HW 4

Anna Fetter

10/29/2024

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?

Student Input.

Equalized odds states that the false positive and false negative rates should be equal across groups. In 4.5.2, the reported data showed different mortgage approval rates for different racial groups. To assess equalized odds, the data would also need to report false positive and false negative rates, meaning people who should have been approved for loans and weren't and those who were approved who shouldn't have been.

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

Student Input

The impossibility result from the incompleteness theorem states that it is impossible to meet all three fairness criteria (independence, separation, sufficiency) except for two fringe cases: perfect classifiers and perfectly equal proportions of ground truth. The impossibility result doesn't hold because when the classifier is perfect, there is no room for disagreement on independence, separation, or sufficiency since the expected and the actual values are the same. Equal class proportions of the ground truth also minimize some of the discrepancies between the fairness criteria, making it possible to satisfy all three.

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?

[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

Student Input

Rawl's Veil of Ignorance is the thought experiment where the participant won't know their circumstances while they sit under this "veil of ignorance" and make societal decisions. The idea is that people would make more just decisions if they weren't influenced by their own status, abilities, or other characteristics. Rawl's Veil of Ignorance would describe a protected class as any characteristics that could lead to disadvantage to a person during the decision process about the rules of society. Removing these protected classes before training the algorithm would be a good first step, but proxy classes that highly correlate to the protected class could just as well disadvantage these people in the protected class.

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.

Student Input

COMPAS is not justified to supplement judge discretion because it fails equalized odds across racial boundaries. COMPAS overpredicts violent crime recidivism for black Americans and underpredicts violent crime recidivism for white Americans. The moral theory of consequentialism would also say that COMPAS is not justified because the consequences of sentencing people to more years in prison based on race or proxies for race may outweigh the societal good of having COMPAS in the first place. Consequentialism would require an examination into the societal consequences of COMPAS correctly predicting rates of recidivism among groups.