Algorithmic Fairness – Impossibility Theorems

Marcello Di Bello - ASU - Fall 2021 - Week #6

Our goal is twofold: one, to gain a better understanding of the formal criteria of algorithm fairness; two, to understand the impossibility theorems about algorithmic fairness.

Chouldechova's impossibility theorem

This theorem[1] relies on two fairness criteria:

[1] Chouldechova (2017), Fair Prediction With Disparate Impact: A Study of Bias in Recidivism Prediction Instruments

1. Predictive parity
2. Classification parity

1. Predictive parity

Predictive parity requires that the Positive Predictive Value (PPV) of the algorithm be the same across groups. We define PPV, as follows:

$$PPV = \frac{TP}{TP + FP}$$

where TP stands for True Positive (i.e. the fraction of people who are correctly labeled 'positive' *out of all people*) and FP stands for False Positive (i.e. the fraction of people who are incorrectly labeled 'positive' *out of all people*).

The denominator refers to all people labeled 'positive', whether correctly (TP) or incorrectly (FP). So, predictive parity—equality in PPV across groups—requires that the fraction of people correctly labeled 'positive' *out of all people labeled positive whether correctly or incorrectly* be the same across groups.

2. Classification parity

Classification parity requires that the False Negative Rate (FNR)—that is, the fraction of people who are incorrectly labeled 'negative' *out of all positive people*—and the False Positive Rate (FPR)—that is, the fraction of people who are incorrectly labeled 'positive' *out of all negative people*)—be the same across groups. Phrased this way, classification parity requires *equality in error rates* across groups.

An equivalent way of defining classification would be this. Classification parity requires that the True Positive Rate (TPR)—that is, the fraction of people who are correctly labeled 'positive' *out of all positive people*—and the True Negative Rate (TNR)—that is, the fraction of people who are correctly labeled 'negative' *out of all negative*

people—be the same across groups. Phrased this way, classification parity requires *equality in correctness rates* across groups.

So classification parity can be understood as equality in FNR and FPR (equality of error rates) or equality in TPR and TNR (equality in correctness rates). Classification parity can also be defined in a number of other ways, all equivalent, keeping in mind that

Any definition of classification parity will include equality in FPR or TNR, along with equality in FNR or TPR.

$$TPR = 1 - FNR \text{ and } TNR = 1 - FPR.$$

The theorem

No algorithm can satisfy both predictive parity and classification parity so long as the base rates of the two groups are different.

Proof

Even though the theorem isn't difficult to prove—it is just a rewording of Bayes' theorem—its significance is profound. It is worth following the proof carefully. The proof consist of two parts.

The **first part** establishes the following claim:

$$\frac{PPV}{1 - PPV} = \frac{p}{1 - p} \times \frac{1 - FNR}{FPR}$$

where p stands for the *base rate* or *prevalence* of the outcome of interest in a given population.

Let's see why this claim holds. By definition, we know that

$$PPV = \frac{TP}{TP + FP} \qquad (*)$$

Notice that

$$TP = p \times TPR \text{ and } FP = (1 - p) \times FPR$$

where TPR is True Positive Rate (defined as the fraction of people who are correctly labeled 'positive' *out of all positive people*) and FPR is False Positive Rate (defined as the fraction of people who are incorrectly labeled 'positive' *out of all negative people*).

After replacing TP by $p \times TPR$ and replacing FP by $(1 - p) \times FPR$ in equation (*) above, we get:

If you know Bayes' theorem, you should recognize it here.

$$PPV = \frac{p \times TPR}{p \times TPR + (1 - p) \times FPR}.$$

We analogously define the complement of PPV, as follows:

$$1 - PPV = \frac{FP}{TP + FP}$$

Thus, by performing the appropriate replacements, we get:

$$1 - PPV = \frac{(1-p) \times FPR}{p \times TPR + (1-p) \times FPR}.$$

Finally, we take the ratio of PPV and its complement, as follows:

The rightmost equality uses the fact that $TPR = 1 - FNR$

$$\frac{PPV}{1-PPV} = \frac{\frac{p \times TPR}{p \times TPR + (1-p) \times FPR}}{\frac{(1-p) \times FPR}{p \times TPR + (1-p) \times FPR}} = \frac{p}{1-p} \times \frac{TPR}{FPR} = \frac{p}{1-p} \times \frac{1-FNR}{FPR}$$

We are now ready to embark on the **second part** of the proof, the one that establishes the impossibility. The equation we just established holds for any group (indices 1 and 2):

$$\frac{PPV_1}{1-PPV_1} = \frac{p_1}{1-p_1} \times \frac{1-FNR_1}{FPR_1}$$
$$\frac{PPV_2}{1-PPV_2} = \frac{p_2}{1-p_2} \times \frac{1-FNR_2}{FPR_2}$$

By inspecting the two equations, we can see that predictive parity (equality in PPV) and classification parity (here understood as equality in error rates, FNR and FPR) cannot be concurrently satisfied whenever the base rate or prevalence p differs across groups.

