Fairness in Machine Learning Using a Logistic Decision Rule

Eric Landgrebe, Vineet Parikh, Christopher Qian

September 2019

1 Introduction

Fairness in machine learning is a topic that has seen an explosion in interest as machine learning becomes more widespread. The use of machine learning in important decisions, such as the COMPRAS model for determining risk of criminal recidivism, has come under scrutiny lately because of the discrepancy in predictions across groups. As it turns out, some definitions of fairness that would be simultaneously desirable are fundamentally incompatible with each other. Our project will be focused on determining how well we can satisfy different fairness constraints, given a certain model.

2 Model

In this project, we will assume a setting where people are associated a set of features, drawn from a probability distribution, where different groups with have different probability parameters. We assume a logistic decision rule determines a particular binary label of a person. That is,

$$P(Y = y|X = x) = \frac{1}{1 + \exp(-y(w^T x + b))}$$

For example, Y could represent whether or not a person will default on their credit card. If a company wants to determine whether or not to approve someone's credit card, setting $\hat{Y} = Y$ gives the Bayes optimal predictor. Of course, in practice, a company will not know the optimal predictor. However, we will assume that they use a logistic decision rule of their own on the same set of features, just with a different weight vector w (indicating a discrepancy in what they believe to be a valuable predictor and what actually is a good predictor).

3 Questions

We will attempt to answer the following questions:

- Is the Bayes optimal predictor necessarily fair?
- How do perturbations in model parameters affect accuracy and fairness?
- How can we optimize accuracy with different fairness metrics/how do classifiers optimized for different fairness metrics compare in terms of accuracy?

In addition, we will test our fairness optimization with a logistic classifier on multiple datasets, including generated multinomial distributions, a (fairly large) Kaggle dataset on predicting whether loans would default provided by ICL, and a (fairly small) German creditability dataset that classifies whether individual recipients would be risky or safe to loan to. We are looking to compare accuracy across optimizing our classifier for these different fairness metrics.

4 Project Value

We believe that our model will present a significant contribution to the fairness literature. Our model is a reasonable approximation for actual decision rules in that it captures how the agents that make a decision would evaluate a candidate. By making this assumption, our theoretical and simulated results will provide valuable insight for how to balance fairness in real world classifiers.

5 References

1. German Creditability Dataset: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
2. Loan Default Prediction Dataset: https://www.kaggle.com/c/loan-default-prediction?fbclid=IwAR2IM0K06xOLfY-
h-PaW9EQAVQ0Kh34NM45L9_2wpYwjbPn5FV8Qb6_q1x4