Fairness, Part III

Giulia Fanti

Fall 2019

Based in part on slides by Anupam Datta

Administrative

- HW4 due Nov. 22 (2.5 weeks from now)
  - Fairness + anonymous communication

- Recitation on Friday (Sruti)
  - Creating fair classifiers

- Feedback included on presentations, please read!

In-class Quiz

- On Canvas

Last time

- Group fairness vs Individual Fairness
  - When does one imply the other?

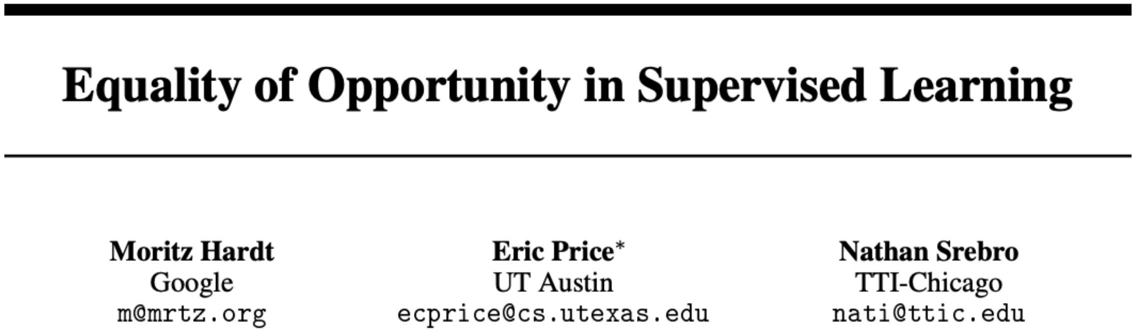
- Equalized odds vs equal opportunity
  - How do we transform an unfair classifier into a fair one?
  - What is the effect of fairness on classifier accuracy?

Today

- Review of equalized odds vs equal opportunity
  - Revisit geometric interpretation
- Disparate impact
  - Metric for measuring
  - How to prevent it
- Overview of fairness techniques & how they relate to each other
- Wrap up Unit 2

