A Mathematical Approach to Algorithmic Fairness in Machine Learning

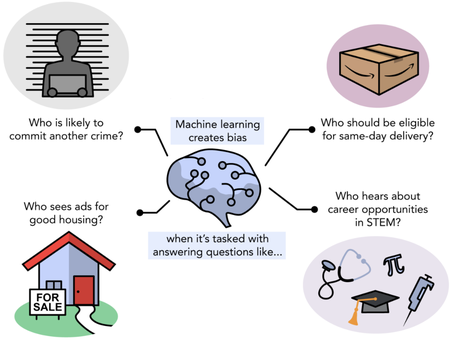
Aditya Mittal

Department of Mathematics; University of California, Davis

Introduction

Machine learning and AI systems are increasingly embedded in real-world decision-making processes that impact consumers. This brings forth a critical concern: *algorithmic fairness*. The challenge here is ensuring that predictions are not disproportionately influenced by sensitive attributes such as race, gender, or age with the goal of promoting equitable outcomes.

Examples of Fairness: Hiring decision, Healthcare, Credit Lending, Risk Assessments.



It is necessary to develop mathematical approaches to quantify biases and implementing strategies to mitigate potential unfairness.

Motivating Example: COMPAS

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm was a commercial machine learning system designed to assess a criminal's risk of recidivism and assist judges in sentencing decisions within the U.S. legal system.

An independent analysis by *ProPublica* alleged significant disparities: Black defendants were disproportionately misclassified as high-risk, while white defendants were more frequently misidentified as low-risk.

A Look into Algorithmic Fairness: Metrics and Implementation

The scope of Algorithmic Fairness includes establishing benchmarks to measure fairness and developing methods in compliance with these metrics.

- **Fairness Metrics:** Fairness criteria can be categorized into two measures: *individual fairness* and *group fairness*.
  - Individual fairness is that similar individuals should be treated similarly.
  - Group fairness requires that predictions remain consistent across different groups as defined by some sensitive attribute(s).
- **Implementing Fairness:** Fairness constraints can be integrated across various stages of the machine learning pipeline.
  - Pre-processing involves removing bias from the original dataset before training the model
  - In-processing refers to incorporating fairness constraints during model training to reduce the predictive power of sensitive variables. A considerable amount of research in this area involves achieving fairness through ridge penalties in linear models
  - Post-processing involves modifying model predictions after training.

Fairness Metrics

Individual Fairness ensures similar individuals should receive similar predictions. If two individuals are close in terms of relevant features, their outcomes should also be close:

$$d(f(x_i), f(x_j)) \leq d(x_i, x_j)$$

Group fairness ensures that predictive decisions do not disproportionately favor or disadvantage certain demographic groups. Several group fairness metrics have been proposed:

1. **Demographic Parity** requires the decision outcome to be independent of the protected attribute (e.g., race, gender):

$$P(\hat{Y} = 1 \mid S = a) = P(\hat{Y} = 1 \mid S = b), \quad \forall a, b \in \mathcal{S}$$

2. **Equal Opportunity** ensures that the True Positive Rate (TPR) is equal across different demographic groups for individuals who should receive a positive outcome.

$$P(\hat{Y} = 1 \mid Y = 1, S = a) = P(\hat{Y} = 1 \mid Y = 1, S = b), \quad \forall a, b \in \mathcal{S}$$

3. **Equalized Odds** requires that both the True Positive Rate (TPR) and the False Positive Rate (FPR) are the same across demographic groups.

$$P(\hat{Y} = 1 \mid Y = y, S = a) = P(\hat{Y} = 1 \mid Y = y, S = b), \quad \forall y \in \{0, 1\}, \quad \forall a, b \in \mathcal{S}$$

4. **Correlation Based Metrics:** requires that correlation between $\hat{Y}$ and S is minimized across all demographic groups.

$$\min_{\hat{Y}} |\text{cor}(\hat{Y}, S)|$$

Implementing Fairness via Ridge Penalty on S

The setting: Let X and S denote matrices of predictors and sensitive features. Our goal is to predicting $\hat{Y}$ while minimize the predictive power of S in our regression model.

