A note on the impossibility of "fairness"

Thomas Miconi

Abstract

Various measures can be used to determine bias or unfairness in a predictor. Previous work has already established that some of these measures are incompatible with each other. Here we show that, when groups differ in prevalence of the predicted event, several intuitive, reasonable measures of fairness (ratio of positive predictions to event occurrence, probability of positive prediction given actual occurrence or non-occurrence, and probability of occurrence given positive or negative prediction) are all mutually exclusive: if *one* of them is equal among groups, *the other two* must differ. The only exceptions are for perfect, or trivial (always-positive or always-negative) predictors. As a consequence, any non-perfect, non-trivial predictor must necessarily be "unfair" under two out of three reasonable criteria. The result applies to all predictors, algorithmic or human. We conclude with possible ways to handle this effect when assessing and designing prediction methods.

1 Introduction

Suppose we must make predictions about the occurrence of a certain event, or condition, when the prevalence of this condition differs among different population groups. Ideally, we would want our predictions to be free of bias or prejudice. But what does it mean for a prediction to be "unbiased"? Simply equalizing the rate of positive predictions among groups is unsatisfactory, if the groups do differ in actual prevalence for the condition. We want our predictions to reflect the different base rates across groups, but not to unjustly flag people due to their group membership.

Several measures can be used to determine whether a predictor is biased, by comparing them across groups (in the following, Y is a binary variable indicating whether a certain event of interest actually occurs in reality, $Pred$ is a binary variable indicating that the event is predicted to occur by the predictor):

- The "harshness ratio", that is, the ratio between the number of predicted and actual positive cases $P(Pred)/P(Y)$ *(Measure 1)*
- The proportion of positive (or negative) predictions for which the event turns out to actually occur: $P(Y/Pred)$ and $P(Y/\neg Pred)$ *(Measure 2)*
- The proportion of actually positive (or negative) cases that elicit a positive prediction: $P(Pred/Y)$ and $P(Pred/\neg Y)$ *(Measure 3)*

These measures may indicate bias because if they differ significantly between groups, we might suspect that the predictor is biased. To illustrate with an example from criminal justice, if members of one group have a much higher probability of eliciting a positive prediction of guilt, in proportion to their probability of actual guilt (that is, if one group has a much higher $P(Pred)/P(Y)$ ratio), we might regard this as prima facie evidence of bias. Similarly, if guilty defendants from group A have a lower chance of eliciting a positive prediction of guilt than guilty members of group B, we would suspect that the predictor is biased in favour of group A. Conversely, if defendants from group A that elicit a positive prediction of guilt have a lower chance of being actually guilty than similarly-predicted members of group B, we would suspect that the predictor is biased against group A.

Ideally, we would want all of these measures to be simultaneously equal across groups. And indeed, if our predictions were absolutely perfect, with zero errors (i.e. $P(Y/Pred) = P(Pred/Y) = 1$ and $P(Y/\neg Pred) = P(Pred/\neg Y) = 0$, for all groups), this would actually be the case: all these measures would be strictly equal across all groups, independent of each group's prevalence for the condition[1]

Unfortunately, for non-perfect predictors, when prevalence differs among groups, some of these measures are known to be incompatible. Kleinber, Mullainathan and Raghavan [3], as well as Chouldechova [1], showed that equalizing Measure 2 and Measure 3 among groups simultaneously was impossible.

The purpose of this note is to show that, for non-perfect predictors and under different prevalences among groups, all three of these measures of fairness are mutually exclusive. That is, equalizing *one* of these three measures across groups guarantees that *both* other measures will differ across groups.

2 Proofs

We assume that $P(Y)$ differs among groups, and that the predictor $Pred$ is neither perfect, nor trivial (always-positive or always-negative).

Suppose $P(Pred/Y)$ and $P(Pred/\neg Y)$ are equal among groups (Measure 3). What can we say about $P(Pred)/P(Y)$ (Measure 1)?

$$\begin{aligned} P(Pred)/P(Y) &= \frac{P(Pred \cap Y) + P(Pred \cap \neg Y)}{P(Y)} \\ &= \frac{P(Pred/Y)P(Y)}{P(Y)} + \frac{P(Pred/\neg Y)P(\neg Y)}{P(Y)} \\ &= P(Pred/Y) + P(Pred/\neg Y)\frac{P(\neg Y)}{P(Y)} \end{aligned} \tag{1}$$

$\frac{P(\neg Y)}{P(Y)}$ is a strictly monotonic (decreasing) function of $P(Y)$. Therefore, if $P(Pred/Y)$ and $P(Pred/\neg Y)$ are equal among groups, and $P(Y)$ differs among group, then expression 1 will differ among groups - unless $P(Pred/\neg Y)$ is 0, which occurs only for perfect predictors or trivial, never-positive predictors. Therefore, equalizing Measure 3 forces a difference in Measure 1 (and, by simple contraposition, vice versa).

From Bayes' theorem, identical non-zero $P(Pred/Y)$ and different $P(Pred)/P(Y)$ immediately implies different $P(Y/Pred)$. Therefore, equalizing Measure 3 also forces a difference in Measure 2.

