Fairness in Machine Learning: Comprehensive Analysis & Improvements

Esha Singh Saurabh Mylavaram Anushree Choudhary
sing0640@umn.edu mylav008@umn.edu choud146@umn.edu

1 Abstract

Research in algorithmic fairness has increased substantially over the years giving rise to new metrics and algorithms to address this issue. In this project, we will try to explore the fairness footprints of popular classification algorithms that we will see in the course. We will tackle the issue of quantifying fairness and also suggest remedial methods to improve fairness of these algorithms.

2 Introduction

In recent years, there has been a surge in the usage of Machine Learning algorithms to simplify our day-to-day decisions in diverse fields. From medical diagnosis of deadly diseases to informing decisions about loan sanctions or bail grants, it can influence the lives of individuals in a profound way. Given this scenario, there is a major concern surrounding algorithms that can not only amplify the inherent human biases which are present in data [1], but also introduce new ones [2]. A fair algorithm can be understood as one that does not privilege or disadvantage an arbitrary group of users based on sensitive information like gender, ethnicity, disability or sexual orientation. Fairness of a decision is evaluated using two distinct notions: disparate treatment[3] and disparate impact[3].

3 Evaluation & Analysis

We aim to explore and evaluate algorithmic fairness of the below stated classification algorithms. For this we will use UCI datasets, and evaluate them using fairness metrics described below.

- **Dataset:** We will make use of the UCI datasets repository from which we have isolated two datasets namely : Breast Cancer Coimbra Dataset[4] and COMPAS Recidivism Dataset[5]
- **Models:** Given our choice of datasets and the target variables of interest, we will limit ourselves to supervised classification algorithms. Specifically we will deal with Support Vector Machines (SVMs) and Logistic Regression.
- **Metrics:** There are numerous ways to quantify fairness. Superficially we can categorise them into 1) statistical measures, 2) definitions based on predicted outcome, 3) predicted and actual outcomes, 4) predicted probabilities and the actual outcome. Based on statistical measures as per [6][3] we can define 8 different metrics amongst which we will explore FPR in our future work. We will also explore demographic parity (category 2), Equality of Opportunity (category 3) and test-fairness (category 4). We would also use Individual fairness to counter drawbacks of group fairness.

4 Proposal to Improve Fairness

We propose to improve fairness in supervised classification tasks in the following ways:

1. Devise ways to re-design/augment algorithms (SVM, logistic) to make them more fair[7]
2. Develop new metrics which quantify fairness with less bias than prior suggested metrics.

5 Alternative project idea

Alternatively, we shall build a recommendation system for Anime series for which data is available on Kaggle[8].

References

[1] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, pp. 4349–4357, 2016.

[2] L. Sweeney, "Discrimination in online ad delivery," *Queue*, vol. 11, no. 3, pp. 10–29, 2013.

[3] Z. Zhong, "A tutorial on fairness in machine learning."

[4] U. of Irvine, "University of irvine machine learning repository."

[5] ProPublica, "Compas recidivism risk score data and analysis," Mar 2019.

[6] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, FairWare '18, (New York, NY, USA), p. 1–7, Association for Computing Machinery, 2018.

[7] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, "Putting fairness principles into practice: Challenges, metrics, and improvements," 2019.

[8] CooperUnion, "Anime recommendations database," Dec 2016.