \text{Cell}_5$$

$$FP = \text{Cell}_4 + \text{Cell}_6$$

$$TN = \text{Cell}_1 + \text{Cell}_3 + \text{Cell}_7 + \text{Cell}_9$$

$$FN = \text{Cell}_2 + \text{Cell}_8$$

Snapchat

$$TP = \text{Cell}_9$$

$$FP = \text{Cell}_7 + \text{Cell}_8$$

$$TN = \text{Cell}_1 + \text{Cell}_2 + \text{Cell}_4 + \text{Cell}_5$$

$$FN = \text{Cell}_3 + \text{Cell}_6$$

- ▶ TP = true positive = present and detected
- ▶ TN = true negative = not present and not detected
- ▶ FP = false positive = not present but detected
- ▶ FN = false negative = present but not detected

Confusion Matrix for Multi Class Classification

Class j output by the pattern recognition system

		'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
True object class i	'0'	97	0	0	0	0	0	1	0	0	1
	'1'	0	98	0	0	1	0	0	1	0	0
	'2'	0	0	96	1	0	1	0	1	0	0
	'3'	0	0	2	95	0	1	0	0	1	0
	'4'	0	0	0	0	98	0	0	0	0	2
	'5'	0	0	0	1	0	97	0	0	0	0
	'6'	1	0	0	0	0	1	98	0	0	0
	'7'	0	0	1	0	0	0	0	98	0	0
	'8'	0	0	0	1	0	0	1	0	96	1
	'9'	1	0	0	0	3	0	0	0	1	95

Confusion Matrix for Multi Class Classification

analysis for the classification of '3'

- ▶ TP = '3' present and '3' detected
- ▶ TN = not present and not detected
- ▶ FP = not present but detected
- ▶ FN = present but not detected

Class j output by the pattern recognition system

		'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
True object class i	'0'	97	0	0	0	0	0	1	0	0	1
	'1'	0	98	0	0	1	0	0	1	0	0
	'2'	0	0	96	1	0	1	0	1	0	0
	'3'	0	0	2	95	0	1	0	0	1	0
	'4'	0	0	0	0	98	0	0	0	0	2
	'5'	0	0	0	1	0	97	0	0	0	0
	'6'	1	0	0	0	0	1	98	0	0	0
	'7'	0	0	1	0	0	0	0	98	0	0
	'8'	0	0	0	1	0	0	1	0	96	1
	'9'	1	0	0	0	3	0	0	0	1	95

Confusion Matrix for Multi Class Classification

analysis for the classification of '3'

- ▶ TP = '3' present and '3' detected
- ▶ TN = not present and not detected
- ▶ FP = not present but detected
- ▶ FN = present but not detected

Class j output by the pattern recognition system

		'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
True object class i	'0'	97	0	0	0	0	0	1	0	0	1
	'1'	0	98	0	0	1	0	0	1	0	0
	'2'	0	0	96	1	0	1	0	1	0	0
	'3'	0	0	2	95	0	1	0	0	1	0
	'4'	0	0	0	0	98	0	0	0	0	2
	'5'	0	0	0	1	0	97	0	0	0	0
	'6'	1	0	0	0	0	1	98	0	0	0
	'7'	0	0	1	0	0	0	0	98	0	0
	'8'	0	0	0	1	0	0	1	0	96	1
	'9'	1	0	0	0	3	0	0	0	1	95

Confusion Matrix for Multi Class Classification

analysis for the classification of '3'

- ▶ **TP** = '3' present and '3' detected
- ▶ **TN** = not present and not detected
- ▶ **FP** = not present but detected
- ▶ **FN** = present but not detected

Class j output by the pattern recognition system

		'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
True object class i	'0'	97	0	0	0	0	0	1	0	0	1
	'1'	0	98	0	0	1	0	0	1	0	0
	'2'	0	0	96	1	0	1	0	1	0	0
	'3'	0	0	2	95	0	1	0	0	1	0
	'4'	0	0	0	0	98	0	0	0	0	2
	'5'	0	0	0	1	0	97	0	0	0	0
	'6'	1	0	0	0	0	1	98	0	0	0
	'7'	0	0	1	0	0	0	0	98	0	0
	'8'	0	0	0	1	0	0	1	0	96	1
	'9'	1	0	0	0	3	0	0	0	1	95

Confusion Matrix for Multi Class Classification

analysis for the classification of '3'

▶ TP = '3' present and '3' detected

▶ TN = not present and not detected

▶ FP = not present but detected

▶ FN = present but not detected

Class j output by the pattern recognition system

	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
'0'	97	0	0	0	0	0	1	0	0	1
'1'	0	98	0	0	1	0	0	1	0	0
'2'	0	0	96	1	0	1	0	1	0	0
'3'	0	0	2	95	0	1	0	0	1	0
'4'	0	0	0	0	98	0	0	0	0	2
'5'	0	0	0	1	0	97	0	0	0	0
'6'	1	0	0	0	0	1	98	0	0	0
'7'	0	0	1	0	0	0	0	98	0	0
'8'	0	0	0	1	0	0	1	0	96	1
'9'	1	0	0	0	3	0	0	0	1	95

True object class

i

Accuracy

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions..

