HW 4

Mason Boyles

10/10/2024

This homework is designed to give you practice working with statistical/philosophical measures of fairness.

The paper linked below[1] discusses potential algorithmic bias in the context of credit. In particular, banks are now regularly using machine learning algorithms to do an initial screening for credit-worthy loan applicants. In section 4.5.2, this paper reports the rates at which various racial groups were granted a mortgage. If we assume that it is a classifier making these predictions[2] what additional information would be necessary to assess this classifier according to equalized odds?

In order to assess this classifier according to equalized odds, we would need to know the false positive and true positive rates of the predictor across the racial groups. These rates should we roughly equal for all groups, or at least, the differences should be lower than ε (usually .2).

Show or argue that the impossibility result discussed in class does not hold when our two fringe cases[3] are met.

When there is a perfect predicting classifier and perfectly equal proportions of ground truth, Independence will hold because there are equal proportions of ground truth, so $P[\hat{y} = 1|S = X]$ will be equal no matter the X. This means the DI score will be 1 and the statistical parity score will be 0. As far as separation, equalized odds will be satisfied because it is a perfect predicting classifier in this scenario meaning the false positive rate will be 0 and the true positive rate will be 1 for all classes. This means that

$$|P[Y = 1|(S = 1 \cap Y = 0)] - P[Y = 1|S \neq 1 \cap Y = 0]| \geq \varepsilon$$

and

$$|P[Y = 1|(S = 1 \cap Y = 1)] - P[Y = 1|S \neq 1 \cap Y = 1]| \geq \varepsilon$$

will both be 0. Finally, this case would satisfy sufficiency because the prediction will not reveal anything about the underlying class label. To show this, we need to show that

$$P(y|S, \hat{y}) == P(y|\hat{y})$$

. This is true because its a perfect classifier so both are 1. Therefore it is possible for all three fairness criteria to be met in this case.

[1] https://link.springer.com/article/10.1007/s00146-023-01676-3

[2] It is unclear whether this is an algorithm producing these predictions or human

[3] a) perfect predicting classifier and b) perfectly equal proportions of ground truth class labels across the protected variable

How would Rawls's Veil of Ignorance define a protected class? Further, imagine that we preprocessed data by removing this protected variable from consideration before training out algorithm. How could this variable make its way into our interpretation of results nonetheless?

It would define a protected class as a group that has historically had a disadvantage when we separate into classes based on a variable. Even if we remove this variable from consideration, it is possible it could make its way into our results due to proxy variables. This was an issue for COMPAS because dispite not taking race into account, it took other variables into account that are highly correlated with race so they essentially act as a variable for race under another name.

Based on all arguments discussed in class, is the use of COMPAS to supplement a judge's discretion justifiable. Defend your position. This defense should appeal to statistical and philosophical measures of fairness as well as one of our original moral frameworks from the beginning of the course. Your response should be no more than a paragraph in length.

I am of the opinion the use of COMPAS to supplement a judges decision is not justifiable. The main reason stems in the fact that it does appeal to separation since the equalized odds measures did not line up to where they should be for one of the most important protected groups (race). From the perspective of virtue ethics, I would say that this goes directly against the virtue of fairness, which is the whole point of the legal system. In addition, since the algorithm is completely black box, it is harder to audit its decision, which I would argue goes against the virtue of transparency. Along with this, if you were to appeal the decision, the algorithm will not be able to take in additional information and will come to the same result in revision. Further, to respond to those who argue that it is okay because the judge can override or appeal the decision made by the algorithm, I would say that regardless, if it is unfair across protected classes, it would cloud a judge's decision and sway them more towards more unfair judgements in the long run. In addition, it provides a lazy fallback solution for situations in which the judge is unsure and thus deincentivises them from digging deep into a person's situation because they can blame bad decisions on the model.