Inverse Probability Weighting for Fairness in Binary Classification Algorithms

Brian Cho
Yale University, S&DS Senior Project

Abstract

We introduce the use of inverse probability weighting (IPW) with propensity scores as a potential method for algorithmic fairness. In this study, we focus on binary classification algorithms with a binary sensitive attribute, such as race or gender. We discuss the results of this re-weighting from both statistical and causal perspectives of fairness, contextualize this method with existing fairness re-weighting methods, and demonstrate its key theoretical properties through simple simulations. While this method has inherent limitations, it introduces a novel method for data pre-processing in the context of algorithmic fairness, and offers a way to remove specific effects of S on other variables in the training data.

Background

Problem Setup

- S is a binary sensitive variable, such as race, gender, or religion
- X are other covariates present.
- Y is a binary outcome variable.

Propensity Scores and Inverse Probability Weighting

Let Z denote a variable in $\{X, Y\}$.

$$e(Z) = \mathbb{P}[S = 1 | Z = z] = \mathbb{E}[S | Z = z]$$

For each observation in the training data, we assign the following weight:

$$W_i = \frac{S_i}{e(Z_i)} + \frac{1 - S_i}{1 - e(Z_i)}$$

Fairness Measurements

Measure	Definition
Calibration.P	$\mathbb{E}[Y \mid S=1, \hat{Y}=1] - \mathbb{E}[Y \mid S=0, \hat{Y}=1]$
Calibration.N	$\mathbb{E}[Y \mid S=1, \hat{Y}=0] - \mathbb{E}[Y \mid S=0, \hat{Y}=0]$
FalsePos	$\mathbb{E}[\hat{Y} \mid S=1, Y=0] - \mathbb{E}[\hat{Y} \mid S=0, Y=0]$
FalseNeg	$\mathbb{E}[\hat{Y} \mid S=1, Y=1] - \mathbb{E}[\hat{Y} \mid S=0, Y=1]$
DemoPar	$\mathbb{E}[\hat{Y} \mid S=1] - \mathbb{E}[\hat{Y} \mid S=0]$

Table 1: Statistical Measures for Fairness (Disparate Impact)

Key Results of IPW

IPW with $e(Z)$ results in **independence** between the variable Z_r and the sensitive attribute S_r in the re-weighted training data.

IPW with $e(Z)$ **removes the total effect** of the sensitive attribute S_r on the variable Z_r in the re-weighted training data.

Base Re-weighting and IPW with $e(Y)$

Base Re-weighting

$$W_i^B = \frac{P_{des}}{P_{obs}} \approx \frac{\mathbb{P}[Y = y_i]\mathbb{P}[S = s_i]}{\mathbb{P}[Y = y_i, S = s_i]}$$

IPW with $e(Y)$

$$W_{S\sim Y} = \frac{1}{\mathbb{P}[S = s_i]} * \frac{\mathbb{P}[Y = y_i] * \mathbb{P}[S = s_i]}{\mathbb{P}[S = s_i, Y = y_i]} = \frac{1}{\mathbb{P}[S = s_i]} * W_{base}$$

Data-Generating Process For Figure 3,4

- $S \sim Bern(p), \quad p \in [0,1]$
- $X \sim N(0,1)$
- $Y \sim \begin{cases} Bern(0.9) & S = 1 \\ Bern(0.1) & S = 0 \end{cases}$

Compared to the existing re-weighting method, IPW with $e(Y)$ only differs by a single term. The less likely group of S is upweighted so that the two groups of S have equal weights.