Last time: Another definition of fair classifiers

- NeurIPS 2016

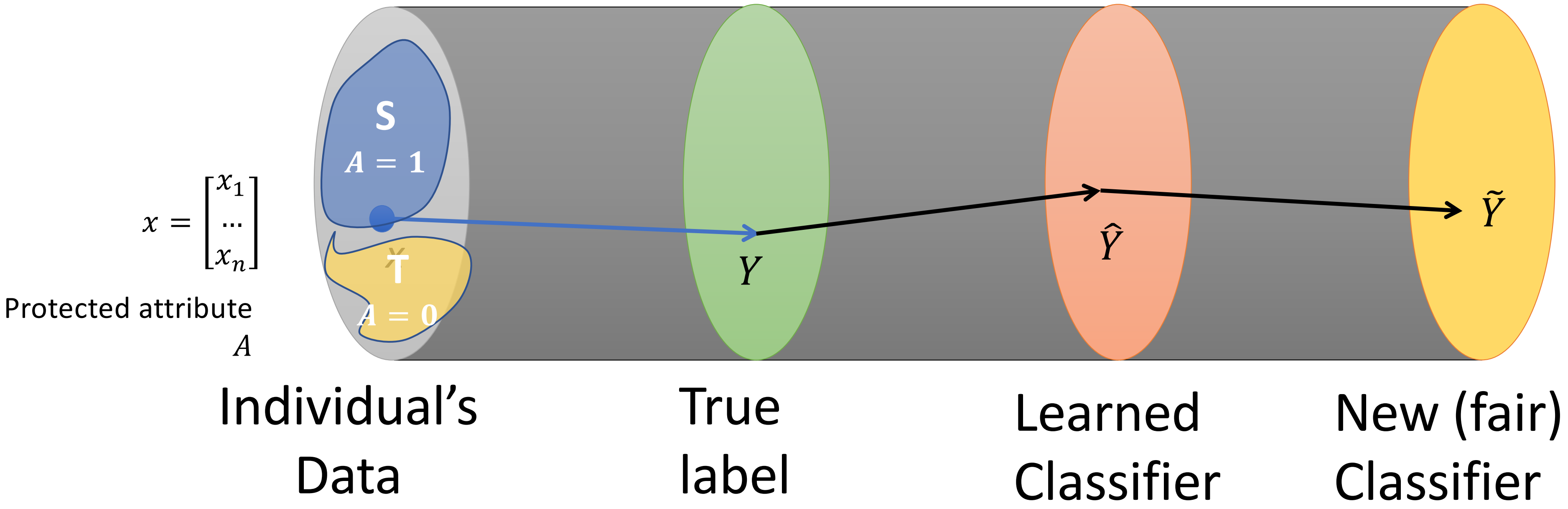


Equality of Opportunity in Supervised Learning

Moritz Hardt
Google
m@mrtz.org

Eric Price*
UT Austin
ecprice@cs.utexas.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu


S
$A = 1$
T
$A = 0$
$x = \begin{bmatrix} x_1 \\ \cdots \\ x_n \end{bmatrix}$
Protected attribute
A
Y
$\hat{Y}$
$\tilde{Y}$
Individual's Data
True label
Learned Classifier
New (fair) Classifier

Equalized odds

- Consider binary classifiers
- We say a classifier $\hat{Y}$ has equalized odds if for all true labels y,

$$P[\hat{Y} = 1 | A = 0, Y = y] = P[\hat{Y} = 1 | A = 1, Y = y]$$

Q: How would this definition look if we only wanted to enforce group fairness?

A: $P[\hat{Y} = 1 | A = 0] = P[\hat{Y} = 1 | A = 1]$

Exercise

Come up with an example classifier that exhibits group fairness but not equalized odds.

Equalized Odds: $P\left[\hat{Y} = 1 | A = 0, Y = y\right] = P\left[\hat{Y} = 1 | A = 1, Y = y\right]$

Group Fairness: $P\left[\hat{Y} = 1 | A = 0\right] = P\left[\hat{Y} = 1 | A = 1\right]$

