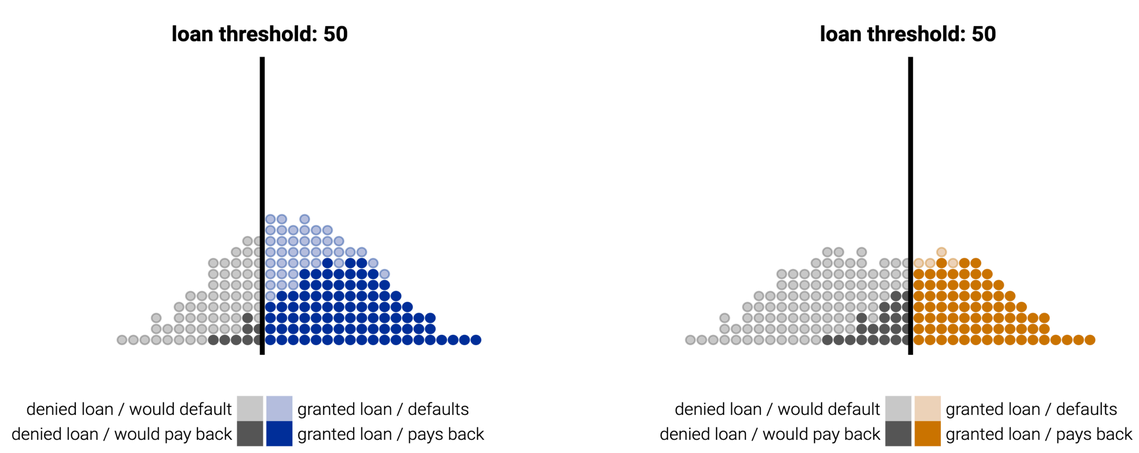
Optimization-Based Approaches for Enforcing Fairness in Machine Learning

Amil Merchant and Alexander Lin
Applied Math 221: Advanced Optimization
Spring 2019, Professor Yaron Singer

Introduction

As machine learning is increasingly adopted for real-world decision making with important societal consequences, it is important to ask if these models are fair and to ensure that they do not discriminate. Consider the following example from credit-loan classification which clearly shows that the blue group is disproportionately able to receive loans.



Dataset

The dataset used for this project was the UCI Adult Income dataset. Providing features such as education, marital status, and type of employment, the goal is to predict whether a person makes over $50,000 a year. The repository contains 32,561 individuals. Categorical variables are 1-hot encoded, leading to a total of 93 attributes in the feature matrix.
Although this dataset itself is not used to make societal decisions, it bears many similarities to credit scoring algorithms and showcases why fairness can be important.

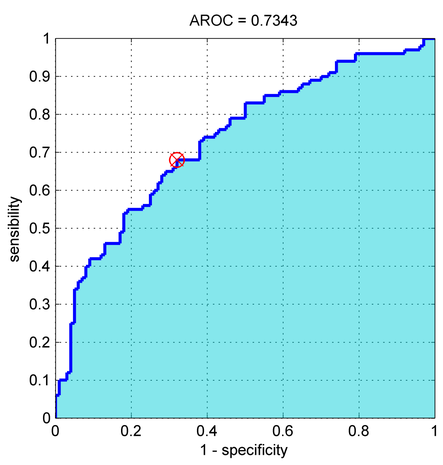
- **X: 14 features (ex: education, employment)**
- **Y: Income >$50,000**
- **Z: Protected attributes of race, sex**

Definitions of Fairness

In the past few years, a number of definitions of fairness have been proposed. In this project, we evaluated models based on the following:

Calibration (AUC-ROC)

This definition aims at maximizing predictive power and the accuracy of the classifier.

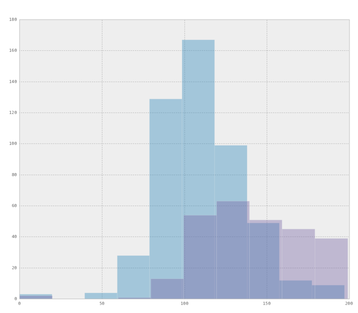


Source: molevole.altervista.org

Demographic Parity

This method aims at satisfying equal prediction distributions (referred to as the q %-rule) for protected groups.

$$\min\left\{\frac{\hat{p}(y=1|z=k_1)}{\hat{p}(y=1|z=k_2)}, \frac{\hat{p}(y=1|z=k_2)}{\hat{p}(y=1|z=k_1)}\right\} \geq \frac{q}{100}$$