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}$$

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

Precision

- **Precision** is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.
- Take it as to find out '*how much the model is right when it says it is right*'.

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall

- 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.
- A measure to see how separate the classes are.

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

F1 Score

- F1 score is calculated by combining the precision and recall metrics; the common approach for combining these metrics is known as the f-score.

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

- 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 ROC curve can help.

$$F_{\beta} = (1 + \beta^2) \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

- $\beta < 1$ focuses more on precision while $\beta > 1$ focuses more on recall.

ROC (Receiver Operating Characteristics) curve

- ▶ We are given a number of test cases for which we know the "truth."
- ▶ For a single setting, we can calculate (test our method for) TP, TN, FP, and FN.
 - TP = true positive = present and detected
 - TN = true negative = not present and not detected
 - FP = false positive = not present but detected
 - FN = false negative = present but not detected
- ▶ $TP + TN + FP + FN = \#$ of normals and abnormals in our study population.

ROC analysis

► True Positive Fraction

- $TPF = TP / (TP + FN)$
- also called *sensitivity*
- true abnormalities called abnormal by the observer

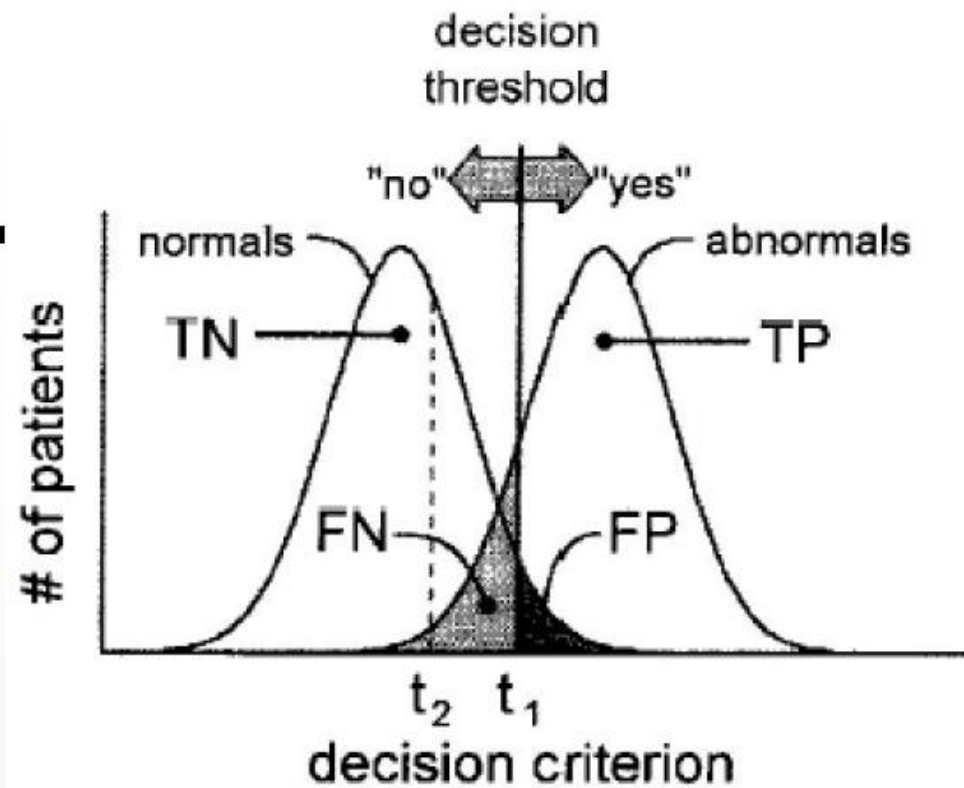
► False Positive Fraction

- $FPF = FP / (FP + TN)$

► *Specificity* = $TN / (TN + FP)$

