Table of contents

- Real-world examples
- Error parity and confusion matrices
- **Other fairness metrics**
- Pareto fronts

A probabilistic interpretation

$$\text{TPR} = Pr(\hat{Y} = 1 \mid Y = 1) = \text{Sens}$$

$$\text{FNR} = Pr(\hat{Y} = 0 \mid Y = 1) = 1 - \text{Sens}$$

$$\text{TNR} = Pr(\hat{Y} = 0 \mid Y = 0) = \text{Spec}$$

$$\text{FPR} = Pr(\hat{Y} = 1 \mid Y = 0) = 1 - \text{Spec}$$

Interlude: base rate fallacies and testing for viruses

- Consider a test that detects presence of a rare virus with sensitivity 99% and specificity 99%
- The virus is rare, it is only ever found in 1% of the population.
- You decide to test yourself. Your test turns out positive. What is the probability you have the virus?

$$
\begin{aligned}
Pr(Y=1 \mid \hat{Y}=1) &= \frac{Pr(\hat{Y}=1 \mid Y=1)Pr(Y=1)}{Pr(\hat{Y}=1)} \\
&= \frac{Pr(\hat{Y}=1 \mid Y=1)Pr(Y=1)}{Pr(\hat{Y}=1 \mid Y=1)Pr(Y=1)+Pr(\hat{Y}=1 \mid Y=0)(1-Pr(Y=1))} \\
&= \frac{\text{Sens}\cdot p}{\text{Sens}\cdot p+(1-\text{Spec})(1-p)} = \frac{0.99\cdot 0.01}{0.99\cdot 0.01+0.01\cdot 0.99} = 0.5
\end{aligned}
$$

- This is known as Positive Predictive Value, or PPV. Analogously we may define NPV: $Pr(Y=0 \mid \hat{Y}=0)$

Conditional independence assumptions

- Let us now introduce a sensitive attribute A. You can take A to be gender, for example.
- Equal opportunity then translates to:

$$Pr(\hat{Y} = 1 \mid Y = 1, A = 1) = Pr(\hat{Y} = 1 \mid Y = 1, A = 0)$$

- And equalized odds to additionally requiring that:

$$Pr(\hat{Y} = 0 \mid Y = 0, A = 1) = Pr(\hat{Y} = 0 \mid Y = 0, A = 0)$$

- These two conditions together imply the following conditional independence assumption:

$$\hat{Y} \perp A \mid Y \qquad \text{(Separation)}$$

Conditional independence assumptions

- Let us now take the opposite approach. What if we insisted instead that:

$$Y \perp A \mid \hat{Y} \qquad \text{(Sufficiency)}$$

- This is demanding that the probability that you are truly positive given you received a positive label should be independent of your gender, i.e., this is a statement about PPV and NPV parity
- Can we simultaneously achieve both?

Theorem. Sufficiency and separation can only both hold if A and Y are marginally independent.

Proof. Exercise (optional).

- So if the creditworthiness is truly independent of gender, then in principle there exist classifiers that create a label which is also genuinely independent of gender, and satisfies these constraints.

Demographic parity

- All conditions considered so far make statements about the accuracy of the algorithm. An alternative perspective is to sidestep that altogether and insist that the proportion of men and women receiving a positive label is equal. This is known as demographic parity, and translates to *independence*:

$$Pr(\hat{Y} \mid A = 0) = Pr(\hat{Y} \mid A = 1) \therefore \hat{Y} \perp A$$

Theorem. Sufficiency and independence can only both hold if A and Y are marginally independent.

Proof. Exercise (optional).

Theorem. If Y is binary, and the classifier label is not independent of Y, separation and independence can only both hold if A and Y are marginally independent.

Proof. Exercise (optional).

Summary

- A large number of possible fairness metrics are available, and many of them are competing in that they cannot hold at the same time unless very strong assumptions about the data hold.
- Some metrics rely on false positive and false negative rates. Others rely on positive or negative predictive values. Finally, demographic parity is concerned with the proportion of positive “decisions” made on each subpopulation, and is not concerned as much with their accuracy.
- A key contribution of machine learning to ethics of decision making is precise definitions.