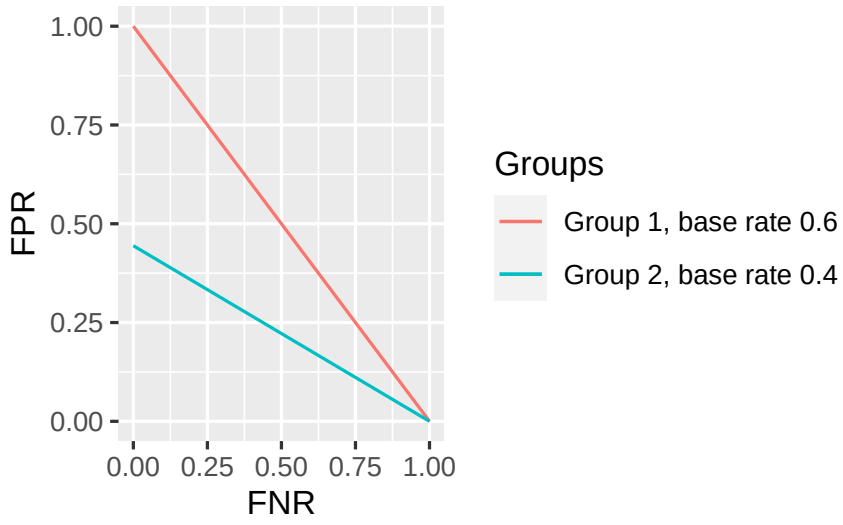
Suppose the base rates of the two groups are different, hence $\frac{p_1}{1-p_1} \neq \frac{p_2}{1-p_2}$. If $\frac{PPV_2}{1-PPV_2} = \frac{PPV_1}{1-PPV_1}$ (predictive parity is satisfied), then $\frac{1-FNR_1}{FPR_1} \neq \frac{1-FNR_2}{FPR_2}$ (classification parity is violated). Conversely, if $\frac{1-FNR_1}{FPR_1} = \frac{1-FNR_2}{FPR_2}$ (classification parity is satisfied), then $\frac{PPV_2}{1-PPV_2} \neq \frac{PPV_1}{1-PPV_1}$ (predictive parity is violated). QED.

Significance

Chouldechova writes:

> differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al. are a direct consequence of applying an RPI [=Recidivism Prediction Instruments] that that satisfies predictive parity to a population in which recidivism prevalencea differs across groups

As the graph below shows, the group with higher prevalence will experience higher FPR no matter the FNR, while holding PPV the same across groups. So the disparities we saw in the COMPAS algorithm cannot—as a mathematical fact—be avoided.



Kleinberg et al's impossibility theorem

This theorem[2] relies on slightly different fairness criteria:

3. *Predictive Calibration* - The fraction of positive people *out of those who are assigned the same risk score* is the same across groups. In other words, the score reflects the same likelihood of the outcome across groups.

4. *Classification Balance* - Positive individuals are assigned the same average risk score across groups (positive balance). Negative people are assigned the same average score across groups (negative balance).

The theorem states that, given different base rates or prevalence across groups, no algorithm can satisfy both predictive calibration and classification balance unless the algorithm achieves perfect prediction. The proof is involved.

[2] Kleinberg, Mullainathan and Raghavan (2016), Inherent Trade-Offs in the Fair Determination of Risk Scores Intuitvely, the risk score mean the same thing (=likelihood of the outcome, say recidivism) across groups.

Predictive calibration and classification balance are similar to predictive parity and classification parity. But they are criteria for the fairness of algorithmic scores rather than algorithmic decisions. These criteria of fairness can be applied independently of the choice of a decision threshold.

The Diagonal

The two impossibility theorems that use four criteria of fairness:

Impossibility	Classification	Predictive
Chouldechova	1.parity	2. parity
Kleinberg	3. balance	4. calibration

But what about *1. classification parity* (equality in *FPR* and *FNR*) and *4. prediction calibration*. Are these two criteria compatible?