Equal opportunity

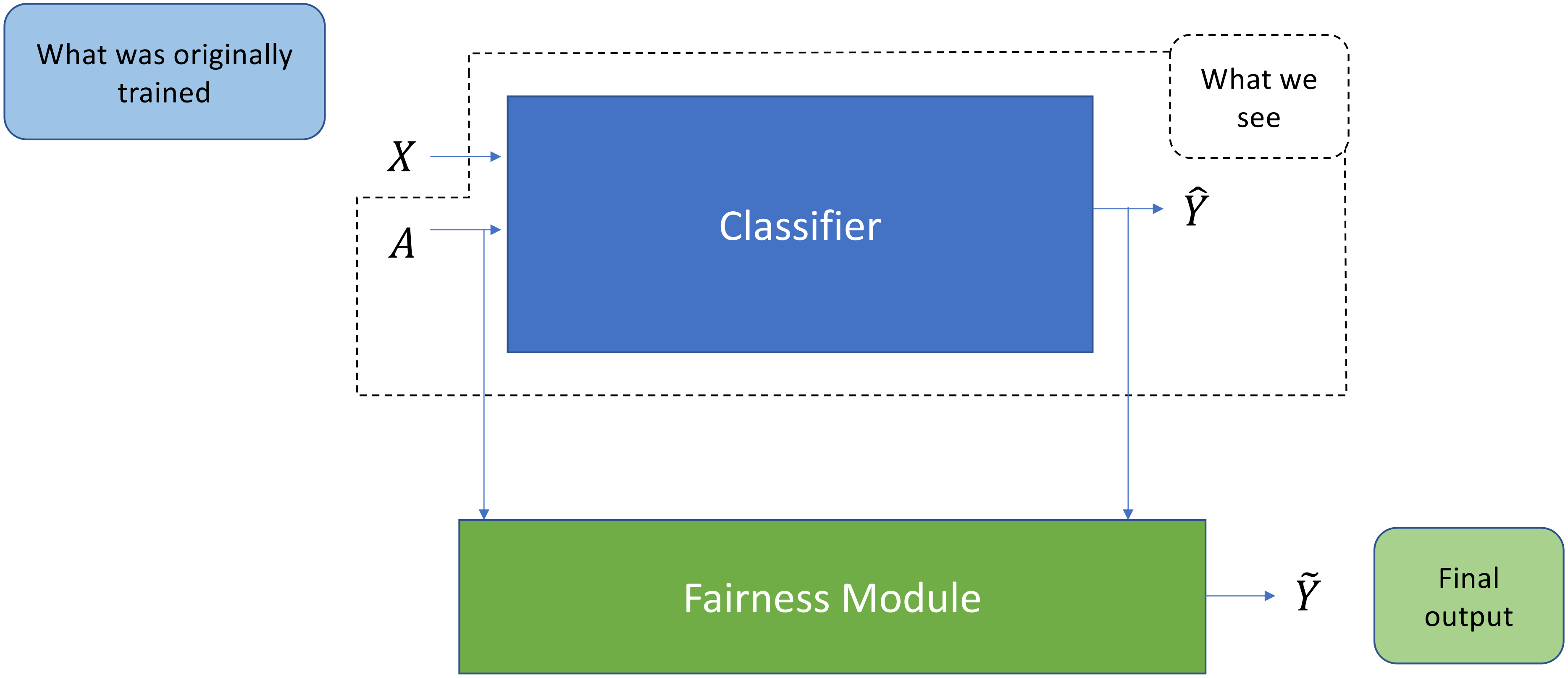
- Suppose $Y = 1$ is the desirable outcome
  - E.g., getting a loan
- We say a classifier $\hat{Y}$ has equal opportunity if

$$P\left[\hat{Y} = 1 | A = 0, Y = 1\right] = P\left[\hat{Y} = 1 | A = 1, Y = 1\right]$$

Interpretation: True positive rate is the same for both classes

Weaker notion of fairness → can enable better utility

So how do we enforce fairness?



How is the fairness module defined?

Protected attribute A

Predicted Label $\hat{Y}$

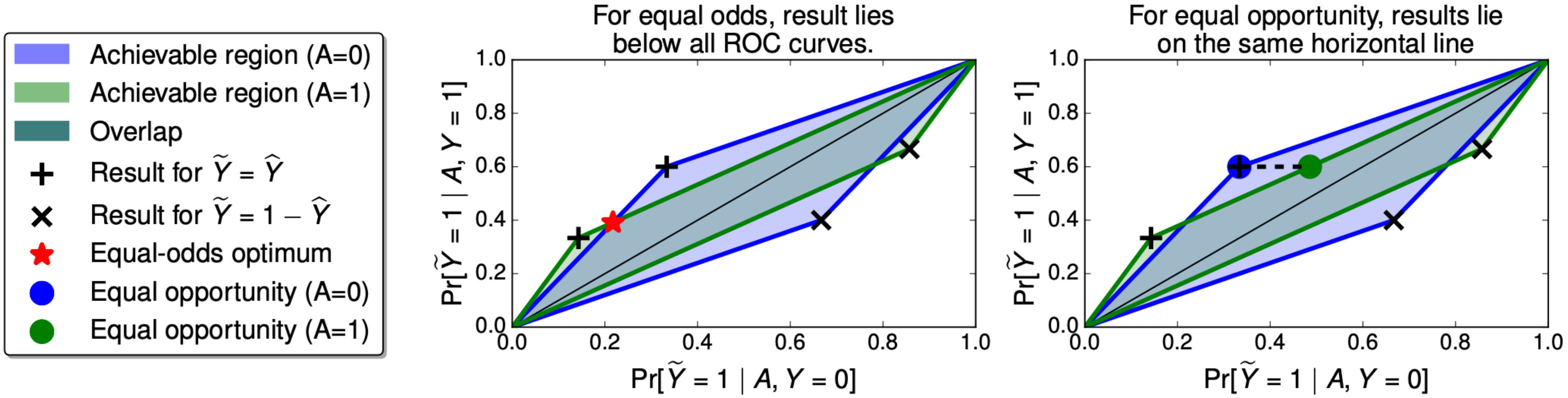
	0	1
0	$p_{00} = P(\tilde{Y} = 1 \vert A = 0, \hat{Y} = 0)$	$p_{01} = P(\tilde{Y} = 1 \vert A = 1, \hat{Y} = 0)$
1	$p_{10} = P(\tilde{Y} = 1 \vert A = 0, \hat{Y} = 1)$	$p_{11} = P(\tilde{Y} = 1 \vert A = 1, \hat{Y} = 1)$

