HW 4

Holden Wright

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?

To assess the classifier's fairness according to equalized odds, you would need data on the true positive rate and false positive rate for the different racial groups in the study. Equalized odds requires that the difference in these rates is less than a chosen epsilon for all groups. This means the classifier's odds of correctly identifying creditworthiness and incorrectly granting credit to unworthy applicants should not differ much across racial groups.

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

The classifier correctly labels all cases, so its predictions perfectly match the true labels for every individual. Since both the true positive rate and false positive rate are identical across groups (100% TPR, 0% FPR), equalized odds is met. Since there are perfectly equal proportions of ground truth class labels across the protected variable independence and sufficiency are both met since Y is completely independent of S.

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?

According to Rawls's Veil of Ignorance, a "protected class" would be any group of people who could be disadvantaged in society due to characteristics like race, gender, socioeconomic status, or ability, which are all concealed behind the veil. If we remove the protected attribute, for example race, from our model, proxy

[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

variables (like geographic location or income) might still introduce bias, as they are correlated with race. This effect could still cause bias in the results, showing that just removing the protected class may not entirely eliminate bias.

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 as a supplement to judicial discretion raises both fairness and ethical concerns. A fault of the COMPAS algorithm is that it has higher false-positive rates among minority groups, which violates the principle of equalized odds. Philosophically, utilitarian frameworks may justify its use because of its overall efficiency. Under utilitarianism if its accuracy is above 50% it would provide an overall net benefit of being right more often than being wrong. However, from a deontology perspective it would be unjustifiable, as it disproportionately hurts protected classes by reinforcing bias.