From the same reasoning that led to expression 1, simply by swapping symbols, we obtain that if $P(Y/Pred)$ and $P(Y/\neg Pred)$ are equal among groups (Measure 2), then $P(Y)/P(Pred)$ is a strictly decreasing function of $P(Pred)$. Now if $P(Y)/P(Pred)$ were equal among groups, since $P(Y)$ differs among groups, then $P(Pred)$ would also need to differ among groups - and thus $P(Y)/P(Pred)$ (as a strictly monotonic function of $P(Pred)$) would also differ among groups, leading to a contradiction. Thus, equalizing Measure 2 forces a difference in Measure 1, and again vice versa from contraposition.

Again, from Bayes' theorem, identical $P(Y/Pred)$ and different $P(Y)/P(Pred)$ immediately implies different $P(Pred/Y)$. Thus, equalizing Measure 2 forces a difference in Measure 3.

Collectively, these proofs substantiate the argument of this paper: equalizing any of these three measures among groups forces both other measures to differ among groups.

3 Discussion

Since several intuitive measures of fairness are mutually exclusive (when populations differ in prevalence and the predictor is neither perfect nor trivial), it follows that any predictor can always be portrayed as biased or unfair, by choosing a specific measure of fairness. Note that this result applies to all forms of prediction, whether performed by algorithms or by humans.

Given the mathematical impossiblity of simultaneously meeting all the measures of "fairness" described above, how can a predictor adjust their behavior? There are multiple options, which are not mutually

[1] Incidentally, checking whether a proposed measure of fairness is met by a perfect predictor is a useful, if weak, criterion for judging such measures; note that equal prediction rates across groups fails to meet it.

exclusive.

3.1 Choose one measure and stick to it

These various measures tend to emphasize different viewpoints and perspectives about fairness, and depending on the situation, it is possible that one meaure could be legitimately preferred to others. The first measure suggests that the predictor should be equally "harsh" across groups, ensuring that the proportion of positive predictions to actual events is similar (ideally, equal to one). This is a highly intuitive notion of fairness at the level of the group.

The second definition may be seen as reflecting the judge's viewpoint: "given the output of the test, what is the chance that the event actually occurs?" Similarly, the third definition seems to reflect the defendant's viewpoint: "Given that I am actually innocent (or not), what is the chance that the algorithm will flag me?"

We note that this last measure, in particular, seems to be of special interest. The so-called Blackstone formulation: "It is better that ten guilty persons escape than that one innocent suffer" can be interpreted as emphasizing $P(Pred/\neg Y)$ as the most important measure. It is also the basis for Hardt's equalized odds criterion [2] (which in the binary case is exactly identical to our Measure 3).

3.2 Improve the test as much as possible

As stated in the introduction, a perfect classifier would simultaneously equalize all these measures across groups. Inspection of the proofs reveals that the dependence of these measures on actual prevalence results from terms of the form $P(Pred/\neg Y)\frac{P(\neg Y)}{P(Y)}$ - i.e. an error term multiplied by a strictly monotonic function of the prevalence. If the error terms ($P(Pred/\neg Y)$, $P(Y/\neg Pred)$, etc.) can be reduced as much as possible, the discrepancy between the various measures will also decrease. Thus, improving the accurracy of the predictors will tend to improve their fairness (or more precisely, the compatibility between various measures of fairness).

3.3 Ignore these measures and reach for a first-principles solution

Ideally, a fair predictor would effectively eliminate group membership as an independent influence on decisions. As a theoretical example, one may consider a multiple-regression approach, in which a regressor for group membership is used while training the predictor, then discarded at decision time (this is a purely theoretical example; we do not claim that such a method is actually feasible or desirable in practice). Under some conditions, a method of this kind would guarantee that group membership does not affect conclusions even if other variables correlate with group membership.

However, besides the fact that it would likely lead to worse predictions (because group membership may itself be a proxy for other predictive variables not included in the model), it would also fail the measures of fairness described above, and thus might still lead to claims of bias.

3.4 Use a detailed cost-benefit approach

In some circumstances, the costs of the various types of errors (false alarms, misses, etc.) can be quantified. In this case, assuming sufficient and reliable data, the objective optimum for each quantity can be effectively computed. Such cost-benefits analyses are particularly relevant in the medical domain, as illustrated by the output of the British National Institute for Health and Care Excellence. However, in other circumstances, the costs and benefits are either difficult to assess, or lie in a moral plane that is not amenable to quantification.

To conclude, we stress that the foregoing should in now way distract from the important goal of dismantling the very real biases and discriminations that affect existing decision mechanisms. To spell out the obvious: just because any decision system will mathematically be "biased" along some measure, does not at all imply that any observed bias is simply a mathematical artifact. Thus, the incompatibility between fairness measures should not be used as a cover for obvious injustices. On the contrary, a better understanding of logical constraints on the outcomes of decision systems can and should inform, and therefore assist, efforts toward a fairer world.

References

[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

[2] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[3] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.