HW 4

Ryan Dee

10/10/2024

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?

Student Input.
Under equalized odds there is only discrimination if the true positive and negative rates are different between certian classes, and we are not sure if they are based on this literature review. We would also need some measure of the difference in the false positive and false negative rates of the predictor. I would assume this predictor would be something like whether someone defaulted on a mortgage. Then we would need the individual decisoin matrices between latino, white, and black people applying for loans.

2

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

Student Input

3

The impossibility result discussed in class refers to the idea that no predicitive model can satisfy equalized odds, and disparate impact.

In instance a the equalized odds is trivially satisfied because the True Positive Rates of both classes are the same: 1, because the classifier makes no mistakes. At the same time it satisfies disparate impact because the positive prediction rate is the same accross groups.

Instance B describes a situation where the ground truth labels for each group is equal. In this instance an accurate predictor that does not discriminate can diprove our impossibility result. Because each group has the same the same base rate, an accurate predictor will give the same True Positive Rate for each group, which satisfies equalized odds, it can also satisfy disparate impact because the positive predictions would be the same for each group.

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?

Student Input Under Rawl's veil of ignorance, we are called to imagine rules that would not unfairly benefit a group because that group would not know they are a member of that class. Under this idea, a protected class may behave differently behind the veil, promoting ideas that they may think will help everyone equally but will actually disadvantage them when they are in the real world. In response to the second quesiton, there are proxies for protected class, things such as zipcode or family history are things that are only really tied to a group label that we would otherwise not consider in our interpretation. This is how we have disparate impact between group labels even when the group labels are not asked for. #

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 From my perspective the use of COMPAS to supplement a Judge's discretion cannot be justifiable from a deontologist point of view. This is because of the idea that the mechanism through which the COMPAS algorithm is produced, a training data of similar cases, is not admissible in court and should not be used to sentence an independent person. At the same time, from a deontologist point of view, this would have to be used for every parole decision under deontology. This is not only unrealistic, but not practical. In addition, the COMPAS algorithm violates the equalized odds standard of parity, and due to its black box nature the uncertainty for why that is the case could not be generalized under deontology. Finally, appealing to Rawl's difference principle, which states there will be disparity amongst people but it should only be permitted given that it benefits the least favored in society. Currently the justice system does not operate in this manner, but the permition of COMPAS in the courtroom would only make this worse as it takes a high level of statistical knowledge to interpret, knowledge that would only be available to the best defense lawyers, making the benefits go to those already favored most by society.

---

1. https://link.springer.com/article/10.1007/s00146-023-01676-3 (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↩