3. Fairness: Definitions and metrics

What is fairness ?

Many definitions of fairness:

- Individual. seeks for similar individuals to be treated similarly
- Group. split a population into groups defined by protected attributes (e.g. women, black) and seeks for some measure to be equal across groups

Individual fairness

Am I as an individual being treated as another with similar qualifications?

- *Fairness through unawareness*.
- *Fairness through awareness.*
- *Counterfactual fairness.*

Fairness through unawareness

- **Definition:** Protected attributes A not explicitly used as features X.
- **Mathematically:** $A \cap X = \emptyset$
- **Potential problem:** proxy in the data (see practice notebook).

Fairness through awareness/ Individual Fairness

- **Definition:** *S*imilar predictions to similar individuals (w.r.t similarity metric).

- **Mathematically:** x,y 2 individuals. Model M must satisfy Lipschitz condition:

$$D(M(x), M(y)) \leq d(x, y)$$

- **Potential problem:** Choice of similarity/distance metric.

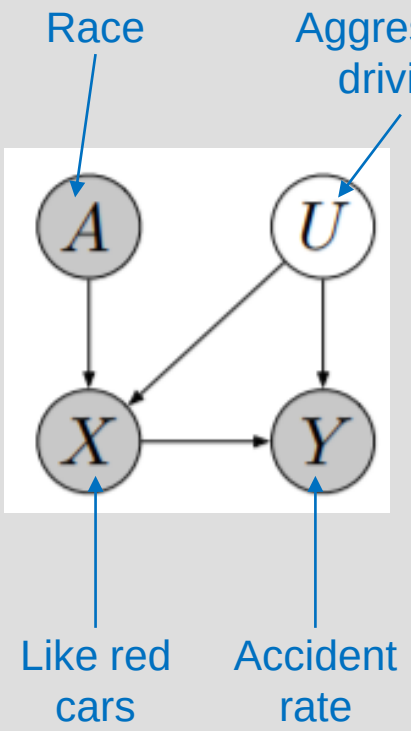
Counterfactual fairness

- **Definition:** Causal graph. Fair if the predicted outcome does not depend on a descendant of the protected attribute

- **Mathematically:** A protected attribute, X other features, U latent background variables. $\hat{Y}$ estimator.

$$P(\hat{Y}_{A<-a}(U) = y|X = x, A = a) = P(\hat{Y}_{A<-a'}(U) = y|X = x, A = a)$$

- **Potential problem:** determining the causal graph



Group fairness

Is the minority group to which I belong being treated as the majority group?

- **Equality of Outcome.** Distribution across groups should be the same *(Statistical Parity, Disparate Impact)*
- **Equality of Opportunity.** *E.g. Equalized Odds*

Equality of Outcome

- **Definition:** Equal likelihood of positive predicted outcome.
- **Mathematically:**

$$\underbrace{P(\hat{Y} = 1|A = male)}_{SR_{male}} = \underbrace{P(\hat{Y} = 1|A = female)}_{SR_{female}}$$

- **Example metrics:**
  - Disparate Impact = SR_{female} / SR_{male} → ideal value of 1
  - Statistical Parity = $SR_{female} - SR_{male}$ → ideal value of 0
- **Potential problem:** Positive discrimination ?

Equality of Opportunity

- **Definition:** This should be equal for all groups:
  - probability of a person in the positive class being correctly assigned a positive outcome
  - probability of a person in a negative class being incorrectly assigned a positive outcome

- **Mathematically:** $\underbrace{P\left(\hat{Y} = 1 \middle| A = m, Y = y\right)}_{TPR_m \text{ if } y = 1 \text{ ; } FPR_m \text{ if } y = 0} = P\left(\hat{Y} = 1 \middle| A = f, Y = y\right), y \in \{0,1\}$

