No nonsense version of the "racial algorithm bias"

by Yuxi_Liu

13th Jul 2019

In discussions of algorithm bias, the COMPAS scandal has been too often quoted out of context. This post gives the facts, and the interpretation, as quickly as possible. See this for details.

The fight

The COMPAS system is a statistical decision algorithm trained on past statistical data on American convicts. It takes as inputs features about the convict and outputs a "risk score" that indicates how likely the convict would reoffend if released.

In 2016, ProPublica organization claimed that COMPAS is clearly unfair for blacks in one way. Northpointe replied that it is approximately fair in another way. ProPublica rebukes with many statistical details that I didn't read.

The basic paradox at the heart of the contention is very simple and is not a simple "machines are biased because it learns from history and history is biased". It's just that there are many kinds of fairness, each may sound reasonable, but they are not compatible in realistic circumstances. Northpointe chose one and ProPublica chose another.

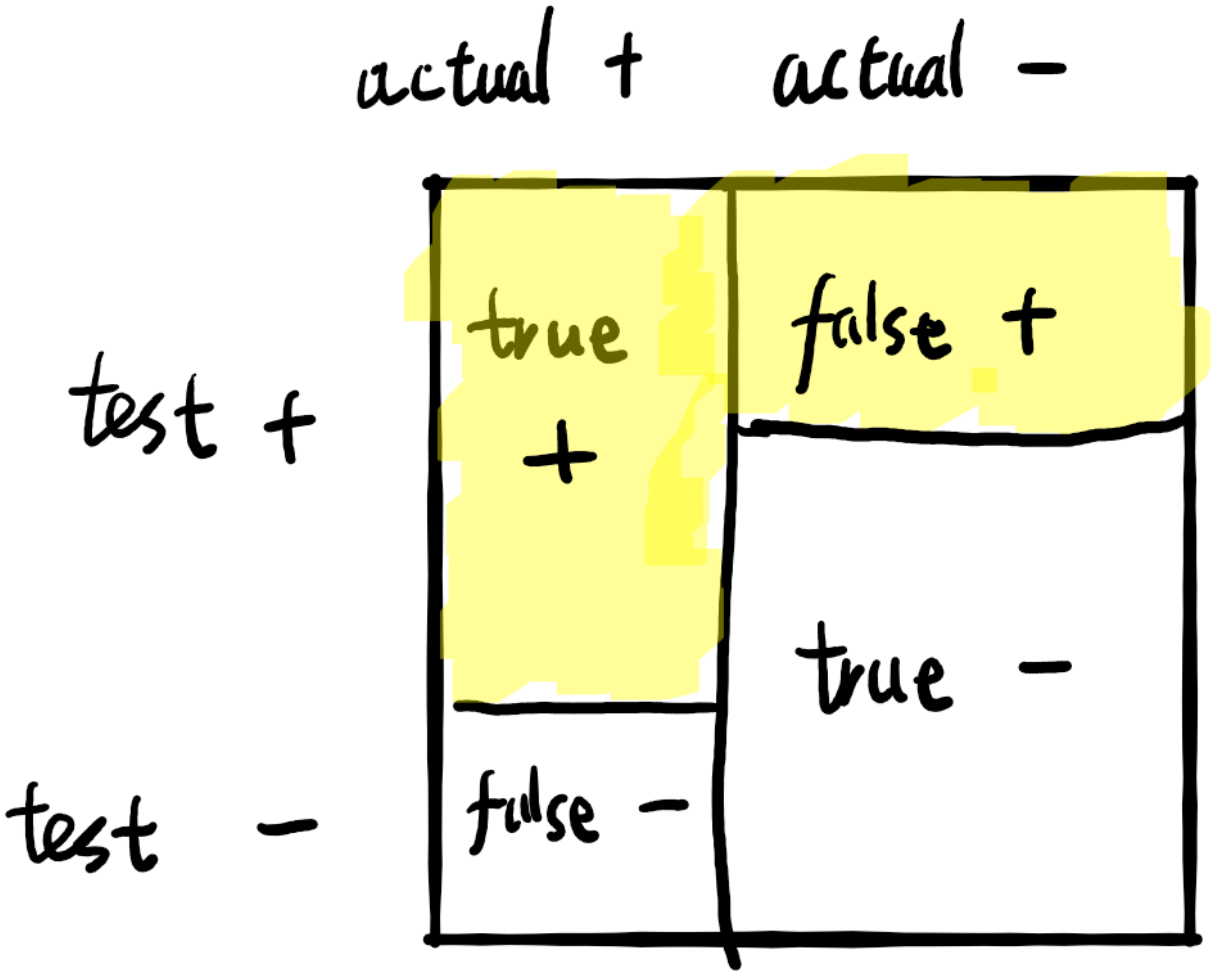
The math

The actual COMPAS gives a risk score from 1-10, but there's no need. Consider the toy example where we have a decider (COMPAS, a jury, or a judge) judging whether a group of convicts would reoffend or not. How well the decider is doing can be measured in at least three ways:

- False negative rate = (false negative)/(actual positive)
- False positive rate = (false positive)/(actual negative)
- Calibration = (true positive)/(test positive)

