Measures of classification performance

Alberto Paccanaro

EMAp – FGV

www.paccanarolab.org

Some images in these slides are from (or adapted from):
A. Geron, Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow, O'Reilly, 2020

The MINST dataset

70000 images

Each image: 28x28 features

Each feature can take values in [0, 255]

A binary classification problem

We want to build a "5-detector" capable of distinguishing between just two classes,

- 5 (C_1 or "positives", P)
- not-5 (C_0 or "negatives", N).

Let us assume we have built our classifier.

How can I evaluate its performance on the test set?

Remember: we are always interested in generalization, i.e. how well it will perform on unseen data (that is the test set)

Confusion matrix

Classifier output: probability of being a 1

0 0.5 1

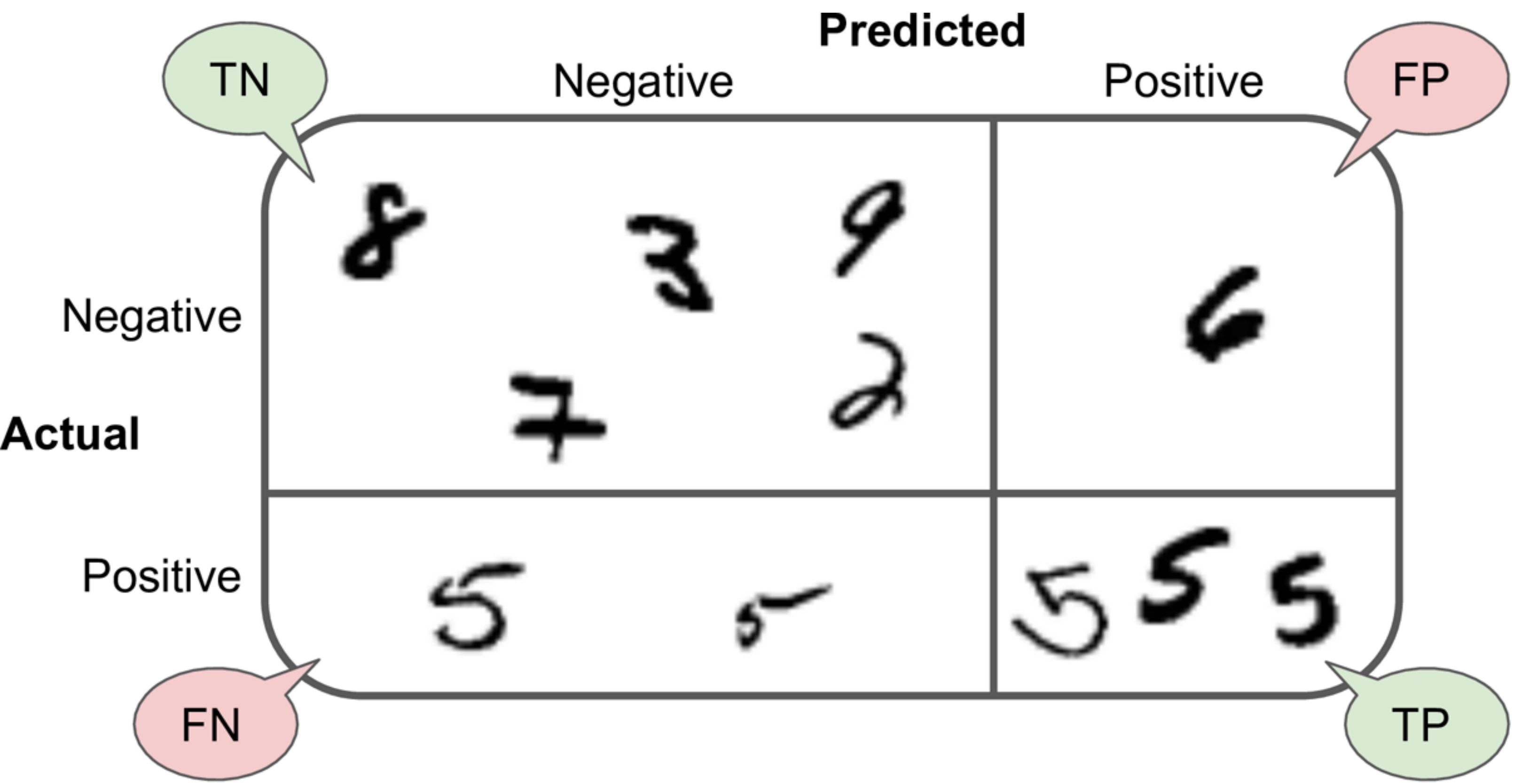
predicted

actual		
	TN	FP *Type 1 error (false alarm)*
	FN *Type 2 error (miss)*	TP

Confusion Matrix

predicted

actual	TN	FP
	FN	TP



Measures

		predicted	
actual		TN	FP
		FN	TP

Accuracy: percentage of correct predictions

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

Precision: the accuracy of the correct prediction *(how precise am I? what is the % of correct, out of all those that I predict as P?)*

