Non-Discriminatory Machine Learning

Naman Goel, Mohammad Yaghini, Boi Faltings

Artificial Intelligence Lab, EPFL, Switzerland

naman.goel@epfl.ch

1. Introduction

Algorithmic decision-making is being used in areas such as calculating criminal recidivism risk scores, stop-and-frisk programs, predictive policing, university admissions, bank loan approvals and jobs/salary screening/recommendation etc. There have been strong evidences that such decisions often show discrimination based on sensitive attributes such as race, gender etc (even while not directly using the sensitive attributes in decision making). Examples include disfavoring a certain race while calculating recidivism risk score or a certain gender while recommending jobs and salaries. This causes machine learning systems to not only appear discriminatory (which may have legal/financial consequences) but potentially create (or increase) imbalance in the society.

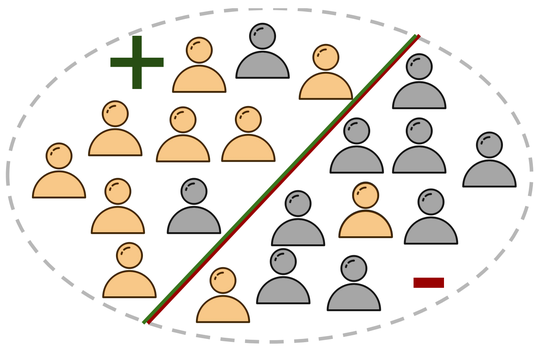


Figure 1: *A discriminatory decision boundary favoring people of different races differently. The challenge is to learn a non-discriminatory decision boundary with minimum loss in utility.*

- Current approaches for learning non-discriminatory models either result in non-convex optimization problems or don't have a clear probabilistic interpretation. We propose a technique to achieve non-discrimination in machine learning without sacrificing probabilistic interpretation and convexity.
- Besides non-discrimination, **fairness** of machine learning systems is another (more subjective) issue. Unlike earlier approaches, our technique also offers a *weighted proportional fairness* interpretation of its decisions.

2. What is non-discrimination ?

For binary classification and binary sensitive attribute :

- **Demographic Parity :**
  Predicted class $(\hat{Y})$ is independent of the sensitive attribute (Z).

$$P(\hat{Y}|Z=b) = P(\hat{Y}|Z=w)$$

- **Equalized Odds :**
  Predicted class $(\hat{Y})$ is independent of the sensitive attribute (Z), conditional on the true class (Y).

$$P(\hat{Y}|Z=b,Y) = P(\hat{Y}|Z=w,Y)$$

Depending on the domain and application, there may be several other definitions of non-discrimination such as equal true positive or false negative rates for all values of the sensitive attribute.

3. Which definition is better ?

Depends on the bias is the world, in which the classifier operates. World bias represents the inherent correlation between the true class and the sensitive attribute in real world.

Observations :

- If there is no bias in the world, any classifier satisfying equalized odds also satisfies demographic parity.
- If the world is biased, the only way to make a classifier satisfying equalized odds also satisfy demographic parity is by making its predictions independent of the truth (which makes the classifier practically useless).

4. Approximately non-discriminatory classifiers (p-rule)

$$min(\frac{P(\hat{Y}=+|Z=b)}{P(\hat{Y}=+|Z=w)}, \frac{P(\hat{Y}=+|Z=w)}{P(\hat{Y}=+|Z=b)}) \geq p$$

Observations :

- If the world satisfies p-rule, any classifier satisfying equalized odds also satisfies p-rule.
- If the world doesn't satisfy p-rule, a classifier satisfying equalized odds must lose some accuracy to satisfy p-rule. The required loss in accuracy depends on how far the world is from satisfying the p-rule.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

5. Weighted Sum of Logs Technique

Maximize weighted sum of logs of probabilities of favoring individuals subject to constraints on accuracy of a logistic regression type classifier. Weights represent historic bias in the training data. Individuals belonging to historically disfavored and minority groups are given more weight than the others. Results in a convex optimization problem in the model parameters (minimizing negative of weighted sum) and preserves probabilistic interpretation.

