HW 4

Jiangyuan Yuan

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?

We need additional information about the true creditworthiness of each loan applicant, such as the actual outcomes and whether they ultimately repay the loan or default, alongside the classifier's predictions for these applicants. Specifically, for each racial group, we require the rates at which the classifier correctly approves creditworthy applicants (true positive rate) and incorrectly approves non-creditworthy applicants (false positive rate). This means obtaining confusion matrices for each group that detail true positives, false positives, and false negatives. With this data, we can compare the true positive and false positive rates across different racial groups to determine if the classifier's predictions are independent of race given the actual outcome.

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

The impossibility result discussed in class, which states that a classifier cannot simultaneously satisfy all of the fairness criteria, does not hold in the two cases. In the first case, a perfect predictor makes no errors, eliminating the trade-offs between these fairness criteria because the prediction perfectly aligns with the actual outcome Y for all individuals, regardless of the protected attribute A; thus, independence (the prediction is independent of A), separation (the prediction is independent of A given Y), and sufficiency (the outcome is independent of A given the prediction) can all be satisfied simultaneously. In the second case, when the base rates P(Y=1 I A=a) are equal across groups, the statistical dependencies that typically force a trade-off between independence, separation, and sufficiency are absent, allowing a classifier to meet all three fairness criteria at once. Therefore, the conflicts leading to the impossibility result are resolved under these specific conditions. #

[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 is defined as any group distinguished by morally arbitrary characteristics, such as race, gender, or socioeconomic status, that individuals behind the veil would seek to protect to ensure fairness since they do not know which position they will occupy in society. Even if we preprocess data by removing protected class variables before training our algorithm, this variable can still indirectly influence results through correlated features acting as proxies. Proxies are variables that are highly related to protected classes. During lecture, we gave the example that many ZIP codes in the US act as a good proxy for race due to various historical factors that have caused many zip codes to be composed primarily of one race.

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.

Based on statistical and philosophical measures of fairness, the use of COMPAS to supplement a judge's discretion is not justifiable. Statistically, COMPAS has been shown to exhibit biases against certain racial groups, violating fairness criteria due to unequal false positive and false negative rates across these groups. From a virtue ethics standpoint, relying on a biased algorithm undermines the practice of judicial virtues such as fairness and impartiality. Under virtue ethics, Judges, as part of the legal system, should be expected to embody these virtues in their decision-making, and dependence on a flawed tool compromises the ethical integrity expected of the courts. Therefore, employing COMPAS conflicts with practicing virtue ethics. As a side note, I believe a 2% increase in accurately predicting recidivism compared to an average person is far from accurate.