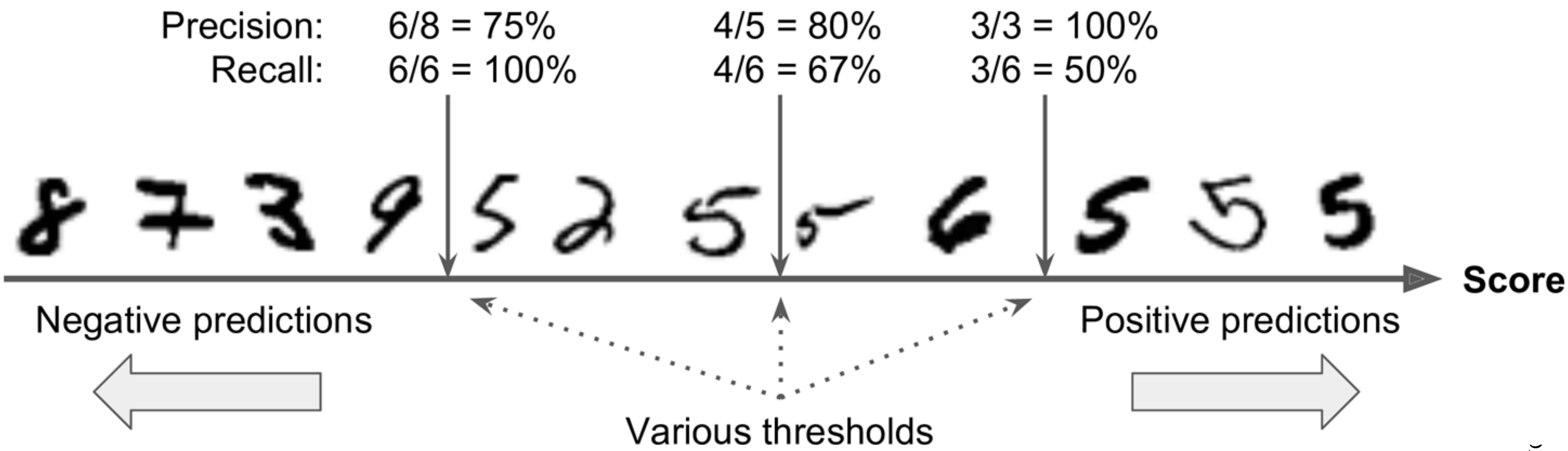
$$\frac{TP}{TP + FP}$$

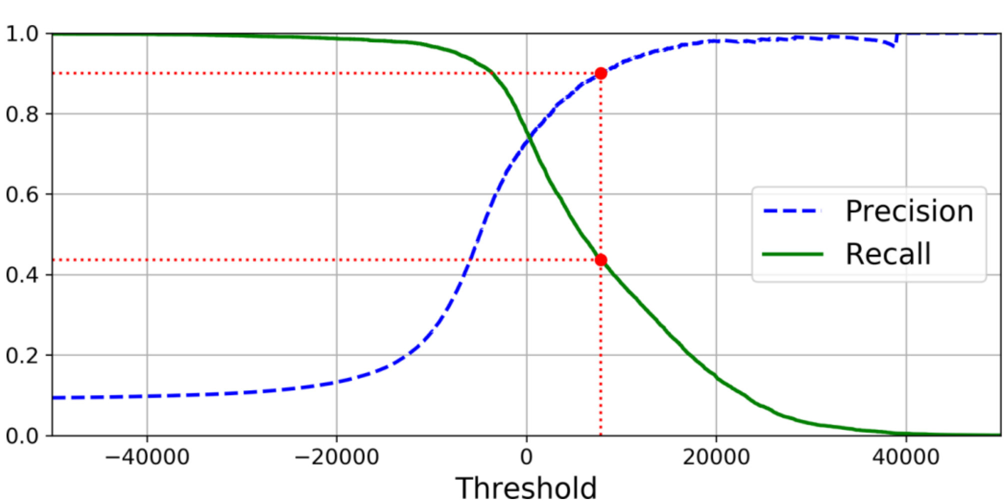
Recall: percentage of positive instances that are correctly predicted *(how much do I "cover" the P? what is the % of correct, out of all those that are P?)*

