HW 4

Swagat Adhikary

10/28/2024

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?

To assess a classifier for equalized odds, we need the True Positive Rate and False Positive Rate for each racial group, as equalized odds requires both rates to be similar across groups. In other words, equalized odds ensures that credit-worthy individuals are equally likely to be approved across groups and non-credit-worthy individuals are equally likely to be approved across groups as well..

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

In the case where we have a perfect classifier and the underlying base rates are equal, all three fairness criteria are met. Independence is met because the perfect classifier will never predict disproportionately for any of the protected variables since their ground truth rates are equal. Separation holds because (by definition of perfect classifier) the false positive rate across protected variables will always be 0 and true positive rate across protected variables remain 100% and thus equal in odds. Sufficiency is present because the classifier's predictions are perfectly aligned with actual outcomes, giving each group a precision of 100%

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 our algorithm. How could this variable make its way into our interpretation of results nonetheless?

Under Rawls's Veil of Ignorance, a protected class would be defined as a group whose characteristics, such as race, gender, or socioeconomic status, should not influence decisions because individuals, acting behind the veil, do not know their own position in society. The veil of ignorance principle calls for fairness by treating everyone as if they could be in any position, making biases against certain groups (protected classes) unfair, as no one would want to risk disadvantaging themselves if they were in that class. Simply preprocessing data by removing this protected class from consideration doesn't stop it from influencing the interpretation of results,

[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

as proxy variables which correlate extensively with the protected variable could still be considered and thus influence the interpretation in a way similar to how the protected variable would.

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.

The use of COMPAS to supplement a judge's discretion is unjustifiable due to its violations of key fairness metrics—such as equalized odds—resulting in disproportionate false positives between Black and White defendants as calculated in class. Statistically, this disparity indicates that COMPAS is not equally reliable across racial groups, undermining its fairness and trustworthiness. From a Rawlsian perspective, using COMPAS without addressing these biases conflicts with the Veil of Ignorance principle, which would reject any system that places one group at an unfair disadvantage. Although COMPAS may provide additional information to judges, its lack of fairness across groups suggests it could reinforce systemic biases rather than mitigate them. To align with both statistical fairness and moral fairness, reforms or alternative unbiased tools are necessary before such an algorithm can ethically support judicial discretion.