- True normals called normal by the observer

- $FPF = 1 - \text{specificity}$



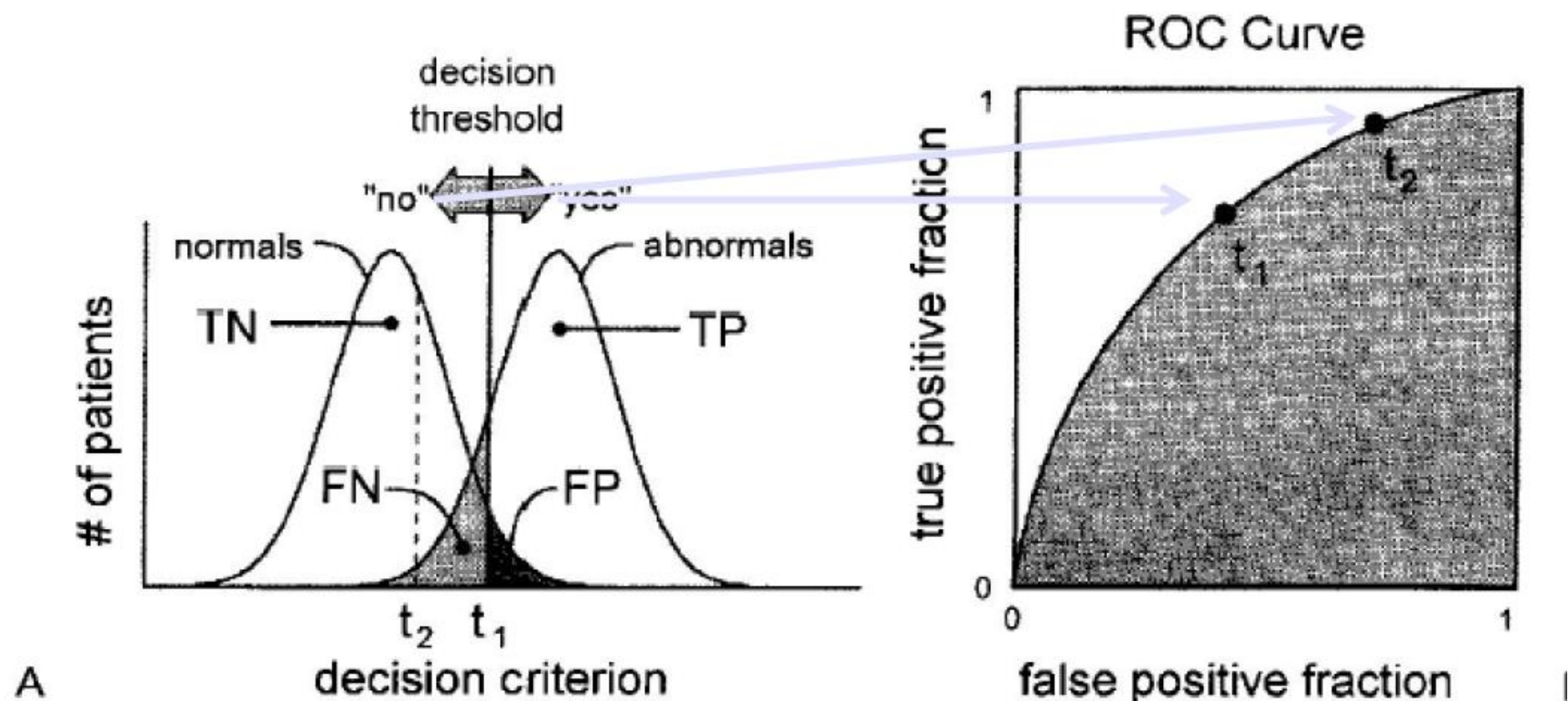


FIGURE 10-38. A: The basics of signal detection theory. The study population includes normals and abnormals, and they are histogrammed according to the *decision criterion* (see text). The diagnostician applies a decision threshold, and calls cases to the right of it abnormal ("yes") and to the left of it negative ("no"). The true positive, true negative, false positive, and false negatives can then be computed. These values are used to calculate the true-positive fraction and the false-positive fraction, which are plotted on the ROC curve **(B)**. The decision threshold at each point (t_1) gives rise to one point on the ROC curve (t_1). Shifting the decision threshold toward the left (to the *dashed line* marked t_2 on the left figure) will increase the true-positive fraction but will also increase the false-positive fraction.

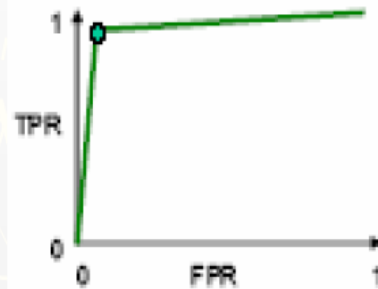
ROC (Receiver Operating Characteristics) curve

- ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values.

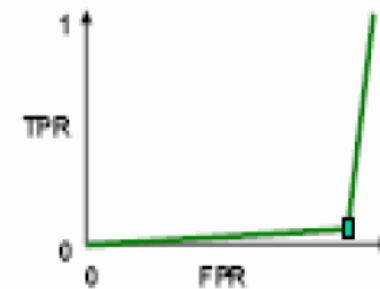
$$\text{True Positive Rate (TPR)} = \text{RECALL} = \frac{TP}{TP+FN}$$

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

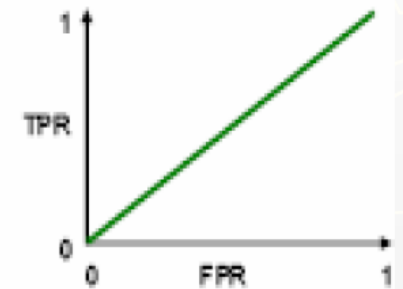
- ROC space: good and bad classifiers.



- Good classifier.
 - High TPR.
 - Low FPR.



- Bad classifier.
 - Low TPR.
 - High FPR.



- Bad classifier (real picture).

Credits

- <https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15>
- <https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826>
- https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#h2_8