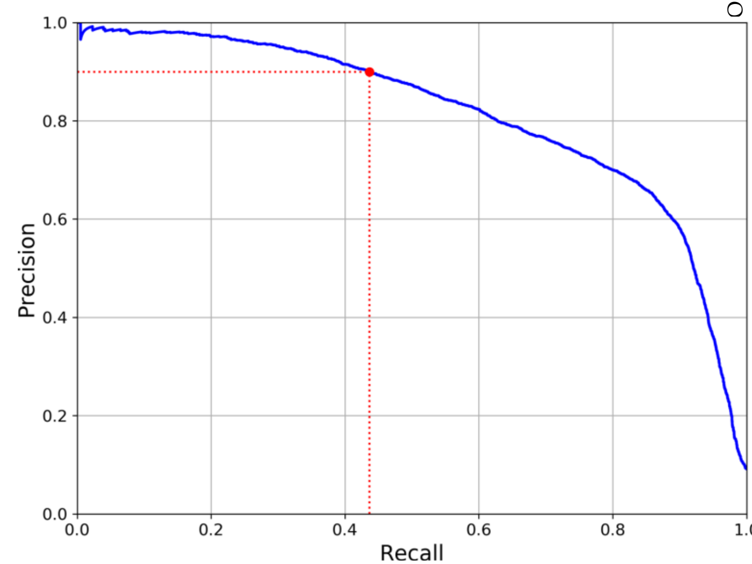
$$\frac{TP}{P} = \frac{TP}{TP + FN}$$

Unfortunately: increasing precision reduces recall, and vice versa…

Precision/recall trade-off







F_1 measure

$$\mathrm{HM}(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

The harmonic mean of precision and recall:

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

It is higher when precision and recall are both high.

These considerations help us to pick a threshold.

Can I use these ideas to compare the performance of classifiers?

The receiver operating characteristic (ROC) curve

	predicted	
actual	TN	FP
	FN	TP

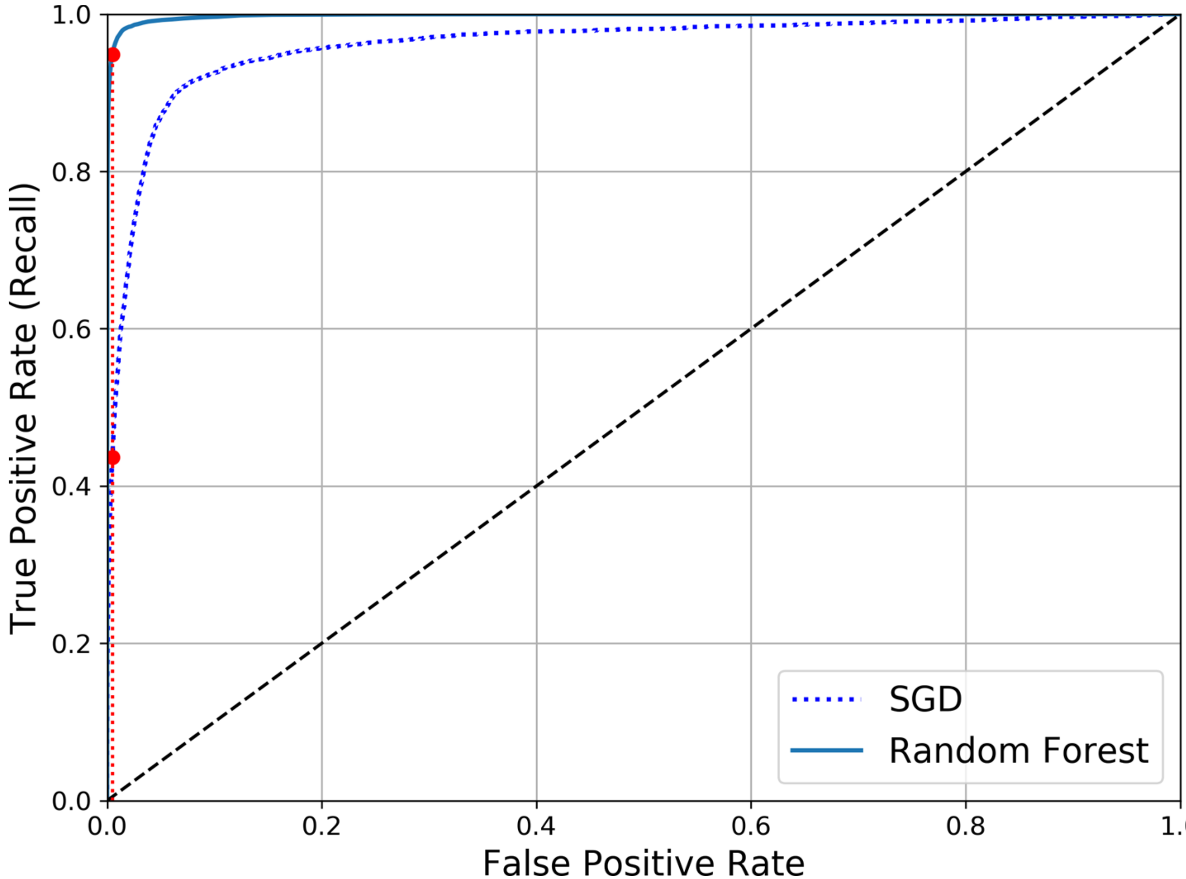
True positive rate (= recall, sensitivity): % of positive instances that are correctly predicted

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

False positive rate: the % of negative instances that are incorrectly predicted.

$$1 - specificity =$$

$$1 - \frac{TN}{N} = \frac{FP}{FP + TN}$$



One way to compare classifiers is to measure the *area under the curve* (AUC).

A purely random classifier will have a ROC AUC equal to 0.5.