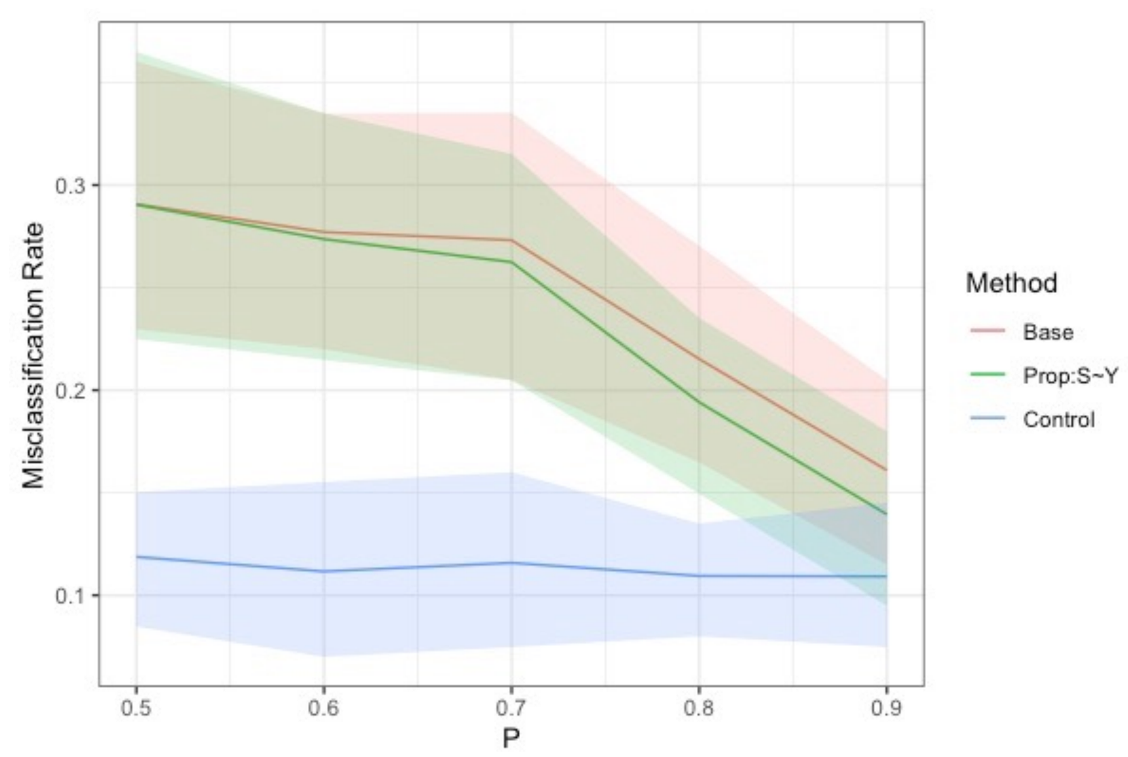


Figure 3: Misclassification Rates (with 90% CF) for Base Re-weighting, IPW with $e(Y)$, and no re-weighting (control)

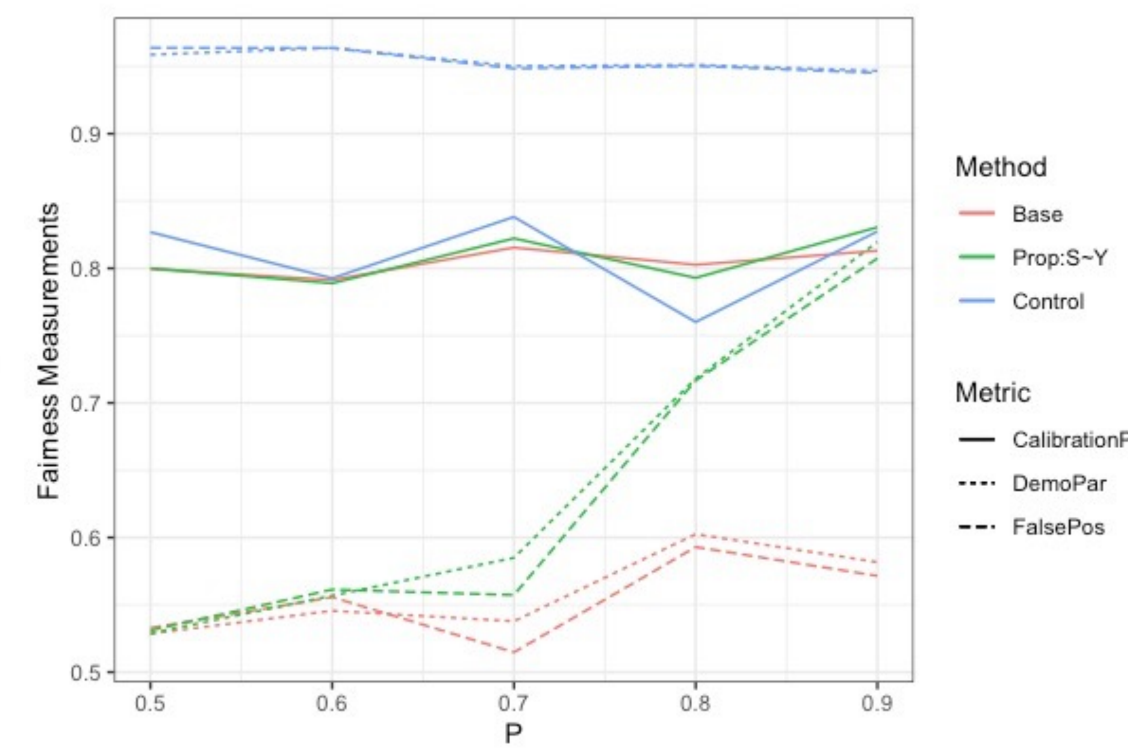
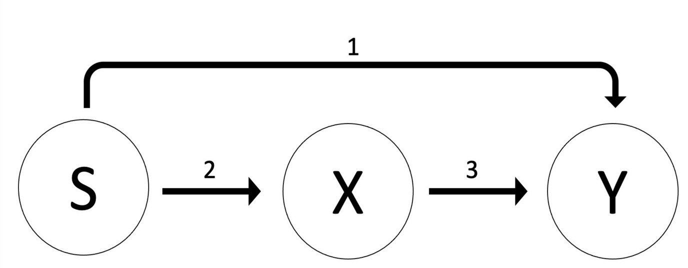


