Table of contents

Measures of classification accuracy: a Sudoku exercise

- An algorithm classifies patients as having a breast tumor, or not.
- 10 patients were classified as having a breast tumor, and 10 were classified as healthy. In reality 12 patients had a tumor.
- Do know enough to compute the accuracy of the classifier?

	Classified healthy	Classified diseased	Total
Healthy			
Diseased			**12**
Total	**10**	**10**	**20**

Measures of classification accuracy: a Sudoku exercise

- An algorithm classifies patients as having a breast tumor, or not.
- 10 patients were classified as having a breast tumor, and 10 were classified as healthy. In reality 12 patients had a tumor.
- Do know enough to compute the accuracy of the classifier?

	Classified healthy	Classified diseased	Total
Healthy	8	0	**8**
Diseased	2	10	**12**
Total	**10**	**10**	**20**

Accuracy = 1-2/20 = 90%

Measures of classification accuracy: a Sudoku exercise

- An algorithm classifies patients as having a breast tumor, or not.
- 10 patients were classified as having a breast tumor, and 10 were classified as healthy. In reality 12 patients had a tumor.
- Do know enough to compute the accuracy of the classifier?

	Classified healthy	Classified diseased	Total
Healthy	0	8	**8**
Diseased	10	2	**12**
Total	**10**	**10**	**20**

Accuracy = 1-18/20 = 10%

Measures of classification accuracy: a Sudoku exercise

- Binary classification is fundamentally a two-dimensional optimization problem.
- False negatives are positive (diseased) examples predicted as “negative” (healthy). False positives are negative examples predicted as positive.

	Classified healthy	**Classified diseased**	**Total**
Healthy	True Negatives	False Positives	**Negatives**
Diseased	False Negatives	True Positives	**Positives**
Total	**Predicted negative**	**Predicted positive**	**Sample size**

Accuracy = (TP + TN)/N

Error Rate = (FP + FN)/N

Accuracy = 1 - Error Rate

Fairness invites us to consider even more dimensions

	Classified healthy	Classified diseased	Total
Healthy	7	1	**8**
Diseased	2	10	**12**
Total	**10**	**10**	**20**

Error rate for men = ¼ = 25%

Error rate for women = 3/16 = 18.75%

Overall error rate = 4/20 = 20%

	Predicted healthy	Predicted diseased	Total per group	Total
Healthy Women	4	1	5	8
Healthy Men	3	0	3	
Diseased Women	2	9	11	12
Diseased Men	1	0	1	
Total	10	10	20	

Tradeoff between false negatives and false positives

- Most classifiers do not actually produce a label directly, but rather a score $s(X)$ on a given object. Sometimes this falls in [0,1] so it acts like a probability, $P(y=1|X) = s(X)$, but for our purposes here this is not necessary.

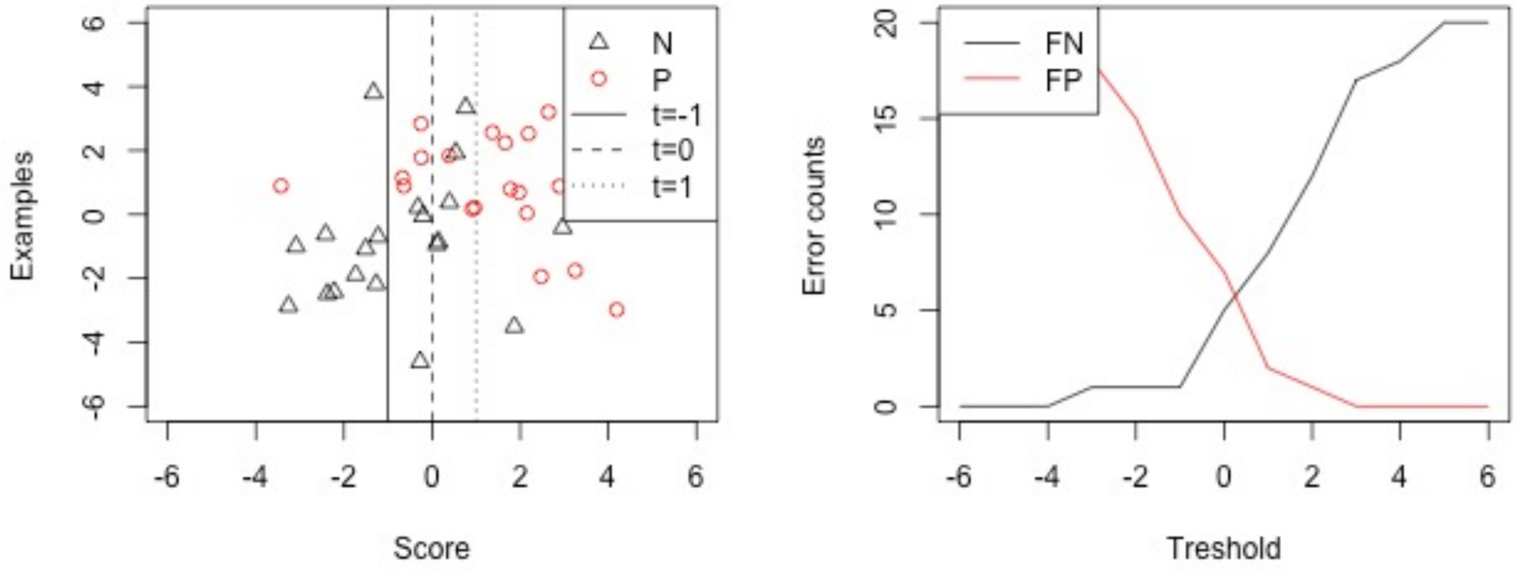
Tradeoff between false negatives and false positives