Note: These four numbers are decoupled!

Geometric Interpretation via ROC curves

- False positive rate for $A = 0$:

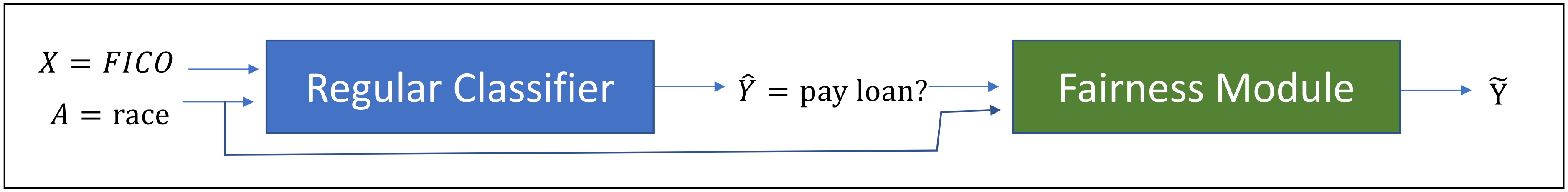
$$P\left(\tilde{Y} = 1 \middle| A = 0, Y = 0\right) = P\left(\hat{Y} = 1 \middle| A = 0, Y = 0\right) p_{10} + P\left(\hat{Y} = 0 \middle| A = 0, Y = 0\right) p_{00}$$



For equal odds, result lies below all ROC curves.

For equal opportunity, results lie on the same horizontal line

Case study: FICO Scores



Baseline fairness techniques

- **Max profit** – no fairness constraint (output $\hat{Y}$)
- **Race blind** – uses same FICO threshold for all groups
- **Group fairness** – picks for each group a threshold such that the fraction of group members that qualify for loans is the same
- **Equal opportunity** – picks a threshold for each group s.t. fraction of non-defaulting group members is the same
- **Equalized odds** –fraction of non-defaulters that qualify for loans and the fraction of defaulters that qualify for loans constant across groups

Profit Results

Method	Profit (% relative to max profit)
Max profit	100
Race blind	99.3
Equal opportunity	92.8
Equalized odds	80.2
Group fairness (demographic parity)	69.8

To summarize…

- Equalized odds and equal opportunity are more practical than individual fairness

- Equal opportunity doesn’t hurt utility too much

- Still requires access to original output labels to evaluate fairness
  - This is also clear from the definition of equalized odds/equal opportunity

Yet another definition, this time based on “Disparate Impact”

Certifying and removing disparate impact*

Michael Feldman
Haverford College

Sorelle A. Friedler
Haverford College

John Moeller
University of Utah

Carlos Scheidegger
University of Arizona

Suresh Venkatasubramanian†
University of Utah

Disparate impact

- Griggs v. Duke Power Co (1971)
  - US Supreme Court: business hiring decision illegal if it results in disparate impact by race

- Leading technique today for determining unintended discrimination in court system
- … but how do we define this?

