HW 4

Pranav Kallem

10/27/2024

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 whether the classifier meets the criterion of equalized odds, we would need information on both the true positive rates and false positive rates across each racial group. Equalized odds is achieved when a classifier provides consistent true positive rates and false positive rates for all groups. This would mean that the likelihoods of a correct and incorrect positive predictions are equal across all racial groups. In this case, the correct positive prediction would be granting a mortgage to a creditworthy applicant, and an incorrect positive prediction would be granting a mortgage to an applicant who will default. Specifically, this requires data on the rate at which each group is correctly classified as creditworthy, which are true positive rates, and the rate of false approvals, which are false positive rates for those who would default. By examining these, one can determine if the classifier's decisions are equitable in terms of equalized odds. This ensures that predictive errors do not disproportionately affect any particular racial group and potentially favor another.

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

The impossibility result, which says that multiple fairness criteria such equalized odds and demographic parity cannot be satisfied at once, would not hold under these two fringe conditions. With a perfect predictor, both true and false positive rates align perfectly with actual outcomes, so fairness metrics are inherently satisfied across groups. Similarly, when class labels are equally distributed across groups, fairness criteria like equalized odds and demographic parity align naturally since each group has the same base rates, avoiding conflicts. These conditions eliminate the trade-offs that typically lead to the impossibility result.

[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

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?

Under Rawls's Veil of Ignorance, a protected class would be defined as a group whose members are shielded from biases based on attributes like race, gender, or socioeconomic status, as these characteristics are ignored to ensure fairness. If we preprocess data by removing this protected variable before training, the variable can still influence results through proxy attributes that correlate with it, such as zip codes or income levels. These proxies allow the protected variable to indirectly affect the algorithm's outcomes, potentially reintroducing biases that impact the protected class even without explicit consideration of the variable.

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 controversial but can be justified if carefully balanced with fairness principles. Statistically, COMPAS can provide consistency in risk assessment, potentially reducing individual bias by offering a data-driven perspective on how likely a convicted criminal is to reoffend. However, fairness issues arise because COMPAS has been shown to have racial bias, failing to meet equalized odds by having different error rates for different racial groups. Philosophically, Veil of Ignorance would argue that justice requires unbiased treatment across demographics, challenging COMPAS's reliability if it indirectly discriminates. However, under utilitarianism, if COMPAS improves public safety overall and judges are trained to interpret its recommendations critically, using it as one tool among many, it may be justifiable as a supplementary measure in decision-making. This use would require transparency and safeguards to ensure it upholds equal treatment and does not undermine judicial fairness.