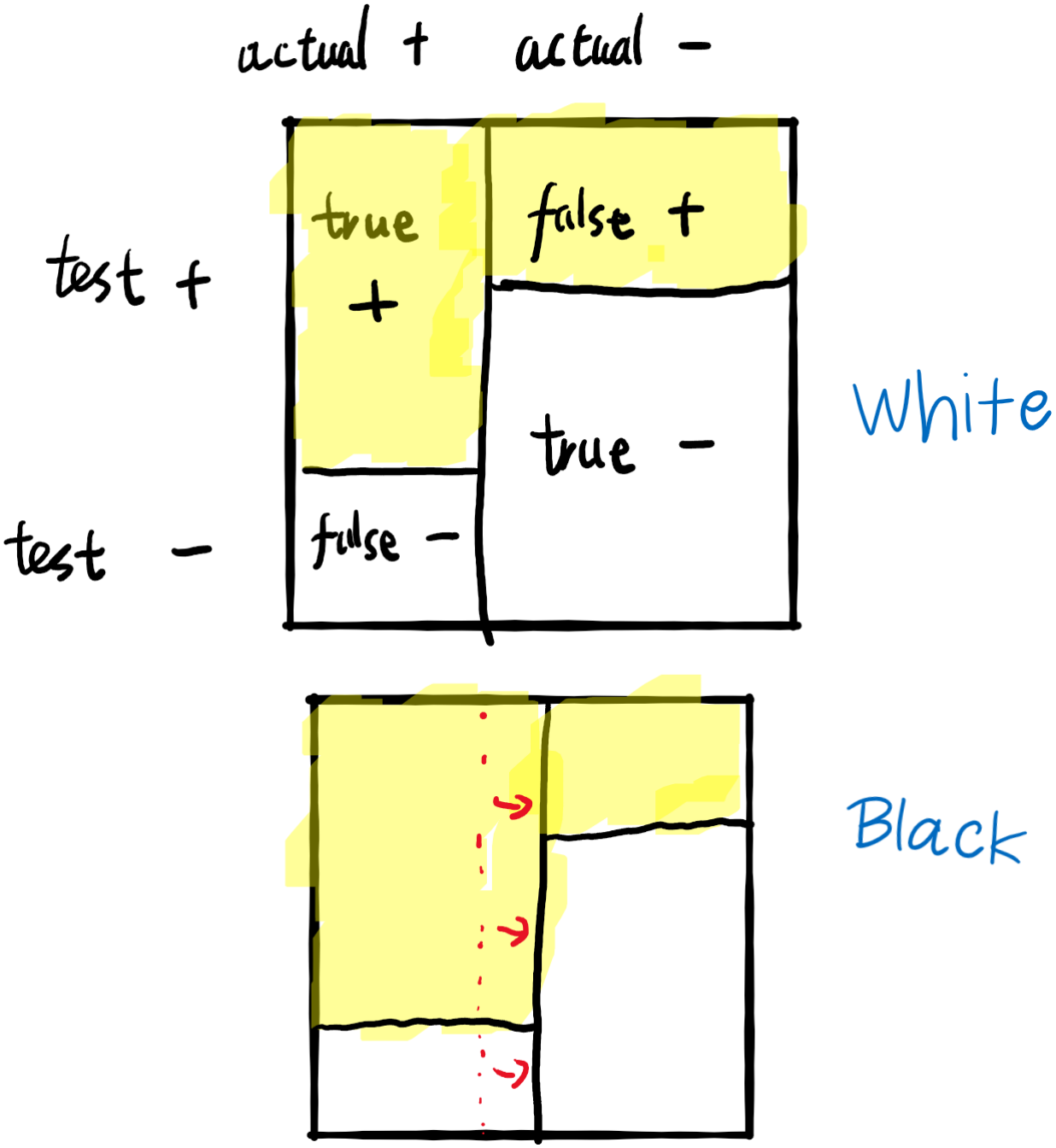
A good decider should have false negative rate close to 0, false positive rate close to 0, and calibration close to 1.

Visually, we can draw a "square" with four blocks:



- false negative rate = the "height" of the false negative block,
- false positive rate = the "height" of the false positive block,
- calibration = (true positive block)/(total area of the yellow blocks)

Now consider black convicts and white convicts. Now we have two squares. Since they have different reoffend rates for some reason, the central vertical line of the two squares are different.



The decider tries to be fair by making sure that the false negative rate and false positive rates are the same in both squares, but then it will be forced to make the calibration in the Whites lower than the calibration in the Blacks.

Then suppose the decider try to increase the calibration in the Whites, then the decider must somehow decrease the false negative rate of Whites, or the false positive rate of Whites.

In other words, when the base rates are different, it's impossible to have equal fairness measures in:

- false negative rate
- false positive rate
- calibration

Oh, forgot to mention, even when base rates are different, there's a way to have equal fairness measures in all three of those... But that requires the decider to be *perfect*: Its false positive rate and false negative rate must both be 0, and its calibration must be 1. This is unrealistic.

In the jargon of fairness measurement, "equal false negative rate and false positive rate" is "parity fairness"; "equal calibration" is just "calibration fairness". Parity fairness and calibration fairness can be straightforwardly generalized for COMPAS, which uses a 1-10 scoring scale, or indeed any numerical risk score.

It's some straightforward algebra to prove that in this general case, parity fairness and calibration fairness are incompatible when the base rates are different, and the decider is not perfect.

The fight, after-math

Northpointe showed that COMPAS is approximately fair in calibration for Whites and Blacks. ProPublica showed that COMPAS is unfair in parity.

The lesson is that there are incompatible fairnesses. To figure out which to apply -- that is a different question.