HW 4

Mackie Jackson

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?

Faber's 2018 analysis showed demographically disparate mortgage approval rates: 71% of whites, 68% of Asians, 63% of Latinos, and 54% of Black individuals were approved. If we assume that a classifier is determining individuals as mortgage-worthy and want to assess the classifier according to equalized odds, we must collect group-specific false negative and positive rates. Equalized odds holds that classification errors are equally distributed across groups, or in this case, racial groups.

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

The impossibility result holds that no more than one of the three fairness metrics (equalized odds, demographic parity, and predictive parity) can hold at the same time for a given classifier. However, when the two aforementioned fringe cases are met, equalized odds would also be met because classification errors (zero in this case) would be the same across the protected variable. Predictive parity would be met because it is a perfect predicting classifier. Finally, demographic parity is met because ground truth class labels are equal 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 our algorithm. How could this variable make its way into our interpretation of results nonetheless?

Rawls' Veil of Ignorance proposes that individuals should make decisions about their society that are dispossessed from their own personal experiences and that have the best possible outcome for all members of a society. To Rawls, a protected class is one that we remove from consideration in classification because we want members of this class to have equal outcomes to those in the non-protected class(es). But even if we remove a protected variable from our training dataset, it might still manifest itself in our interpretation of

[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

results through correlated proxy variables. For example, a mortage qualifier classification algorithm might claim to be race-blind, but may disparately impact Black or Hispanic individuals based on considerations of income and home location.

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.

Utilitarian thought proposes that the correct action is the one that maximizes pleasure, and so the use of COMPAS is correct if it does so as well. In order to determine if a judge's use of COMPAS is morally justifiable, we might simply look to the ratio of correct decisions made with and without algorithmic supplementation. Why? This would measure if the algorithm is maximizing societal pleasure by preventing recidivism more than a judge would unassisted. We might also consider the application of COMPAS as an appeal to fairness as equality, as it will make predictable decisions in a way that individual judges with emotional and personal bias will not. We know that any real classifier is unable to adhere to more than one of the three fairness metrics. In COMPAS's defense, Northpointe has argued that it partially satisfies predictive parity because the risk of recidivism is the same across protected and non-protected classes and it satisfies equalized odds because the difference in white and Black defendant classification error rates are statistically insignificant.