Mitigating Algorithmic Bias

Agnes Xu

May 16, 2024

Background: Bail Decision and COMPAS

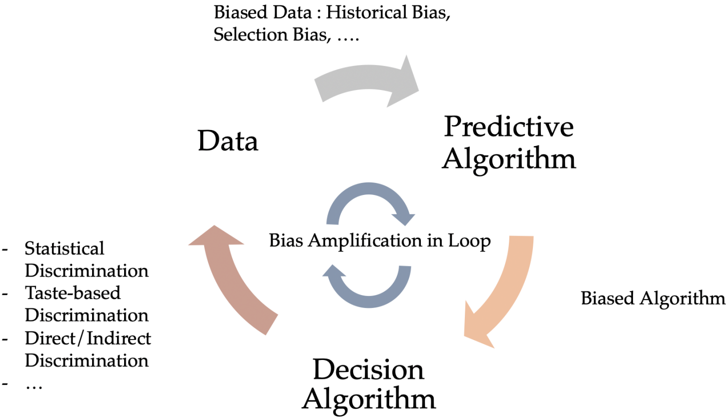
- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is one of the most popular scores used nationwide and is increasingly being used in pretrial and sentencing, the so-called "front-end" of the criminal justice system.

- Studies have found that, when using COMPAS, black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk. (unequal false positive rate)

- Discussion has been going on about algorithmic bias problem like this, including what is fairness, why the algorithm is biased, and how to mitigate such bias.

Background: Algorithmic Bias

How the bias occur and accumulate overtime



Biased Data : Historical Bias, Selection Bias, ….

Data

Predictive Algorithm

Bias Amplification in Loop

Biased Algorithm

Decision Algorithm

- Statistical Discrimination
- Taste-based Discrimination
- Direct/Indirect Discrimination
- …

Background: Algorithmic Bias

- *Bias-in-Bias-out* (data → predictive algorithm): Training data is corrupted by historical bias and bias happening in the measuring, collecting and labelling process. (→ pre-process techniques)

- *Algorithmic Bias* (predictive algorithm → decision algorithm) : Predictive algorithm performing biased towards some unfavoured group. (→ in- and post- process techniques)

- *Feedback Loop* : Enhance and spread bias. (→ break the loop: collect random data / reinforcement learning)

Background: Fairness

Different notions of predictive fairness fall into different level of population:

- *Individual Fairness* : Give similar predictions to similar people
- *Group Fairness* : Treat different groups equally
- *Subgroup Fairness* : Balancing fairness at group and individual level by picking a group level fairness constraint and test it over a large collection of subgroups.

Background: Group-level Fairness

group fairness is often framed in terms of protected attributes while allowing for differing treatment based on a set of qualifications.

Notion	Formula
Demographic Parity	$P(\hat{Y} \mid A=0) = P(\hat{Y} \mid A=1)$
Conditional Parity[1]	$\Pr\left(\hat{Y}=1 \mid A=a, \mathbf{z}=z\right) = \Pr\left(\hat{Y}=1 \mid A=a', \mathbf{z}=z\right)$
Equalized Odds[2]	$P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=0)$
Calibration	$P(\hat{Y}=1 \mid A=a, Y=y) = P(Y=1 \mid A=a, Y=y)$

- **Incompatibility** : Kleinberg et al. (2016), Pleiss et al. (2017)
- **Motivation** : Proportional representation is often desirable its own right; alternatively, the absence of proportional allocation of goods can signal discrimination in the allocation process, typically against historically mistreated or under-represented groups. (Dwork and Ilvento, 2018)
- **Critics** : One can maintain group level fairness while discriminate against individuals through redlining and other strategic design.

[1] Ritov et al. (2017)

[2] and a bunch of balance conditions for terms in confusion matrix

Mitigating Bias

Existing methods to obtain fair predictive models can be classified by the stage of training where they try to intervene:

- **Pre-process**
- **In-process**
- **Post-process**

Or the class of algorithms:

- **Regularization**
- **Adversarial Learning**
- **Representation Learning**
- **Causal Modeling**

Pre-process

Methods in this category include changing the training data before feeding it into a machine learning algorithm and simulating debiased data:

- **Re-weigh or adjust labels** : Pedreschi et al. (2008), Kamiran and Calders (2009), Luong et al. (2011), Kamiran and Calders (2012)
- **Modify feature representations** : Zemel et al. (2013), Feldman et al. (2015), Louizos et al. (2016), Calmon et al. (2017), Samadi et al. (2018)
- **Generate fair synthetic data** : Xu et al. (2018)

In-process

Main works on algorithmic fairness focus on the training process, where fairness notions are used as constraint in optimization problem or regularization term

- **Fairness as Optimization Constraint** : Zafar et al. (2017), Bechavod and Ligett (2010), Woodworth et al. (2017)
- **Fairness as Regularizer** : Kamishima et al. (2011), Bechavod and Ligett (2017), Quadrianto and Sharmanska.(2017)
- **Group-specific Classifier** : Dwork et al. (2018)
- **Tree-based Method** : Kamiran et al. (2010), Kamiran et al. (2013), Valdivia et al. (2020), Raff et al. (2017), Fantin (2020)

Research Question

- Can simply removing sensitive variables really solve the bias? Is there any chance of reconstructing those variables from remaining variables?
- Which measurement of fairness makes more sense? Is there a "hyper" decision rule for the fairness criteria?
- How can we improve the model without directly involving sensitive variable and incorporating our knowledge in the training process?
  - Representation learning methods like PCA and auto-encoders to recode the features
  - Include fairness decision standard through ensemble learning with a human decision model.
  - Include the fairness notions into the penalty term.
  - Regularize the predictive algorithm with l1 error and the distance of the statistical model and human decision model.

