

# Introduction to Machine Learning

## Evaluation: Measures for Binary Classification: ROC Measures

### Learning goals

- Understand why accuracy is not an optimal performance measure for imbalanced labels
- Understand the different measures computable from a confusion matrix
- Be aware that each of these measures has a variety of names

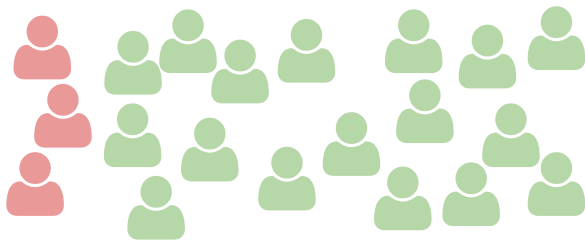
		True Class $y$		
		+	-	
Pred.	+	TP	FP	PPV = $\frac{TP}{TP+FP}$
$\hat{y}$	-	FN	TN	NPV = $\frac{TN}{FN+TN}$
		TPR = $\frac{TP}{TP+FN}$	TNR = $\frac{TN}{FP+TN}$	Accuracy = $\frac{TP+TN}{TOTAL}$

# CLASS IMBALANCE

- Assume a binary classifier diagnoses a serious medical condition.
- Label distribution is often **imbalanced**, i.e, not many people have the disease.
- Evaluating on mce is often inappropriate for scenarios with imbalanced labels:
  - Assume that only 0.5 % have the disease.
  - Always predicting “no disease” has an mce of 0.5 %, corresponding to very high accuracy.
  - This sends all sick patients home → bad system
- This problem is known as the **accuracy paradox**.

# CLASS IMBALANCE

Classifying all observations as “no disease” (green) yields top accuracy simply because the “disease” occurs so rarely → accuracy paradox.



# IMBALANCED COSTS

- Another point of view is **imbalanced costs**.
- In our example, classifying a sick patient as healthy should incur a much higher cost than classifying a healthy patient as sick.
- The costs depend a lot on what happens next: we can well assume that our system is some type of screening filter, and often the next step after labeling someone as sick might be a more invasive, expensive, but also more reliable test for the disease.
- Erroneously subjecting someone to this step is undesirable (psychological, economic, medical expense), but sending someone home to get worse or die seems much more so.
- Such situations not only arise under label imbalance, but also when costs differ (even though classes might be balanced).
- We could see this as imbalanced costs of misclassification, rather than imbalanced labels; both situations are tightly connected.

# IMBALANCED COSTS

**Imbalanced costs:** classifying incorrectly as “no disease” incurs very high cost.



- Problem: if we were able to specify costs precisely, we could evaluate or even optimize on them.
- This important subfield of ML is called **cost-sensitive learning**, which we will not cover in this lecture unit.
- Unfortunately, users find it notoriously hard to come up with precise cost figures in imbalanced scenarios.
- Evaluating “from different perspectives”, with multiple metrics, often helps to get a first impression of system quality.

# ROC ANALYSIS

- **ROC analysis** is a subfield of ML which studies the evaluation of binary prediction systems.
- ROC stands for “receiver operating characteristics” and was initially developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields – still has the funny name.



http:

[//media.iwm.org.uk/iwm/mediaLib//39/media-39665/large.jpg](http://media.iwm.org.uk/iwm/mediaLib//39/media-39665/large.jpg)

# LABELS: ROC METRICS

From the confusion matrix (binary case), we can calculate "ROC" metrics.

		True Class $y$		
		+	-	
Pred.	+	TP	FP	$\rho_{PPV} = \frac{TP}{TP+FP}$
$\hat{y}$	-	FN	TN	$\rho_{NPV} = \frac{TN}{FN+TN}$
		$\rho_{TPR} = \frac{TP}{TP+FN}$	$\rho_{TNR} = \frac{TN}{FP+TN}$	$\rho_{ACC} = \frac{TP+TN}{TOTAL}$

- True positive rate  $\rho_{TPR}$ : how many of the true 1s did we predict as 1?
- True Negative rate  $\rho_{TNR}$ : how many of the true 0s did we predict as 0?
- Positive predictive value  $\rho_{PPV}$ : if we predict 1, how likely is it a true 1?
- Negative predictive value  $\rho_{NPV}$ : if we predict 0, how likely is it a true 0?
- Accuracy  $\rho_{ACC}$ : how many instances did we predict correctly?