Formal definition (80% rule)

Suppose $A = 1$ is the minority (protected) class, as before. We say mechanism M has disparate impact if

$$\frac{Pr(Y = 1|A = 1)}{Pr(Y = 1|A = 0)} \leq \tau = 0.8$$

Q: What does this look like?

A: Group fairness

Q: More or less stringent?

A: Less stringent, group fairness requires ratio to be 1

How do you estimate this?

- Can we predict sensitive attribute A from non-sensitive attributes X?
- Balanced error rate:

$$BER(f(X), A) = \frac{P(f(X) = 0|A = 1) + P(f(X) = 1|A = 0)}{2}$$

A dataset is ϵ-predictable iff there exists classification algorithm $f(X)$,

$$BER(f(X), A) \leq \epsilon$$

What does this mean?

How does all this relate to disparate impact?

- **Theorem:** A dataset with fraction β minority samples receiving outcome $Y = 1$ has disparate impact iff if is $\left(\frac{1}{2} - \frac{\beta}{8}\right)$-predictable.

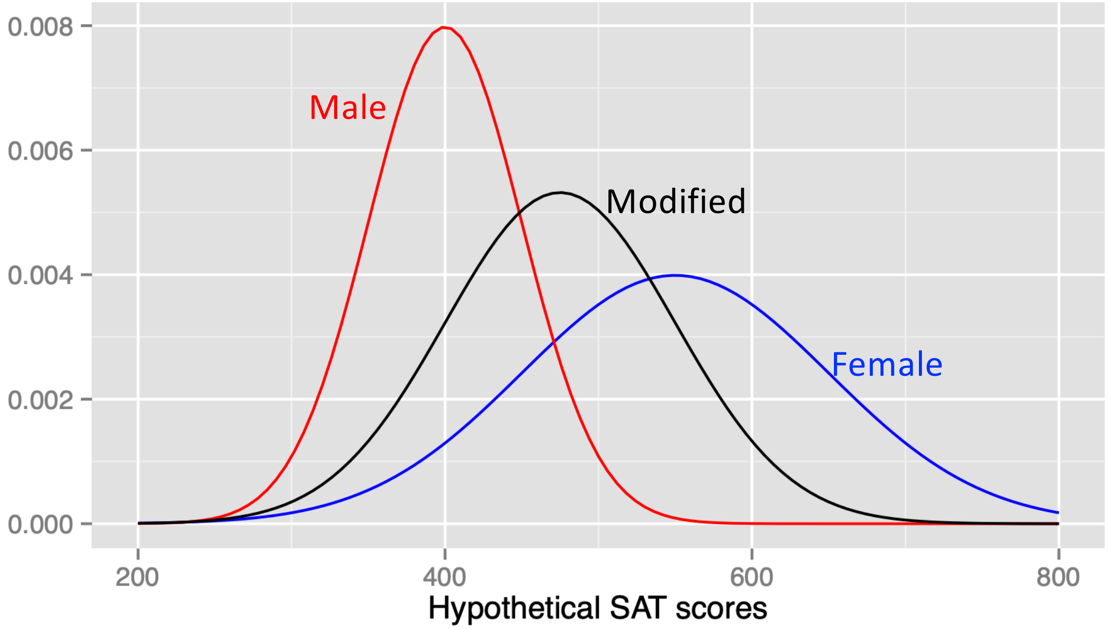
Idea: Look for classifier with low BER!
(authors use SVMs)

If BER < $\frac{1}{2} - \frac{\beta}{8}$ then we have disparate impact.

How do we get rid of disparate impact?

- Dataset:
  - SAT score: X
  - Gender: A ← sensitive
  - Admission status: Y

- Idea: change X while keeping A fixed
- Ensure that 95th percentile in original remains 95th percentile in the new distribution



Conditioned on new SAT score, what is the distribution over sensitive attribute?