- **Example metric:** Average Odds Difference

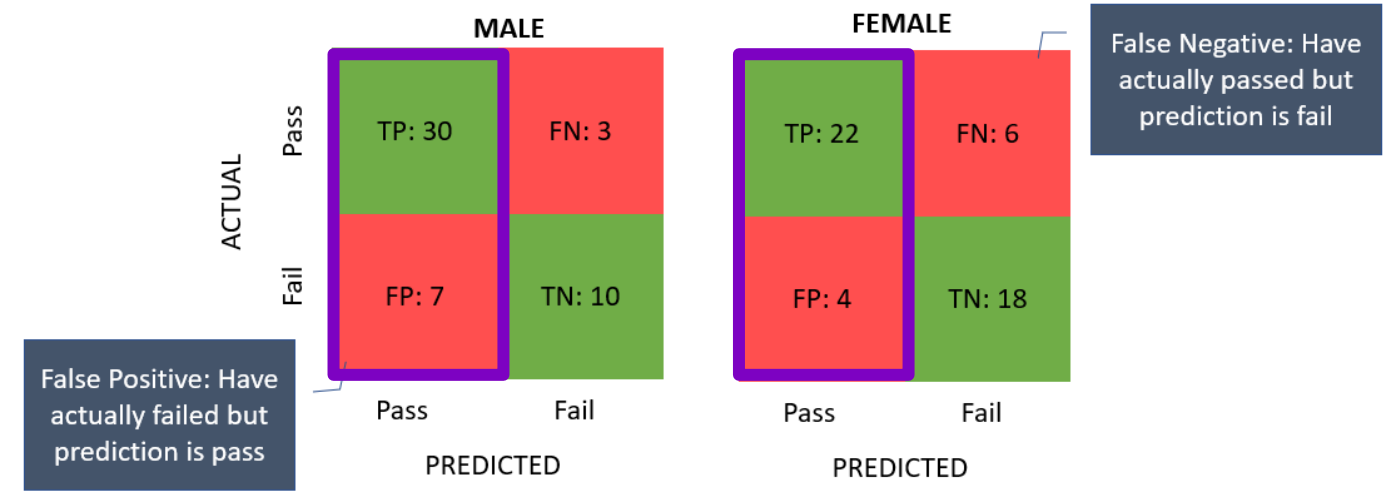
$$AOD = \frac{1}{2}\left[\left(FPR_f - FPR_m\right) + \left(TPR_f - TPR_m\right)\right]$$

- **Potential problem:** Where does ground-truth come from, is it a valid unbiased measure?

Group fairness - Illustration through an example

Equality of outcome →

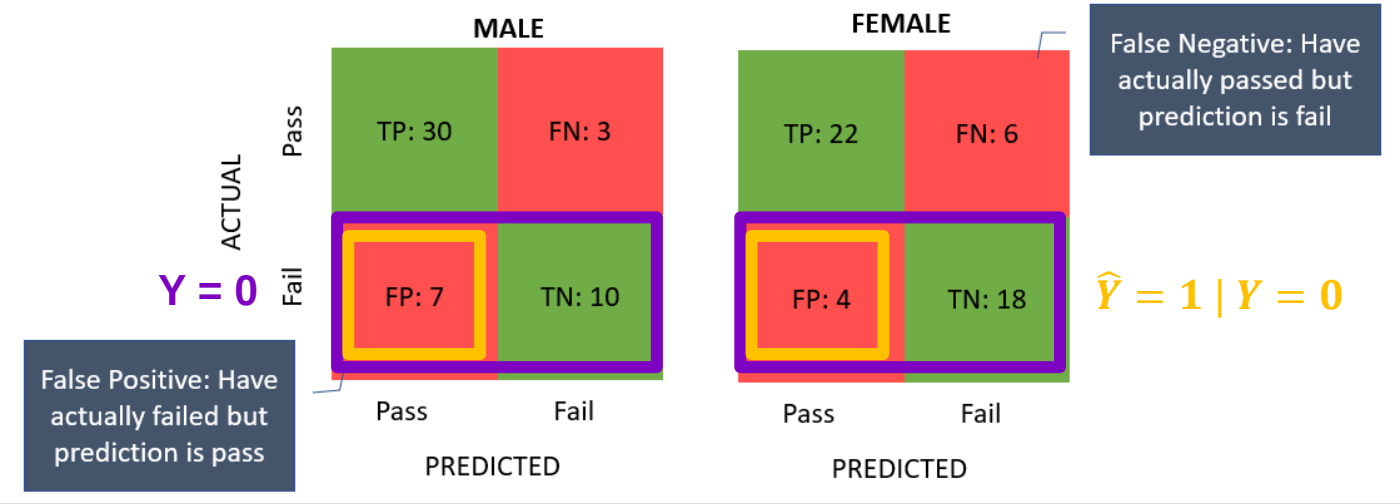
$$P(\hat{Y} = 1|A = male) = P(\hat{Y} = 1|A = female)$$