# LABELS: ROC METRICS

Example:

		Actual Class $y$			
		Positive	Negative		
$\hat{y}$ Pred.	Positive	<b>True Positive</b> (TP) = 20	<b>False Positive</b> (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = <b>10%</b>	
	Negative	<b>False Negative</b> (FN) = 10	<b>True Negative</b> (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) $\approx$ <b>99.5%</b>	
		True Positive Rate = TP / (TP + FN) = 20 / (20 + 10) $\approx$ <b>67%</b>	True Negative Rate = TN / (FP + TN) = 1820 / (180 + 1820) = <b>91%</b>		

[//en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)

https:



# MORE METRICS AND ALTERNATIVE TERMINOLOGY

Unfortunately, for many concepts in ROC, 2-3 different terms exist.

		True condition				
		Total population	Condition positive	Condition negative		
Predicted condition				Prevalence $= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$	
	Predicted condition positive	True positive, Power	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$	
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$	F <sub>1</sub> score = $\frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$		

► Clickable version/picture source

► Interactive diagram

# LABELS: $F_1$ MEASURE

- It is difficult to achieve high **positive predictive value** and high **true positive rate** simultaneously.
- A classifier predicting more positive will be more sensitive (higher  $\rho_{TPR}$ ), but it will also tend to give more *false* positives (lower  $\rho_{TNR}$ , lower  $\rho_{PPV}$ ).
- A classifier that predicts more negatives will be more precise (higher  $\rho_{PPV}$ ), but it will also produce more *false* negatives (lower  $\rho_{TPR}$ ).

The  $F_1$  **score** balances two conflicting goals:

- 1 Maximizing positive predictive value
- 2 Maximizing true positive rate

$\rho_{F_1}$  is the harmonic mean of  $\rho_{PPV}$  and  $\rho_{TPR}$ :

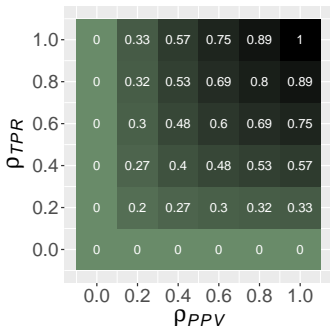
$$\rho_{F_1} = 2 \cdot \frac{\rho_{PPV} \cdot \rho_{TPR}}{\rho_{PPV} + \rho_{TPR}}$$

Note that this measure still does not account for the number of true negatives.

# LABELS: $F_1$ MEASURE

$F_1$  score for different combinations of  $\rho_{PPV}$  &  $\rho_{TPR}$ .

→ Tends more towards the lower of the two combined values.



- A model with  $\rho_{TPR} = 0$  (no positive instance predicted as positive) or  $\rho_{PPV} = 0$  (no true positives among the predicted) has  $\rho_{F_1} = 0$ .
- Always predicting “negative”:  $\rho_{F_1} = 0$ .
- Always predicting “positive”:  
$$\rho_{F_1} = 2 \cdot \rho_{PPV} / (\rho_{PPV} + 1) = 2 \cdot n_+ / (n_+ + n),$$
which will be small when the size of the positive class  $n_+$  is small.

# WHICH METRIC TO USE?

- As we have seen, there is a plethora of methods.  
→ This leaves practitioners with the question of which to use.
- Consider a small benchmark study.
  - We let  $k$ -NN, logistic regression, a classification tree, and a random forest compete on classifying the `credit risk` data.
  - The data consist of 1000 observations of borrowers' financial situation and their creditworthiness (good/bad) as target.
  - Predicted probabilities are thresholded at 0.5 for the positive class.
  - Depending on the metric we use, learners are ranked differently according to performance (value of respective performance measure in parentheses):

metric	TPR	2 (0.8777)	3 (0.8647)	1 (0.9257)	4 (0.8357)
	TNR	4 (0.3764)	2 (0.4797)	3 (0.4072)	1 (0.4911)
	PPV	4 (0.7665)	1 (0.7947)	3 (0.7842)	2 (0.7925)
	F1	3 (0.8179)	2 (0.8279)	1 (0.8488)	4 (0.8130)
	AUC	4 (0.7092)	2 (0.7731)	1 (0.7902)	3 (0.7293)
	ACC	4 (0.7270)	2 (0.7490)	1 (0.7700)	3 (0.7320)
		k-NN	logistic regression	random forest	CART
		learner			

# WHICH METRIC TO USE?

- We need not expect overly large discrepancies in general, but neither will we always see an unambiguous picture.
- Different metrics emphasize different aspects of performance.  
→ The choice should be made in the domain context.
- For practitioners it is vital to understand what should be evaluated exactly, and which measure is appropriate.
  - Regarding credit risk, for instance, defaults are to be avoided, but not at all cost.
  - The bank must undertake a certain risk to remain profitable, so a more balanced measure such as the  $F_1$  score might be in order.
  - On the other hand, a system detecting weapons at an airport should be able to achieve very high true positive rates, even if this comes at the expense of some false alarms.