HW 4

Neel Singh

10/29/2024

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?

In order to assess this classifier based on equalized odds, we would need information on the true positive rates for all separate racial groups, false positive rates for all separate racial groups, data for the ground truth on creditworthiness for each applicant, and potential variables that could confound the relationship between race and creditworthiness. This would allow us to determine whether the classifier's errors are distributed roughly equally across groups, satisfying equalized odds.

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

The impossibility result states that it is impossible to simultaneously satisfy multiple fairness criteria in a classification model where the ground truth outcomes differ between protected groups. It does not hold when we have a perfectly predicting classifier or perfectly equal proportions of ground truth class labels across the protected variables.

In the first case, a perfect predictor, or one that has 100% accuracy, inherently meets all fairness criteria because it produces error-free classifications across all groups. By default, this satisfies equalized odds and predictive parity. Equalized odds required a classifier to have roughly equal true positive and false positive rates across groups. If a classifier is 100% accurate, then the true positive rate is automatically 100% for all groups, and false negative rate is automatically 0% for all groups, thus achieving equalized odds. Predictive parity requires that each group has the same probability that a predicted positive outcome is correct. If a classifier is 100% accurate, then all predicted positives outcomes are correct, so this rate is the same for all groups.

In the second case, if the same percentage of positive and negative outcomes exist across demographic, then equalized odds and predictive parity are also automatically satisfied. Since there is no inherent imbalance for

[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

a classifier to deal with, the fairness criteria will be satisfied because the classifier's predictions will reflect the uniformity in the distribution of the data. Since each group has the same ground truth proportions, a classifier than is accurate across groups will have similar true and false positive rates for each group, satisfying equalized odds. By default, if ground truth is the same for all groups, the probability of a true positive is the same for all groups if the classifier is accurate across groups. Both of these conditions remove the conflicts between different fairness constraints, which make the impossibility result null.

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?

Rawl's Veil of Ignorance would define a protected class as any group that could be at risk of discrimination or unequal treatment due to characteristics that are beyond their control. This could include, but is most definitely not limited to, race, gender, socioeconomic status, and disability. If we preprocess data by removing a protected variable before training, this variable could still influence our results through proxies. These proxies are other variables that are correlated with the protected variable. For example, income and education could share patterns with the removed protected variable. Thus, if they are included in the model, the model may be biased against members of the protected class when implemented, even through the protected variable was not explicitly included.

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.

COMPAS is an algorithm that provides judges with insights into how likely recidivism is for a parolee. This algorithm offers the premise of increasing public safety by preventing individuals who are likely to commit recidivism from being released. It also offers the prospect of more standardized judicial decision making across different judges and cases. This can contribute to fairness by reducing individual judge's biases that might affect sentencing decisions. This aligns with fairness as defined through Rawl's Veil of Ignorance. Under the Veil of Ignorance, society would ideally choose principles that are fair and just to everyone, assuming they have an equal chance of ending up in a random position in society. Thus, sentencing decisions should treat individuals as equally as possible, and attempt to standardize treatment of individuals across different judges. By using COMPAS, an algorithm without human biases can provide a standardized recommendation to all judges, contributing to the equality that Rawl's Veil of Ignorance prescribes. While the Veil of Ignorance justifies the implementation of COMPAS, the societal impacts of COMPAS's use are further justified from a utiltarian standpoint. Utilitarianism seeks to maximize the "good" experienced by society. If the COMPAS algorithm can improve public safety through providing more accurate recidivism predictions and preventing crime, then its use can be justified. Prisoners make up a minuscule segment of society. Any "bad" done to prisoners by keeping them locked up is outdone by the "good" done to the rest of society by preventing further crime.