HW 4

Luca Buchoux

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?

For the algorithm not to violate equalized odds, the probability of the algorithm classifying an observation given that it belongs to some group and that it should (or shouldn't) be truly classified to that group must be within a certain threshold of the same probability given that it is not part of that same group. In this section, we learn the rate at which different demographic groups are approved for a mortgage with a clear trend. In order to determine if these rates violate equalized odds, we would first need to choose an epsilon and then break down the given rates for each demographic into the correct classification and misclassification rates.

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

In the first fringe case, we assume that the response Y is not independent of the variable S. Now for the sake of contradiction, assume we satisfy sufficiency and separation. Thus Y is independent of S given our prediction Yhat and Y is independent of Yhat given S. Equivalently, we can say S is independent of Yhat given Y and S is independent of Y given Yhat. This is the same as saying S is independent of Yhat and Y. I.E. S is independent of Yhat and S is independent of Y. Thus we contradicted our assumption that Y is not independent of S.

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

Rawl's Veil of Ignorance would not define protected classes, because the whole thought behind it is that decisions are made without regard to any class or status, only on what is right and wrong. If a protected class were omitted from the training of an algorithm, it could still be used to determine other sources of bias by bringing the variable back after we make our predictions. If we still see some sort of bias in the predictions, there could be other variables which are still hampering our algorithm.

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.

In the algorithm's current state, I would argue that the use of COMPAS is not justifiable. While the defense is accurate in saying that it is impossible to achieve all statistical measures of fairness, it can not be ignored that violating equalized odds has significant negative impact on one group of people over another. The solution to that argument should not be to merely accept it, but improve the model in a way that increases its performance and decreases the negative real world effects. And while you can also argue that COMPAS is simply an aide to a judge, tools like it can be easily be relied too heavily upon or even subconsciously impact the decisions made. Taking a page from consequentialism here, the negative impacts that COMPAS would have on those affected deem it unjustifiable for use as is.