What to expect

The previous research used theory of fairness as decision rule or constraint, and we're trying to directly let the theory incorporated in the training process as a penalty term, adjusted embedded variables, or part of the ensemble models. In this way we could expect lower cost at the decision level and higher generalizability.

Data

Statistical Modelling:

1. COMPAS(Correctional Offender Management Profiling for Alternative Sanctions): Includes race, age, sex, number of prior convictions, and COMPAS violent crime risk score (a discrete score between 1 and 10)

2. MORF (Massive Open Online Courses Replication Framework)[3] : Experiment data of online learner's features and whether they finish the MOOC.

Modelling Human Decision:

- Online experiment: Asking people to score the likelihood of defendants commiting crime or application candidates performing well after being enrolled. Interviewee are shown generated features of potential agents in the decision making scenario.

[3] Gardner and colleagues (2019)

Fair Representation Learning

dimension reduction methods, protected attributes as sensitive variable correlated with objective

Zemel et al. (2013) : fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group.

Samadi et al. (2018) : fair PCA, to maintain similar fidelity for different groups and populations in the dataset in PCA pre-process.

Fair VAE

protected attributes as nuisance variables

Louizos et al. (2016) : introduce an additional penalty term based on the “Maximum Mean Discrepancy” (MMD) penalizing difference between variational posterior $q(\mathbf{z} \mid \mathbf{x}, A = 0)$ and $q(\mathbf{z} \mid \mathbf{x}, A = 1)$.

Liu and Burke (2018) : representation-learning task as an optimization objective that minimize the loss of the mutual information between the encoding and the sensitive variable.

Davies and Goel (2018) : flexibly fair representation learning by disentangling information from multiple sensitive attributes.

Regression with Fairness Regularizer

In-process, regularization

Berk et al. (2017) : Three fairness regularizer for (linear and logistic) regression and by varying the weight on the fairness regularizer, we can compute the efficient fronter of the accuracy-fairness trade-off which is measured by *Cost of Fairness (PoF)*

Fairness as Constraint Optimization

Zafar et al. (2017) : introduced measure of decision boundary fairness and mechanism to maximize accuracy of convex margin-based fair classifiers (logistic regression and SVM) with disparate treatment and disparate impact constraint.

Woodworth et al. (2017) : estimate an empirical risk minimizing predictor $\hat{Y}$ subject to approximate non-discrimination constraint (relaxation of equalized odds based only on a second-moment condition instead of full conditional independence)

Methodology: Regularization

- Suppose the statistical model is $f(x)$ and the average predicted difference between two models is $p(x)$, then regularize $h(x) = f(x) + ap(x)$, where a is a hyperparameter to be tuned.
- In this way the final model is trained to be including the knowledge from theoretical model to produce more robust result against model misspecification.
- We can apply the similar logic here by setting $p(x)$ to be some adjusted outcome under selected fairness notions or human decision model, and tune the algorithm in the direction that not only reduces variance and bias, but also the deviation from fair status.

Methodology: Ensembling

- Suppose the statistical model is $f(x)$ and the human decision or theoretical model is $p_i(x)$. Here we allow more than one potential human decision models. Then the linear ensemble learning predictor is $\alpha f(x) + \sum_{i=1}^{N} \beta_i p_i(x)$, where $\alpha + \sum_{i=1}^{N} \beta_i = 1$. The hyperparameters will be gained using cross-fitting.
- Or we can nonparametrically ensemble the models using random forest, as would allow for varying weights in the independent variable space.
- This would guarantee that, even if all candidate models are misspecified, the ensembled model will outperform all of them.
- Inspired by the generalized doubly robust estimator, we can also combine parametric models using their GMM distance.

Methodology: Generalized Doubly Robust

Suppose $m \in \{f, g\}$, the true parameters satisfy a set of $\ell_m \times 1, \ell_m > p_m$ moment conditions:

$$\mathbb{E}\left[\psi_m\left(x, y; \theta_m\right)\right] = 0$$

Given a sample of n i.i.d observations, we can construct the following (adjusted) moment distance functions:

$$Q_m\left(\theta_m\right) = \kappa_m^{-1} \bar{\psi}_m\left(\theta_m\right)' \Omega_m \bar{\psi}_m\left(\theta_m\right)$$

, where $\bar{\psi}_m\left(\theta_m\right) \doteq \frac{1}{n} \sum_{i=1}^n \psi_m\left(x_i, y_i; \theta_m\right), \Omega_m$ is a $\ell_m \times \ell_m$ positive definite weight matrix, and $\kappa_m = \ell_m - p_m$ is the degree of freedom of the χ^2 statistic that the unadjusted Q_m equals if m is the true model.