First, we can write:

$$X = B^T S + U$$

where B is the solution to the least squares problem: $B_{\text{OLS}} = (S^T S)^{-1} S^T X$. We can define the residuals:

$$\hat{U} = X - B_{\text{OLS}}^T S$$

By properties of OLS, residuals and regressors are uncorrelated. Thus, S and $\hat{U}$ are orthogonal; i.e. COV(S, $\hat{U}$) = 0.

B_{OLS}^T can then be interpreted as the component of X that is explained by S, and $\hat{U}$ as the component of X that cannot be explained by S (the de-correlated predictors).

Formulating the Mathematical Problem

We may now define our model as follows:

$$y = \alpha^T S + \beta^T \hat{U} + \epsilon$$

We aim to predict $\hat{Y}$ while minimizing the predictive power of α coefficients.

We can formulate this problem as:

$$\min_{\alpha,\beta} E\left((y - y_b)^2\right) \quad \text{s.t.} \quad \|\alpha\|_2^2 \leq t(r) \quad \text{where} \quad t(r) > 0$$

or equivalently,

$$(\alpha_b^{\text{FRRM}}, \beta_b^{\text{FRRM}}) = \arg\min_{\alpha,\beta} \|y - S\alpha - Ub\beta\|_2^2 + \lambda(r)\|\alpha\|_2^2$$

To solve this problem:

$$\begin{bmatrix} \widehat{\alpha}_{\text{FRRM}} \\ \widehat{\beta}_{\text{FRRM}} \end{bmatrix} = \left(\begin{bmatrix} \mathbf{S}^\top \\ \widehat{\mathbf{U}}^\top \end{bmatrix} \begin{bmatrix} \mathbf{S} & \widehat{\mathbf{U}} \end{bmatrix} + \begin{bmatrix} \lambda(r)\mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{S}^\top \\ \widehat{\mathbf{U}}^\top \end{bmatrix} \mathbf{y}$$

$$= \begin{bmatrix} \mathbf{S}^\top \mathbf{S} + \lambda(r)\mathbf{I} & 0 \\ 0 & \widehat{\mathbf{U}}^\top \widehat{\mathbf{U}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}^\top \\ \widehat{\mathbf{U}}^\top \end{bmatrix} \mathbf{y}$$

$$= \begin{bmatrix} (\mathbf{S}^\top \mathbf{S} + \lambda(r)\mathbf{I})^{-1} \mathbf{S}^\top \mathbf{y} \\ (\widehat{\mathbf{U}}^\top \widehat{\mathbf{U}})^{-1} \widehat{\mathbf{U}}^\top \mathbf{y} \end{bmatrix}.$$

The β_{FRRM} can be estimated in closed form, only depending on $\hat{U}$, and do not change as r varies. The α_{FRRM} depend on S and also on r through $\lambda(r)$, and they must be estimated numerically.

Introducing TowerDebias

My current research with Dr. Norm Matloff in the Department of Computer Science.

$$E(Y|X) = E[E(Y|X, S)|X]$$

towerDebias estimates $E(Y|X)$ by modifying the predictions of an algorithm designed to predict $E(Y|X, S)$. The Tower Property in probability theory is key here: averaging $E(Y|X, S)$ over S while fixing X gives us $E(Y|X)$. Since the latter does not depend on S, we have effectively removed the influence of S.

References

- Dwork, Cynthia, et al. "Fairness Through Awareness."
- Hardt, Moritz, et al. "Equality of Opportunity in Supervised Learning."
- Scutari, Marco. "fairml: A Statistician's Take on Fair Machine Learning Modelling."
- Scutari, M., Panero, F., & Proissl, M. (2020). "Achieving fairness with a simple ridge penalty."

Mentor: Jonathan Forstater

Math Directed Reading Program

adimittal@ucdavis.edu