HW 5

Ava Klissouras

11/4/2024

The decision of whether or not to utilize the COMPAS algorithm is a difficult and complex one that hinges upon a variety of evidence, both statistical and philosophical. Ultimately, when considering this evidence, it is my professional opinion that the algorithm should not be used, as it violates measures of statistical fairness and philosophical moral criteria, and reduces the transparency of decisions made in court, ultimately undermining the design of the United States legal system.

The COMPAS algorithm may initially seem appealing, as it appears to have the capacity to make a judge's life easier. After all, having a machine learning companion to "check your work" or guide you to the correct choice in a particularly difficult case might seem quite useful. However, consider that the algorithm is not always perfectly reliable. While it is true that no Machine Learning algorithm is perfect, the COMPAS algorithm only has around a 60% accuracy rate. Typically, for Machine Learning algorithms, an accuracy of 80-90% is preferred. Consider that flipping a coin would yield 50% accuracy. Would you allow the outcome of a flipped coin to persuade your decision whether or not to convict a defendant? I would assume the answer is "no." Would you place a defendant's fate in the hands of an algorithm with only 10% higher accuracy than the flip of a coin?

Not only is the COMPAS algorithm's accuracy less-than-ideal, but it disproportionately makes mistakes when it comes to predictions for Black individuals, predicting that they will recidivate at a higher rate than they actually do. To understand how one determines this, we must first dive into a bit of statistical theory– just enough to give context to our assertion that the algorithm disparately harms Black individuals.

You might ask, how could this algorithm predict recidivism rates differently across different races, if race is not included as a variable in the data used to train the algorithm? There may be other variables (called *proxies*) in the data– for instance, zip code– that, while not directly indicating race, may be closely correlated with it. The algorithm can rely on these variables in generating a prediction, causing the prediction to inadvertently take race into account without its explicit presence in the data.

The first measure of statistical fairness that our algorithm violates is *independence.* Statistical independence is violated when an algorithm's prediction is not independent of a "protected variable"– namely, a variable that is *not* supposed to be used in determining the outcome. In our case, this variable is race. In the case of COMPAS, two specific measures of independence were violated. The first is *disparate impact*, which is violated when the probability of being flagged for recidivism when you are not in the protected class (White), divided by the probability of being flagged for recidivism when you are in the protected class (Black), falls short of a certain threshold (typically 0.8). Another type is *statistical parity*, which is violated when the difference between the aforementioned probabilities (rather than the quotient) exceeds a certain threshold (typically 0.2).

You might correctly argue, what if the "true" proportions of recidivism differ in Black and White individuals? Then wouldn't independence be violated, no matter how perfectly the algorithm performs? This is where another measure of statistical fairness comes in, called *separation.* Separation is violated when the algorithm's prediction is not independent of race, conditioned on whether someone has truly recidivated (their "real-life" action, not the algorithm's prediction). One measure of separation is *equalized odds*, which is the false positive rate in the protected class (Black), minus the false positive rate in the non-protected class (White). You could interpret a false positive as predicting that someone would recidivate when they actually did not.

If this difference exceeds a certain threshold (usually 0.2), the algorithm is said to violate separation. This is what happened for the COMPAS algorithm.

You may have heard that the COMPAS algorithm was deemed legal due to something called the *Incompleteness Theorem.* This is because there is a third measure of statistical fairness that the COMPAS algorithm *does* satisfy, called *sufficiency.* Sufficiency is satisfied when whether someone has truly recidivated is independent of their race, conditioned upon the algorithm's prediction. This theorem states that it is impossible to simultaneously satisfy independence, separation, and sufficiency unless a classifier is perfect and the proportions of recidivism are exactly the same across all races- so of *course* the COMPAS algorithm cannot satisfy all of them. However, this is where we must apply some philosophical knowledge, which, when combined with the fact that the COMPAS algorithm violates two of the three measures of statistical fairness, may help you understand why the use of the COMPAS algorithm is still not ideal.

The COMPAS algorithm violates a philosophical measure of fairness called "fairness as equality." Fairness as equality states that the benefits of a good should be allocated equally to everyone. The US legal system is supposed to treat everyone fairly and confer the same rights to all citizens, as detailed in the U.S. Constitution; therefore, one would expect that any algorithm used as a tool in court should do the same. However, the COMPAS algorithm over-predicts recidivism for Black individuals at a greater rate than for White individuals, thereby conferring less benefits to Black individuals than White individuals and undermining the philosophical measure of fairness that arguably influences the design of the United States legal system. All individuals are entitled to the same benefits, such as a fair trial, the right to counsel, and innocence until proven guilty, regardless of race. Yet, the COMPAS algorithm undermines this principle by conferring benefits disparately across races.

The use of the COMPAS algorithm also violates a moral framework called Deontology. Deontologists believe that the morality of an action is governed by the intention behind it. More specifically, an act is moral if it (1) can be universalized without logical contradiction (i.e. if everyone did it, would it still be useful/helpful/logical?), and (2) treats people as ends, rather than mere means (i.e. does not "use" people). The COMPAS algorithm violates the first criterion. Suppose the appeal of using COMPAS is increasing efficiency in court (we already know it does not decrease bias, so arguably its merit is the potential to help a judge quickly make decisions instead of deliberating over them). If every judge used COMPAS to improve the speed/ease with which they make decisions, the use of COMPAS would not confer an advantage. Simply, court cases would all take less time and/or brainpower to settle, but because of this, more court cases could be squeezed into a day's time, and judges would have the same amount of work as they did before. Therefore, we reach a logical contradiction. As for the second criterion, individuals who have previously recidivated (or not recidivated) are used as means rather than ends. These individuals unknowingly give up their personal data, such as zip-code, age, and gender, to be used in the model, without receiving any benefits in exchange for their use.

All in all, based on statistical measures of fairness, philosophical measures of fairness, and a philosophical moral framework, the use of the COMPAS algorithm is arguably unjust. If my above explanations have not swayed you, consider the design of the United States legal system, of which you are an integral part. Judges must make a decision based on evidence synthesized in a standardized way, adhering to a process intended to make the legal system fair for everyone. However, the COMPAS algorithm is what we call a "black box"– we do not have access to exactly how the algorithm makes a prediction, and what variables it includes/excludes (apart from that it does not directly take race into account). Does using such a nebulous decision-making method undermine the meticulous construction of our legal system? The use of the COMPAS algorithm does have some benefits; it can increase efficiency, provide a way for judges to "check their work," and does not violate *all* measures of statistical fairness. It also might circumvent the sometimes-flawed moral intuition that may lead a judge to a non-objective decision. This is why some people believe it is perfectly reasonable to use. In the end, it is up to you to decide, but I hope that my advice has been helpful.