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 COMPAS model for determining risk of criminal recidivism, has come under scrutiny lately because of the discrepancy in predictions across groups. Most fairness metrics primarily focus on how they could be useful as a conceptual form of defining how abstractly "fair" an individual classification or regression system would appear, based on the predictions a system would make, but it's impossible to fully optimize for and satisfy every definition of fairness. Kleinberg et al. prove that specific definitions of fairness/fairness metrics are incompatible with each other (specifically equal odds and calibration), putting into frame the need to understand how fairness metrics can interplay and clash with each other [1]. Heidari et al. have focused on understanding and surveying the wide number of fairness metrics out there, and conclude that for fairness metrics reflecting irreconciliable moral assumptions, it's naturally impossible to ensure that a system could be optimized to guarantee both [4]. Gajane takes this problem a step further, analyzing and showing that individual fairness metrics have their own flaws [2]. Even simply attempting to optimize fairness metrics across individual groups of the data, without considering how fair a classifier would be when applied on intersections of these groups, could lead to issues such as "fairness gerrymandering" as Kearns et al show [3]. While having many fairness metrics can be necessary and can provide nuanced information on how fair a system can be, it's also necessary to understand how individual metrics of fairness both theoretically and practically interact with each other, such that individuals who wish to optimize systems for one definition of fairness understand the inherent tradeoffs of this optimization. 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, especially with regards to perturbing model parameters for one fairness metric and seeing the effect on a different fairness metric?
- 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 distributions with features drawn from different distributions like mulitnomial and Guassian, 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, and while existing testing frameworks such as AEQUITAS, by Udeshi et al, can be leveraged to optimize for arbitrary fairness metrics, we plan to consider significantly simpler (and more easily optimizable) fairness metrics within our experiments [5].

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, and for understanding the inherent trade-offs that different fairness metrics may have.

5 Datasets

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

References

[1] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *CoRR* abs/1609.05807 (2016). arXiv: 1609.05807. URL: http://arxiv.org/abs/1609.05807.

[2] Pratik Gajane. "On formalizing fairness in prediction with machine learning". In: *CoRR* abs/1710.03184 (2017). arXiv: 1710.03184. URL: http://arxiv.org/abs/1710.03184.

[3] Michael J. Kearns et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: *CoRR* abs/1711.05144 (2017). arXiv: 1711.05144. URL: http://arxiv.org/abs/1711.05144.

[4] Hoda Heidari et al. “A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity”. In: *CoRR* abs/1809.03400 (2018). arXiv: 1809.03400. URL: http://arxiv.org/abs/1809.03400.

[5] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. “Automated Directed Fairness Testing”. In: *CoRR* abs/1807.00468 (2018). arXiv: 1807.00468. URL: http://arxiv.org/abs/1807.00468.