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \sum_{m=1}^{N} w_{g(m)} \cdot \log \hat{P}_m^+(\theta) \\ \text{subject to} \quad & \mathcal{L}(\theta) \leq (1+\delta)\mathcal{L}(\theta^v) \end{aligned}$$

- $w_{g(m)}$: empirical estimate of bias in the training data against m's group (race/gender etc).
- $\hat{P}_m^+$: probability of the classifier favoring m.
- θ : model parameters vector of the non-discriminatory classifier to be learned.
- θ^v : model parameters vector of vanilla classifier.
- $\mathcal{L}(\theta)$: classifier's loss (negative log likelihood).
- δ : threshold of tolerance for increase in classifier's loss.

6. Fairness Interpretation

Proportional fairness of a classifier (motivated by rate control in networks [2]) requires that the aggregate of proportional changes in favor given to individuals by any other *allowed* classifier, as compared to a proportionally fair classifier, is negative. Weighted proportional fairness weighs the individual terms in the aggregate according to the different costs paid by the corresponding individuals in history. The weighted sum of logs technique satisfies the notion of weighted proportional fairness, with allowed classifiers being within the threshold of loss tolerance. The fairness notion satisfies interesting properties such as Pareto optimality.

7. Datasets

- **ProPublica's COMPAS Dataset** : Whether individual recidivated within 2 years or not. Not recidivating is '+ve' class.

Race	+ve	-ve
White	**61%**	39%
Black	**49%**	51%

Bias in Training Data

Race	+ve	-ve
White	**74%**	26%
Black	**48%**	52%

Bias in Logistic Regression Classifier

- **Adult Dataset** : Whether individual has high income ($\geq$ 50K USD) or not. High income is '+ve' class.

Gender	+ve	-ve
Male	**31%**	69%
Female	**12%**	88%

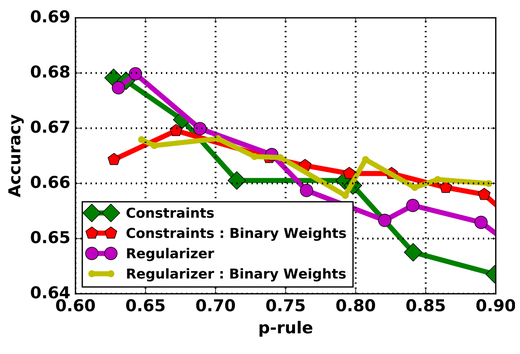
Bias in Training Data

Gender	+ve	-ve
Male	**24%**	76%
Female	**8%**	92%

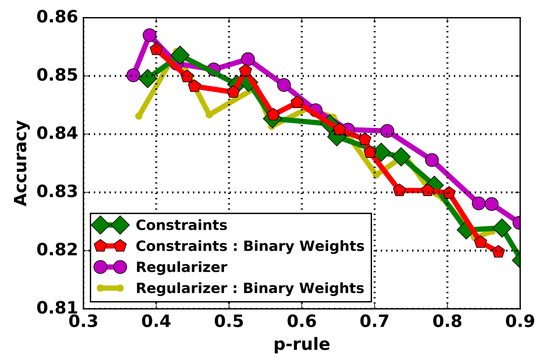
Bias in Logistic Regression Classifier

8. Results

- Comparison of our different versions of weighted sum of logs technique [1]



COMPAS Dataset



Adult Dataset

- Our method achieves state of the art accuracy-(non)discrimination trade-off without sacrificing probabilistic interpretation, convexity and while providing proportional fairness guarantee.

9. Conclusions

Possible to avoid (or reduce) gender, race, religion based unintentional discrimination in machine learning without sacrificing interpretability and computational efficiency. The accuracy loss suffered by non-discriminatory classifier depends on how biased the real world is.

References

[1] Goel, Naman, et al. "Non-Discriminatory Machine Learning through Convex Fairness Criteria." AAAI Conference on Artificial Intelligence (AAAI). 2018.

[2] Kelly, F. 1997. "Charging and rate control for elastic traffic." Transactions on Emerging Telecommunications Technologies 8(1):33–37.