Methodology: Generalized Doubly Robust

$$h(\widehat{f}(x), \widehat{g}(x); w) = w_f f(x) + w_p p(x)$$

where

$$w_f = \frac{Q_p\left(\hat{\theta}_p\right)}{Q_f\left(\hat{\theta}_{\mathcal{M}}\right) + Q_p\left(\hat{\theta}_p\right)}, w_p = 1 - w_f$$

Methodology: Linear Ensemble

$$h(\widehat{f}(x), \widehat{p}(x); w) = w_0 + w_1 \widehat{f}(x) + w_2 \widehat{p}(x)$$

Stacking: Use LOOCV to find the optimal $w = (w_0, w_1, w_2)$

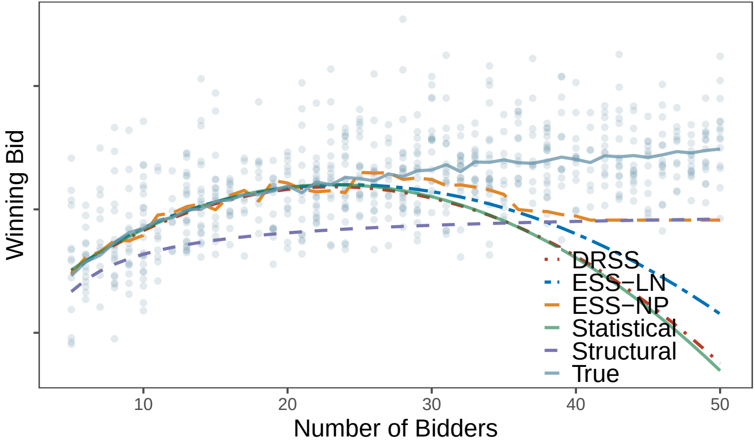
$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - w_0 - w_1 \widehat{f}^{-i}(x_i) - w_2 \widehat{p}^{-i}(x_i) \right]^2$$

- Under a simplex constraint on w, i.e $w_j \geqslant 0, \sum_j w_j = 1$ it is the jacknife model averaging method of Hansen and Racine (2012).
- Can use sample spliting or cross-fitting in place of LOOCV.

Methodology: Non-parametric

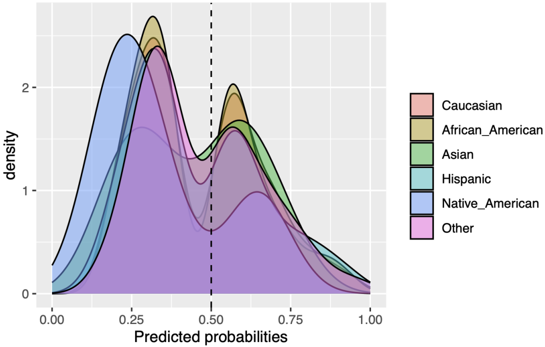
- h is nonparametric
- Use a random forest model for h, built by bootstrap aggregating large number of decision trees.
- Each decision tree is constructed by adaptively partitioning the predictor space formed by $\widehat{f}(x)$ and $\widehat{p}(x)$, which, implies a partition of the underlying input space x.
- Allows us to adaptively assign different weights to different regions of the input space depending on which of the statistical or the structural performs better.

Methodology: Simple Illustration



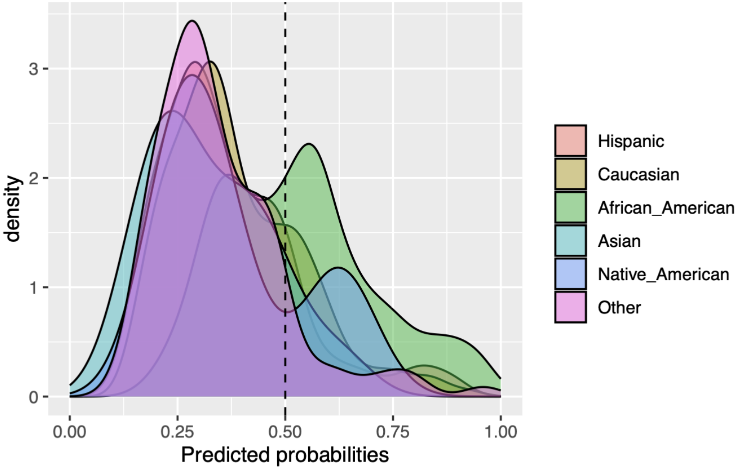
Sample Outcome: Representation Learning

A sample visualization of model performance would be like this:



The graph is made with R package "fairness". It shows the fairness score, measured in predicted possibility of different groups, of the revised model with PCA embedding all variables including sensitive ones. We can see that in this way, the distribution of risk across groups are modelled in quite similar shape. But this results in

Sample Outcome: Original COMPAS predictor



Thanks for listening

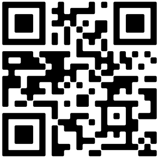


Figure: GitHub Repository