Midterm Project

Musie Gebreegziabher

03/22/2024

Introduction

With the rapid expansion of machine learning, and in alignment with the themes discussed in this paper, the definition of what constitutes statistical and philosophical fairness in policy and decision-making environments has become increasingly divergent. One example brought up in the paper was an AI system used to measure "recidivism risk" in bail decisions. What researchers found was that the system was bias against certain racial groups. This example, among others, illustrates a current flaw in AI and machine learning systems perpetuating societal errors from both recent and distant history.

This paper, titled 'Statistical Equity: A Fairness Classification Objective,' presents an interesting stance and methodology in machine learning. By implementing an augmented version of statistical parity that considers present outcomes, this paper adopts an approach that acknowledges historical biases inherent in the data. It combines the principles of equity and affirmative action, while also ensuring equality among different groups. It advocates for compensating historical biases and leveraging disadvantaged groups, offering a unique approach to addressing equity concerns and diverging from definitions such as "differential privacy" and "fairness through unawareness" due to the requirement of sensitive information.

This methodology holds an interesting stance by questioning how a classifier can generate predictions that benefit the majority. It offers a fresh perspective that both statistically and philosophically advocates for equity, wherein the model provides leverage to ensure groups receive appropriate resources to pursue their goals. This paper introduces the concept of fairness termed "Statistical Equity," defined as follows: "A predictor is considered statistically equitable among demographic groups if it satisfies the equation $p(\hat{Y} |A = a) + p(Y |A = a) = p(\hat{Y} |A = b) + p(Y |A = b)$." That is, statistical equity is achieved when the sum of predicted probabilities and actual outcomes for positive events is equal across different demographic groups.

Methods and Results

In evaluating equity as a fairness definition statistically, the paper uses a fairness gain equation where the fairness gain for a given loss function (Equity, Parity, Classification loss) is determined relative to a basic classifier. In other words, the paper states, "measures how effective a method was in reducing disparities among demographics compared to a classifier with no constraints."

The paper employs an objective function comprising two components combined: the fairness objective, aimed at promoting equity among groups, and the classification objective, which utilizes loss functions to optimize predictive accuracy. The experiment was conducted on three loss functions, equity loss, representing the paper's proposed approach, aiming to rectify historical biases by equalizing the sum of historical and future outcomes between groups, effectively enforcing affirmative action to achieve equilibrium in the objective function. Parity loss which mirrors the statistical parity of notion discussed in the discussion in which historical biases are not taken into consideration. Finally, classification loss(cross-entropy) which is a loss function that contains only the cross-entropy loss with no fairness gains. In the classification objectives, there is a hyperparameter that controls the importance of the fairness constraint over the classification objective. It was hypothesized that the Equity classification objective would achieve the highest gain in fairness and that the higher the hyperparameter value the more we can expect fairness gain and accuracy degrade.

The experiment was conducted on two datasets, the COMPAS and Adult dataset. The COMPAS dataset contains information about defendants from Broward County. In this dataset the sensitive attribute was gender which was used to predict if a criminal would re-offend in two years. The Adult dataset contains information about individuals who either make 50k a year or more or not and gender was the protected attribute in the classification loss used to predict if a criminal would reoffend in two years. In both datasets, the paper found that using the equity loss in classification will result in gain in fairness. It was also found that the higher the hyperparameter value the more degrade there was in accuracy more gain in fairness. However, through MannWhitney U significance test, it was shown insignificant in the COMPAS dataset for low to mid hyperparameter values. For the Adult dataset, although the degrade in accuracy was shown to be statistically significant, the test accuracy loss was reasonable considering the price of fairness we get through the gain in fairness. From the overall results, hyperparameter values between 0.3-0.5 when using the Equity objective show to be the most effective in terms of gain in fairness and maintaining a reasonable test accuracy.

Additionally, the paper tested the effect of the statistical equity model on the feedback loop. This loop is a phenomenon in which biased information is magnified through different systems, amplifying historical biases. The results of this showed that using the paper's statistical equity notion can bring equality, equity, and fairness in long run and mitigate the negative effects of the feedback loop phenomenon.

In a more philosophical approach, the paper also looks a public perception of Equity through 150 workers taking a survey. In the surveys, participants were asked to go through four scenarios and select their preferred fairness solution for each scenario. In Scenario 1, participants were asked to rate pictures of equity and equality and chose their preferred picture. In Scenario 2, participants were asked to s rate loan distribution mechanisms. One is based on equity, which considers each student's past history of receiving a scholarship (equity). Another simply proposes to equally distribute the loan among all the students (parity). In scenario 3, respondents were asked to rate the government subsidized housing distribution systems proposed in the survey— one based on equity considering how houses were historically distributed across different races (equity). The other proposes to equally distribute houses across different racial categories (parity). In scenario 4, respondents were asked to rate college admission systems—one based on equity considering if the student is a first-generation college student (equity). The other equally admits students from first generation and non-first generation backgrounds (parity). The results of this showed that there are some cases in which our notion of fairness is strongly preferred by a large margin, and some other cases where preference is given to the parity notion. It was evident that strong preference is given to the notion introduced in the paper for scenarios 1 and 2, and despite the fact that scenarios 3 and 4 are not over-preferred for the notion, there are still considerable number of people who gave preference to the notion in these scenarios.

Normative Consideration

The statistical and philosophical notion of fairness in the rapid expansion of machine learning algorithms in policy and decision-making environments and in general is one of huge/top normative concern. In the example earlier of an AI system used to measure "recidivism risk" in bail decisions, we see the societal and personal impacts these systems can have. It's important to define what is just and equal, and although this paper does not offer a comprehensive answer, it serves as a significant step in the right direction. Moreover, the paper acknowledges the need for future research to delve deeper into how the concept of equity intersects with other established definitions of fairness.