- To then assign a label on X, we can threshold by a value t, and let the label be 1 if s(x) > t. For very large values of t, most examples will be assigned a negative value, which makes the probability of false negatives higher. For very small values of t, most examples will be assigned a positive label.

$$\hat{y}_i = \begin{cases} 1, & \text{if } s(x_i) > t \\ 0, & \text{otherwise.} \end{cases}$$

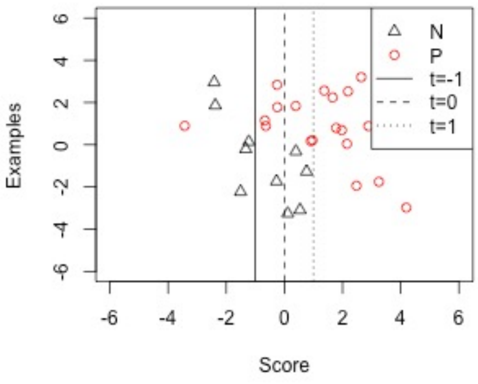
Note that when s(X) is really a probability it might seem natural to choose t = 0.5, but if we care about FNs, say, more than about FPs, we can still modify t.

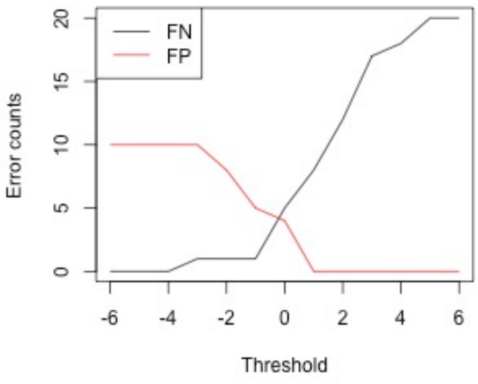
Tradeoff between false negatives and false positives

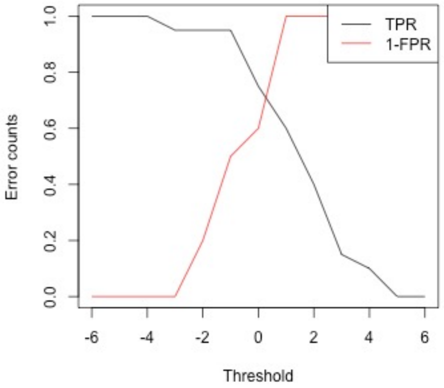


True positive rates and false positive rates

- TPR = TP/(TP+FN) also known as sensitivity and is 1-FNR
- TNR = TN/(TN+FP) also known as specificity and is 1-FPR







Error parity can be broken down into four different tests.

- Error rate is equal across both groups (sometimes known as Error Parity)
- FNR is equal across both groups (sometimes known as Equal Opportunity)
- FPR is equal across both groups (sometimes known as Predictive Equality)
- Both FPR and FNR are equal across both group (sometimes known as Equalized Odds)

	Predicted healthy	**Predicted diseased**	**Total per group**	**Total**
Healthy Women	4	$FP_w = 1$	$N_w = 5$	8
Healthy Men	3	$FP_m = 0$	$N_m = 3$	
Diseased Women	$FN_w = 2$	9	$P_w = 11$	12
Diseased Men	$FN_m = 1$	0	$P_m = 1$	
Total	10	10	20	

- $FPR_w = FP_w/N_w = 1/5 = 0.2$
- $FPR_m = FP_m/N_m = 0/3 = 0.0$
- $FNR_w = FN_w/P_w = 2/11 = 0.18$
- $FNR_m = FN_m/P_m = 1/1 = 1.0$

This classifier certainly does not satisfy predictive equality and is also violating equal opportunity. No error parity holds.

Who is being disadvantaged?

Wait – why "equal opportunity"?

- FNR parity is sometimes called "equal opportunity". This is under the assumption that a positive label confers an advantage, for example, it represents:
  - The decision to grant a loan
  - The decision to admit someone to a university
  - …
- In such cases, FNR parity, or, equivalently, TPR parity ensures that, say, that the percentage of men that are truly creditworthy and are given a loan equals that for women.
- Generally, some care is needed when defining what is a positive and negative label.
- In the medical setting, in many cases analogy still holds, as a *diagnosis is an opportunity to treat.*

Imperial College London

Summary

- Errors come in two flavors: false positives and false negatives. This is captured in a confusion matrix.
- Classifiers are able to trade them off each other, depending on the relative cost of misclassification.
- To account for imbalanced datasets, we typically use sensitivity and specificity to reason about these tradeoffs, or, equivalently, false positive rates and false negative rates.
- Ensuring equal accuracy across both groups is one possible fairness metric. But a more comprehensive one is to reason separately about false positive and false negative rates.
- We have therefore introduced four different fairness metrics:
  - Error parity (Error Rate)
  - Equal opportunity (FNR)
  - Predictive equality (FPR)
  - Equalized odds (FNR and FPR)