Fairness Checking

1 Introduction

[Deb: Rough sketch of the introduction:

- Usual Motivation for fairness
- Why do we care about distributional robustness in this setting?
- Connection with verification literature
- Related work – (a) fairness in classification, (b) distributionally robust optimization,
- Contributions – (a) distributional robustness for fairness, (b) general framework to design such a classifier, (c) faster approximate minmax solution, (d) simulation results

] Nowadays, AI systems are increasingly used in various high-stakes decision making scenarios. Applications include bail decision, credit approval, and housing allocation, to name a few. Often these applications use learning algorithms trained on past biased data, and such bias is often reflected in the eventual decision. For example, [3] show that popular word embeddings implicitly encode societal biases, such as gender norms. Similarly, [4] evaluate existing facial recognition systems and find that they perform better on lighter-skinned subjects as a whole than on darker-skinned subjects as a whole with an 11.8% - 19.2% difference in error rates. To mitigate these biases, there have been several approaches in the ML fairness community to design fair classifiers [7, 6, 2].

However, the literature has largely ignored the robustness of such fair classifiers. As an example, we consider the performance of the optimized pre-processing algorithm [5] on the popular COMPAS dataset [1]. As a metric of fairness we consider the notion of *demographic parity* (DP), which measures the difference in accuracy between two protected groups. Figure shows two situations – (1) unweighted training distribution (in blue), and (2) weighted training distributions. The optimized pre-processing algorithm [5] is almost fair on the unweighted training dataset (DP ≤ 0.02). However, it shows demographic parity of at least 0.1 on the weighted dataset, despite the fact that the marginal distributions of the features look almost the same for the two scenarios. This example motivates our work and we aim to design a fair classifier that is robust to such perturbations. We also show how to construct such reweighted examples.

Nonetheless, since different algorithms adopt different definitions of fairness and provide different trade-offs with respect to accuracy and utility, it is neither legal nor ethical to enforce businesses to use such algorithms. In this paper, we approach this problem with a perspective from the literature of automated verification, and aim to build tools that can verify whether an algorithm satisfies a given fairness criteria irrespective of the particular algorithm or dataset used. We show using these tools that, although current group fairness algorithms may mitigate fairness for a specific distribution of data, slight perturbations to that data's distribution result in violations of the fairness criteria.

Preprint. Under review.

2 Problem and Definitions

3 Meta Algorithm

4 Approximate Fair Classifier

5 Faster Approximate Fair Classifier

6 Experiment

7 Conclusion

References

[1] Compas dataset. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis. Accessed: 2019-10-26.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[4] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[5] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*, pages 3992–4001. 2017.

[6] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of Opportunity in Supervised Learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[7] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *International Conference on Machine Learning*, pages 325–333, 2013.