$$\begin{cases} SR_{male} = 37/50 = 0.74 \\ SR_{female} = \dfrac{26}{50} = 0.52 \end{cases} \Rightarrow \begin{cases} SP = 0.52 - 0.74 = \mathbf{-0.22} \\ \text{DI} = 0.52/0.74 = \mathbf{0.7} \end{cases}$$

Equality of opportunity →

$$P(\hat{Y} = 1 | A = male, Y = 0) = P(\hat{Y} = 1 | A = female, Y = 0)$$

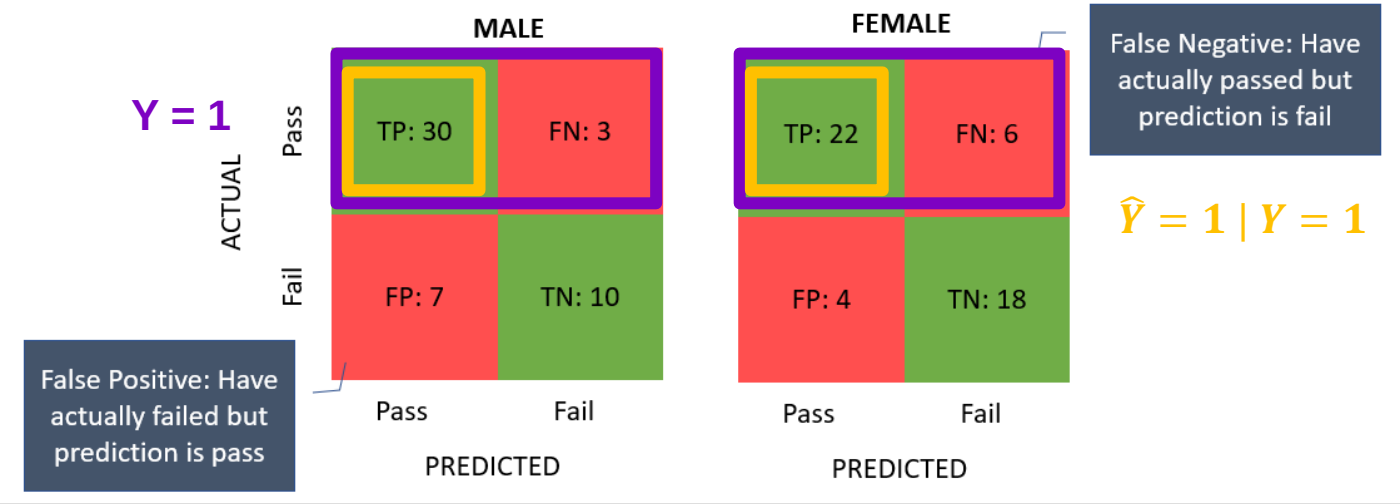


$$P(\hat{Y} = 1 | A = m, Y = 0) = FPR_{male} = \frac{FP}{FP+TN} = 7/17 = 0.41$$

$$P(\hat{Y} = 1 | A = f, Y = 0) = 4/22 = 0.18$$

Equality of opportunity →

$$P(\hat{Y} = 1 | A = male, Y = 1) = P(\hat{Y} = 1 | A = female, Y = 1)$$



$$P(\hat{Y} = 1 | A = m, Y = 1) = TPR_{male} = \frac{TP}{TP+FN} = 30/33 = 0.91$$

$$P(\hat{Y} = 1 | A = f, Y = 1) = 22/28 = 0.79$$

Average Odds Difference

$$\boldsymbol{AOD = \frac{1}{2}[(0.18 - 0.41) + (0.79 - 0.91)] = -0.175}$$

Reading/References

Supporting Notebook

Definitions of fairness
(Verma & Rubin, 2018) Fairness Definitions Explained
(Mehrabi et al., 2021). A survey on bias and fairness in machine learning

Metrics
https://www.holisticai.com/open-source