HW 4

Uma Nair

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?

It is important to clearly define what exactly equalized odds means in order to correctly answer the question above. Equalized odds is a fairness criterion that requires that a classifier's false positive and false negative rates be the same across different groups (ex., based on race, gender, etc.). In the scenario of banks using AI/predictive statistical models to determine mortgage loan eligibility, concerns about racial discrimination can be particularly significant. According to the equalized odds criterion, a classifier should have the same false positive rate (wrongly approving a loan) and false negative rate (wrongly denying a loan) across different racial groups. It would become problematic if a specific racial group was denied a loan significantly more times than another race, showing disparities and bias in the model. There are multiple times of additional information that would be critical to assessing the classifier. For example, demographic data is needed; there needs to be clear information regarding an applicant's race and socioeconomic status. Socioeconomic status is important since it can contain an applicant's information on income, employment history, and other indicators that might impact creditworthiness. Additionally, data on actual loan performance (such as whether or not the loan was repaid) is important to establish ground truth for evaluating model predictions. Additionally, there needs to be a large enough, adequate, and representative sample size. There should be equal or close to equal proportions of each race in the data sets that the model is being trained on. It is also to underscore the importance of historical data, especially in terms of racial bias. Datasets and data that has been collected over the years may inherently be biased based on time periods and the political/social climate of when it was collected.

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

To argue that the impossibility result does not hold in the fringe cases, we must to examine the conditions under which the impossibility result applies and how these fringe cases circumvent those conditions. The

[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

impossibility result in fairness states that statisticians cannot simultaneously satisfy multiple fairness criteria, such as equalized odds, demographic parity, and individual fairness, in all situations. This issue is particularly prominent for feature variables and skewed observations such as race or gender. A perfect predicting classifier is one that accurately predicts the outcomes for all individuals in the data-set. This means that there are no false positives or false negatives. Since the classifier perfectly predicts outcomes, the false positive and false negative rates for each group would be zero. Therefore, equalized odds is satisfied, and there are no errors to compare across groups. If the model predicts outcomes based solely on valid features, and if these features do not correlate with race or protected attributes, it inherently avoids discrimination. In the case of a classifier with perfect prediction accuracy, the concept of the impossibility result does not hold, since the classifier can be accurate and fair across all groups. In the case of perfectly equal proportions of ground truth class labels, this concept refers to a situation where the distribution of class labels (ex., approved vs. denied) is perfectly equal across different groups defined by a protected variable (ex., race, gender). If the ground truth class labels are equally distributed across groups, then any classifier that randomly assigns outcomes in accordance with these proportions will satisfy demographic parity. Since each group has the same proportion of positive outcomes, the classifier will be designed to maintain this balance.In this scenario, if the classifier is also trained on features that do not introduce bias, the false positive and false negative rates can be aligned across groups, thus satisfying equalized odds as well. With an equal distribution of outcomes across groups, it is possible to construct a model that meets both demographic parity and equalized odds. The impossibility result fails because the assumption that the groups have different class distributions does not hold. With that being said, the impossibility theorem does not apply to either fringe case.

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?

Rawls's Veil of Ignorance is a philosophical concept that suggests we should design societal rules without knowing our own place within that society (e.g., race, gender, socioeconomic status). Under this veil, decision-makers would aim to create fair principles that protect the most minority groups, leading to an equitable society. Essentially, one must strip themselves of all characteristics that could potentially cause them to make biased decisions in order to make fair decisions for the good of society and those who are already disadvantaged. In terms of the Veil of Ignorance, a protected class would be defined by characteristics that are vulnerable to discrimination or that might lead to unfair and unequal treatment in society. This includes groups distinguished by attributes such as race, gender, socioeconomic status, age, sexual orientation, disability, etc. The primary goal of the Veil of Ignorance is to prevent any discrimination caused by bias from any one of these protected classes, ensuring a fair and just society. When we preprocess data for a model by removing a protected variable, we aim to eliminate potential bias in the algorithm's training phase. However, there are several ways in which this variable can still influence our results, such as through proxy variables, historical bias in data, and the interpretation and deployment of the model. Other features in the data-set may serve as proxies for the removed protected variable. For example, zip code or income level could correlate strongly with race, allowing for indirect discrimination. The algorithm might learn patterns associated with these proxy features that still reflect biases against the protected class. If the training data contains historical bias (ex., past discriminatory practices, such as redlining/gerrymandering), the model can learn these biased patterns even without explicit references to the protected variable. This might result in biased predictions that disadvantage certain groups. In regards to the deployment and interpretability of the model, decision-makers might inadvertently introduce biases based on their interpretations of the results. If certain outcomes align with societal stereotypes or prejudices, those biases may be projected even if the algorithm was initially trained without the protected variable. This is why attention to detail and caution are necessary when interpreting algorithmic results to ensure that they align with the principles of fairness and justice.

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 to supplement a judge's discretion is not justifiable, as it risks exacerbating existing biases and undermining the principles of fairness and justice. Statistically, COMPAS has been criticized for its lack of transparency and the potential for biased outcomes, particularly regarding racial disparities in risk assessments. This raises significant concerns about equalized odds, as minority groups may be disproportionately flagged as high-risk based on flawed data. Philosophically, we must prioritize the protection of the most disadvantaged; relying on COMPAS could lead to decisions that perpetuate systemic injustices rather than rectify them. While the intent to use data-driven tools in judicial decision-making should be praised, the inherent risks of bias and inequality make COMPAS an inappropriate supplement to judicial discretion. Although that COMPAS should be used as a supplement, it is difficult to say whether or not judges wil use the algorithm to override the algorithm's decision (even though in particular scenarios, it may be more accurate than the judge), perpetuate or affirm their own biases, or the judge may base their decision solely on COMPAS's output for the sake of efficiency. When dealing with models, transparency is very important. A company or whoever creates a particular model should be very transparent about the possible limitations and biases that the model can perpetuate. It is clear that Northpointe did not properly inform judges and those in the legal system that would be using their classification algorithm, which goes against the idea of fairness and transparency in machine learning. A judge is better off going with their decision, as they can sympathize with whoever is on trial and have years of experience, training, and expertise to support their decision. In terms of ethics, COMPAS goes against both maxims of Deontology. It would be difficult to say it is acceptable to universalize the usage of COMPAS, since it is not accurate enough of the time. Additionally, all of the individuals included in the data-sets used to train COMPAS were used as a means to an end to incriminate other people who were also in the same situation as them, most likely without consent. Deontology states that actions must be evaluated based on whether they respect the rights and dignity of all individuals, regardless of the consequences. The reliance on COMPAS, an algorithm that has been shown to perpetuate biases against marginalized groups, undermining this principle by potentially violating the rights of individuals unfairly labeled as high-risk due to an inaccurate classification. Moreover, Deontology stresses the importance of transparency and accountability in decision-making processes. Since COMPAS's data is "black-box", this goes against this principle of transparency and accountability that is crucial for models, as mentioned before.