Spot the mistakes

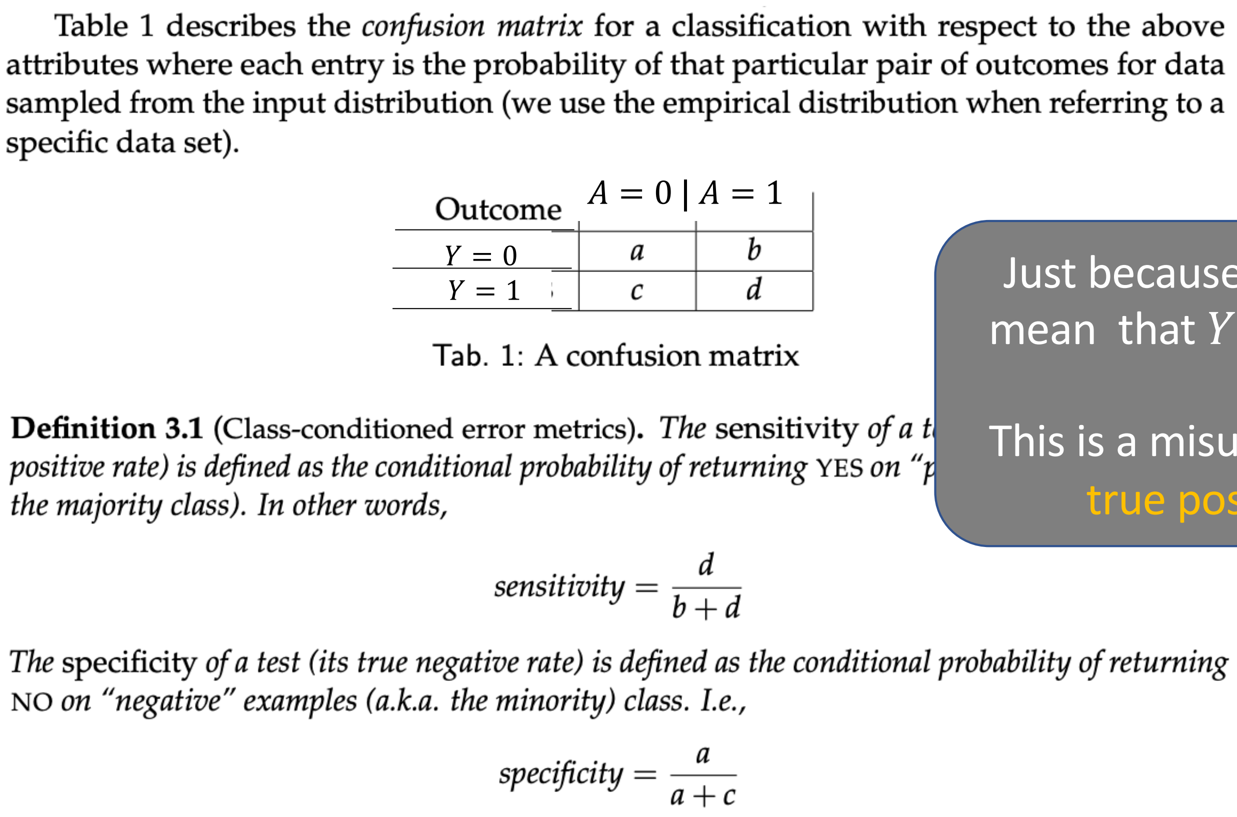


Table 1 describes the *confusion matrix* for a classification with respect to the above attributes where each entry is the probability of that particular pair of outcomes for data sampled from the input distribution (we use the empirical distribution when referring to a specific data set).