Figure 4: Fairness Metrics Across Base Re-weighting, IPW with $e(Y)$, and no re-weighting (control)

The Case of Mediation: IPW with $e(Y)$ and $e(X)$

Data-Generating Process For Figure 5, 6

- $S \sim Bern(0.5)$
- $X \sim \begin{cases} Bern(0.8) & S = 1 \\ Bern(0.2) & S = 0 \end{cases}$
- $Y \sim Bern(0.8(j * X + (1-j) * S)),\ j \in [0,1]$



As j increases, the indirect effect of S on Y (pathways 2 and 3) increases, and the direct effect of S on Y (pathway 1) decreases. Mediation offers a simple example where IPW with $e(Y)$ and $e(X)$ behave similarly.

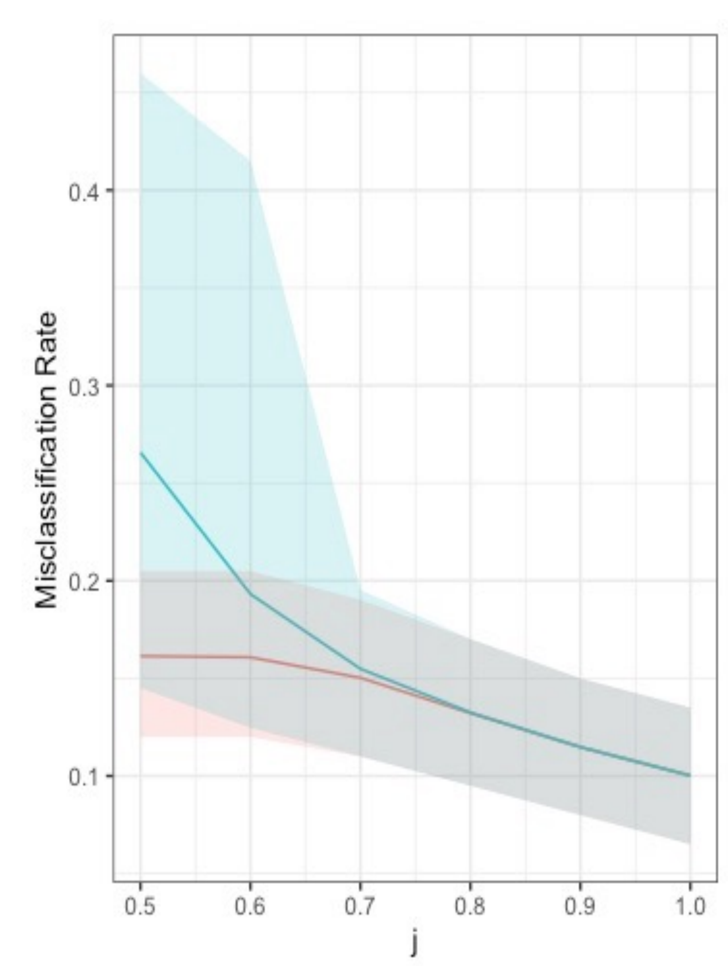


Figure 5: Misclassification Rates for IPW with $e(Y)$ and IPW with $e(X)$

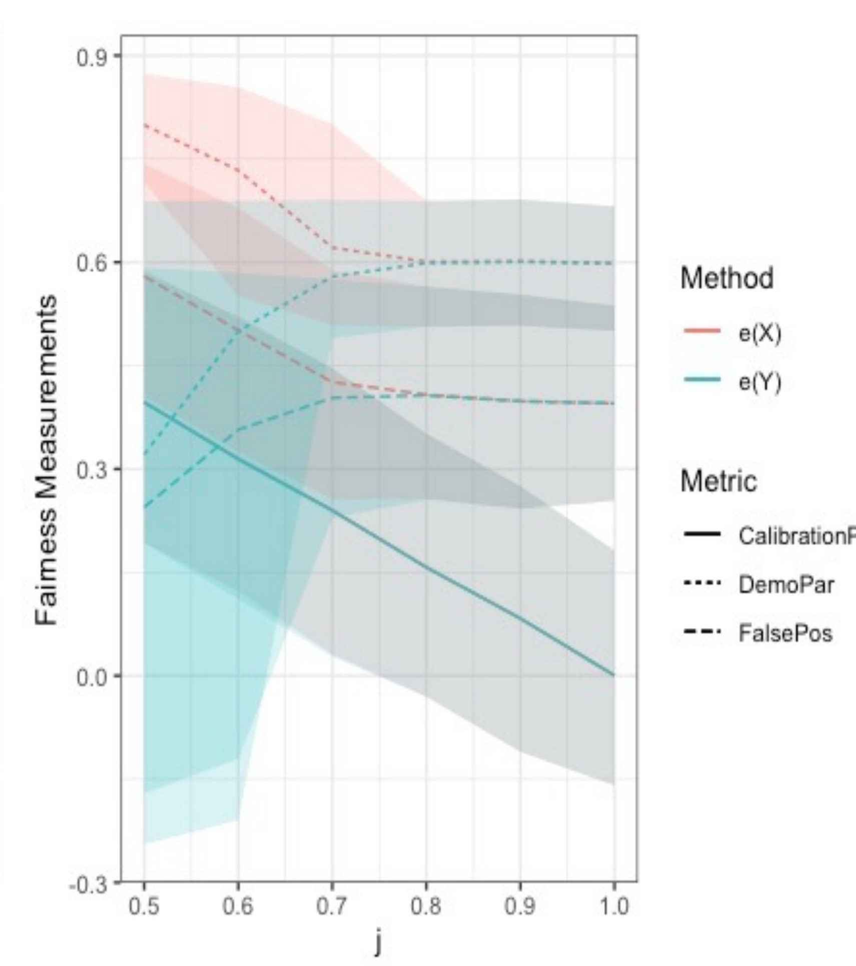


Figure 6: Fairness Metrics for IPW with $e(Y)$ and IPW with $e(X)$

Limitations

Necessary Assumptions

1. The correct propensity scores are needed for these results to hold.
2. All subjects have a nonzero probability of S=1 or S=0.
3. We require a representative sample of the population.

Use-Case Considerations

- Note that re-weighting may sacrifice accuracy for fairness, as we see in the simple simulations. Whether or not this is desirable is highly dependent on the algorithm's use-case.
- While IPW re-weighting guarantees certain properties, it may also change other relationships in the data depending on the choice of Z.

Conclusions

- **Key Results.** With the necessary assumptions, IPW with propensity scores removes the relationship between S and Z in the training data by making these variables independent and removing the total effect of S on Z.
- **Relationship with existing methods**. Compared to the existing re-weighting method proposed by Kamiran et. al, IPW with $e(Y)$ assigns weights that only differ by a single term. Other methods of IPW with a different choice of Z produce different results.
- **Limitations of This Method.** These IPW methods require strong assumptions and may not be suitable for certain use-cases.

Acknowledgements

My senior project was completed with the support of my advisor, Professor Jasjeet Sekhon, and Theo Saarinen. Their guidance and contributions to this project were invaluable towards its completion.