HW 4

Dylan Doby

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?

Student Input.

To assess the classifier according to the equalized odds criterion, it is necessary to gather additional information about both the true positive rates (TPR) and false positive rates (FPR) for each racial group. The TPR represents the proportion of eligible applicants correctly approved, while the FPR represents the proportion of ineligible applicants incorrectly approved. Equalized odds require that the differences in TPRs and FPRs across racial groups are within a small margin, E, to ensure that the classifier's performance does not systematically favor or disadvantage any group. This means we must compare the rates at which applicants from each racial group are correctly and incorrectly approved, considering the ground truth outcomes (whether they are truly eligible or not). Without this data, it's not possible to fully evaluate whether the classifier adheres to the standard of equalized odds and thus meets this measure of algorithmic fairness.

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

Student Input

The impossibility result—where satisfying all 3 fairness criteria (independence, separation, and sufficiency) simultaneously in a single classifier is generally infeasible—does not hold when either of the two fringe cases are met.

Case (a): Perfect Predicting Classifier

In this scenario, the classifier always makes accurate predictions. That is, the predicted outcome, denoted as Y_A_hat, matches the true outcome, denoted as Y, for all observations. With perfect prediction, Y_A_hat

[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

becomes fully dependent on Y, which ensures that the classifier satisfies both separation and sufficiency. Separation is maintained because Y_A_hat remains independent of the protected attribute, S, conditional on the true outcome, Y. Sufficiency is also satisfied because the true outcome, Y, is independent of the protected attribute, S, given the predicted outcome, Y_A_hat. Additionally, independence is inherently satisfied as well, since the predictions Y_A_hat perfectly reflect true eligibility and do not exhibit bias related to S.

Case (b): Perfectly Equal Proportions of Ground Truth Class Labels Across the Protected Variable

When the distribution of true outcomes, Y, is identical across groups defined by the protected attribute, S, it becomes possible to satisfy independence, separation, and sufficiency concurrently. Since the base rates are equal across groups, statistical parity (a form of independence) is maintained by default. In this case, separation and sufficiency are also achievable, as the consistent base rates eliminate disparities in true positive rates and false positive rates across groups. This uniformity in the underlying data distribution allows for a classifier that performs equally well across groups without conflicting fairness criteria.

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?

Student Input

Under Rawls's Veil of Ignorance, a protected class would be defined based on individuals' characteristics that could influence unjust advantages or disadvantages. The Veil of Ignorance principle emphasizes fairness by requiring decision-makers to assume they don't know their own position in society. This would mean treating variables like race, gender, socioeconomic status, and other identity markers as potentially leading to bias or structural disadvantage, thereby classifying them as protected.

Even if a protected variable is removed during preprocessing, its effects can still manifest indirectly through correlated variables. This phenomenon, often referred to as redundant encodings, occurs when other variables in the dataset are highly correlated with the removed protected attribute. For example, a variable like ZIP code can indirectly capture information related to race or socioeconomic status, reintroducing bias into the model's predictions or its interpretation. This suggests that even without explicit consideration of the protected attribute, the classifier could still produce biased outcomes, thus undermining efforts to ensure fairness.

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.

Student Input

The use of COMPAS to supplement a judge's discretion is not justifiable, given both statistical and philosophical concerns about fairness. Statistically, COMPAS fails to satisfy two key fairness criteria: independence (statistical parity and disparate impact) and separation (equalized odds), despite meeting sufficiency. The failure to achieve independence suggests that its predictions systematically disadvantage certain racial

groups, as disparate impact occurs when approval rates differ significantly across protected groups. Similarly, violating equalized odds indicates inconsistencies in true positive and false positive rates across groups, suggesting that the algorithm may unfairly penalize vulnerable populations even when they have similar eligibility to others. This bias remains a significant concern because COMPAS relies on proxies like ZIP code, which can inadvertently reinforce existing social inequities. Philosophically, COMPAS does not align with Rawls's Difference Principle, which requires that any disparities benefit the most vulnerable groups. In practice, its use may exacerbate existing inequalities rather than mitigate them, particularly for racial minorities disproportionately affected by the criminal justice system. From a deontological perspective, the algorithm's black-box nature further violates the moral obligation for transparency and accountability in life-altering decisions. Kantian duty-based ethics demands that individuals be treated as ends in themselves, not merely as means to a broader goal. However, COMPAS reduces defendants to risk scores, denying them the dignity of fully understanding or challenging how their outcomes are determined. Additionally, deontology insists that actions be universally justifiable, but COMPAS's reliance on opaque proxies prevents consistent and fair decision-making across cases. Ultimately, while COMPAS may enhance efficiency, the potential for unjust outcomes and lack of interpretability undermine its ethical use in parole hearings.