HW 4

Razmin Bari

10/28/2024

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

1

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?

Equalized odds account for the ground truth. In this case, the information about if each applicant would have been granted mortgage if they were evaluated per the mortgage-granting criteria individually (without using the algorithmic classifier) and without any inherent bias on the evaluator's part is required.

2

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

If the classifier is perfect, there would not be any false negatives or false positives. Therefore equalized odds will be equal to 0 and thus will not be violated. If there are perfectly equal proportions of ground truth class labels as well, then statistical parity will be 0 and the independence criterion will be met. This also means that the true outcome class is independent of the protected variable once conditioned on the predicted class (since the predicted and true class is the same) and so the sufficiency criteria is met.

3

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?

A protected class would be defined as the most vulnerable and, hence, most disadvantaged group of people by virtue of having one common characteristic that predisposes them to that disadvantage.

If there is a variable whose values for an individual is very closely related or even dependent on the value of the removed protected variable, then the closely related variable can act as a proxy and induce roughly the same bias.

[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

4

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 low accuracy rate and bias (equalized odds criterion unmet) means that it does not provide sufficiently valuable information to begin with. If COMPAS is used merely as a supplement to the judges' decision, it would not circumvent the judges' bias anyway. If a judge heavily bases their decision on this, it would override the US legal system's procedures in that the judge would be unable to provide reasoning rooted in evidence that can be disseminated to the public. Moreover, the idea that other people's data/history can be used to pass judgement on another person with similar background means that there may be people who may be denied parole even if they were fully deserving of it. Conversely, defendants who should be denied parole may end up getting it early. This consequence would lead to a more unsafe society - an ultimately bad consequence. Therefore, by consequentialist arguments, COMPAS use is unjustified.