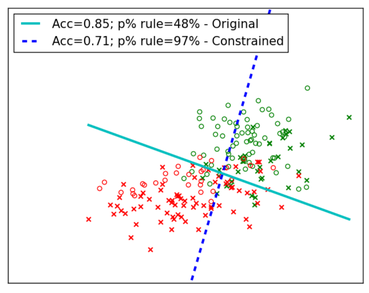


Source: machinelearningmastery.com

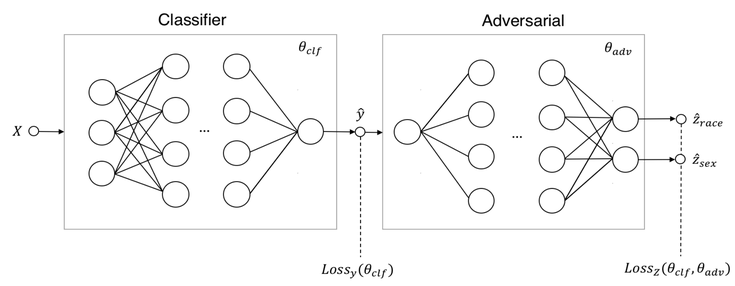
Methods

We explore two methods and apply them to neural network classifiers, which have proven to be successful in achieving high accuracy in many domains, yet are not traditionally constrained to satisfying demographic parity.

Regularizing Decision Boundary Covariance

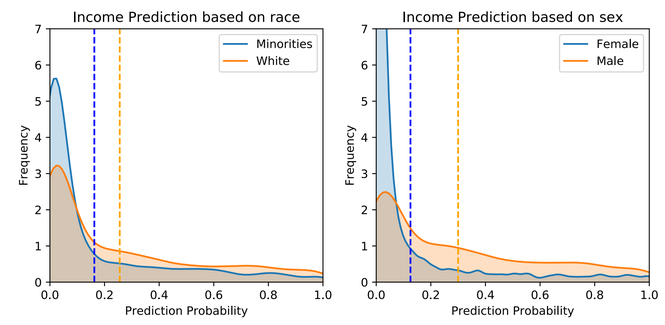


Optimization through Adversarial Networks

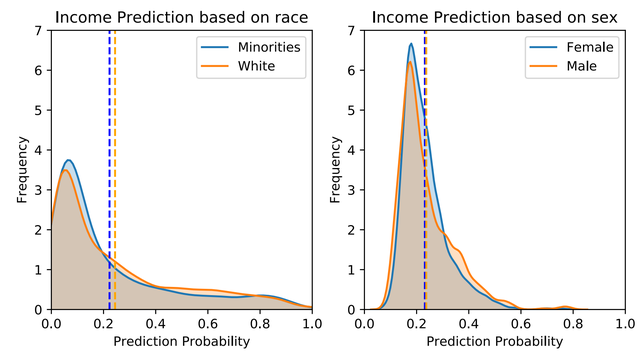


Results

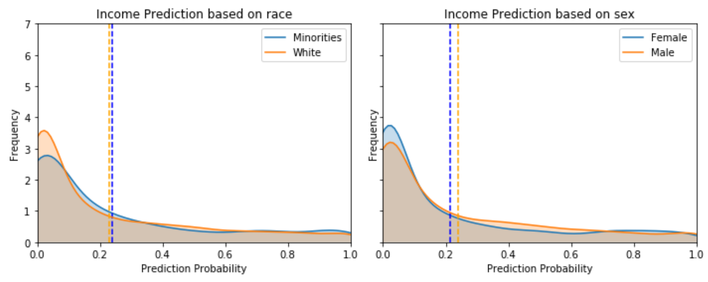
Original (Vanilla Neural Network)



Regularizing Decision Boundary Covariance



Optimization through Adversarial Networks



Conclusions

By slightly sacrificing predictive power, the two correction methods proposed above made significant progress in ensuring demographic parity. These methods are interesting approaches to removing discrimination from machine learning models. Future work aims to compare these methods as well as explore how these models perform on larger real-world datasets.

References

[1] Wadsworth, Christina, Francesca Vera, and Chris Piech. Achieving Fairness through Adversarial Learning. In FAT ML, 2018.

[2] Tonk, Stijn. Fairness in Machine Learning. https://github.com/equialgo/fairness-in-ml.

[3] Zafar, Muhammad et al. Fairness Constraints: Mechanisms for Fair Classification. In AISTATS, 2017.

Thank you to Yaron Singer for his support during this project.