Outcome	$A = 0$	$A = 1$
$Y = 0$	a	b
$Y = 1$	c	d

Tab. 1: A confusion matrix

Definition 3.1 (Class-conditioned error metrics)**.** *The* sensitivity *of a t*
positive rate) is defined as the conditional probability of returning YES *on “p*
the majority class). In other words,

$$sensitivity = \frac{d}{b+d}$$

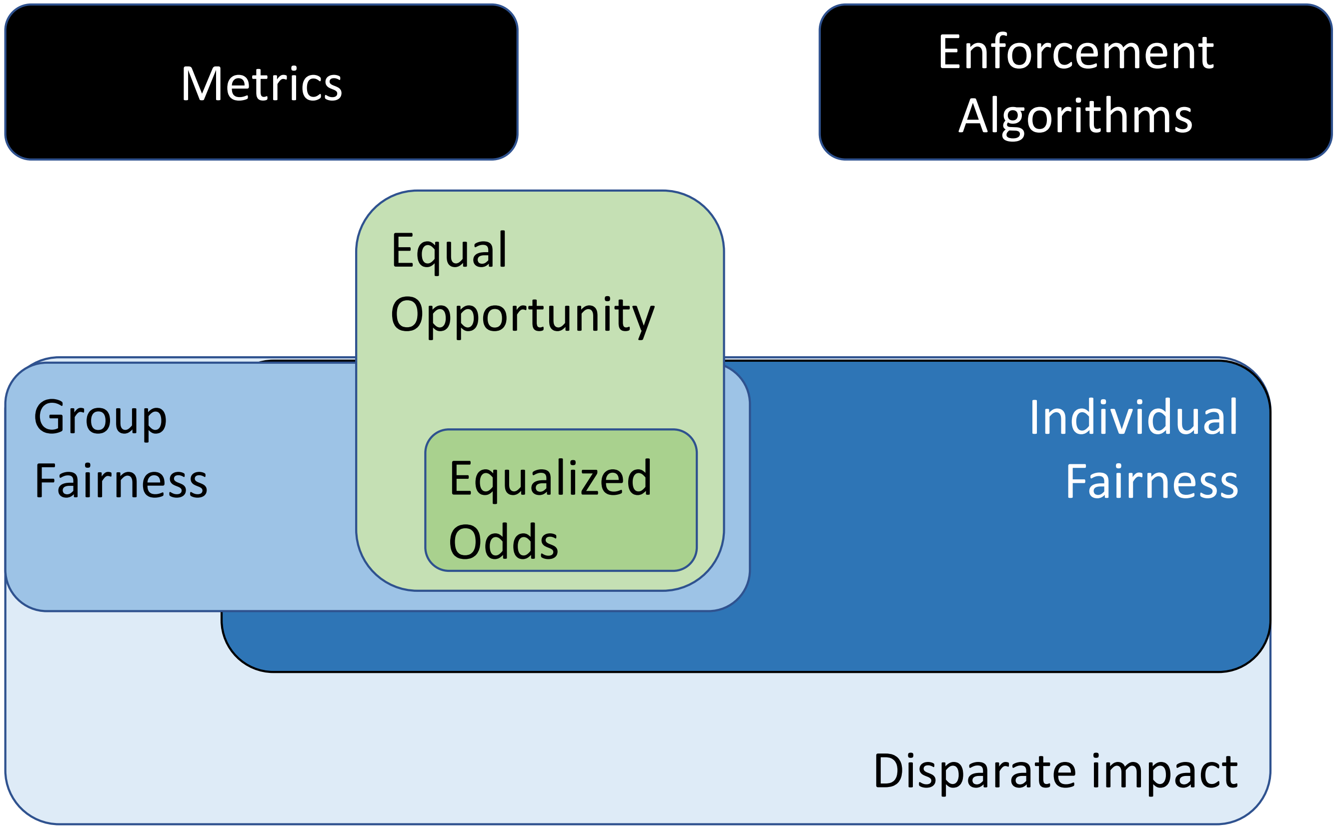
The specificity *of a test (its true negative rate) is defined as the conditional probability of returning* NO *on “negative” examples (a.k.a. the minority) class. I.e.,*

$$specificity = \frac{a}{a+c}$$

Just because $A = 1$ doesn’t mean that Y should equal 1!

This is a misunderstanding of true positive rates.

Fairness: High-Level View



Metrics

Enforcement Algorithms

Equal Opportunity

Equalized Odds

Group Fairness

Individual Fairness

Disparate impact