Fairness Checking

October 20, 2019

1 Introduction

Nowadays, AI systems are increasingly being used for making various high-stakes decision making. Applications include bail decision, credit approval, housing allocation etc. These applications use learning algorithms and if such algorithms are trained on past data which is almost always biased, such bias is often reflected in the eventual decision. For example, Bolukbasi et al. [Bol+16] show that popular word embeddings implicitly encodes societal bias. Similarly, Buolamwini and Gebru [BG18] evaluate existing facial recognition systems and find that they perform better on lighter subjects as a whole than on darker subjects as a whole with an 11.8% - 19.2% difference in error rates. There have been several approaches to design fair classifiers [Zem+13; HPS+16; Aga+18]. Since different algorithms adopt different definitions of fairness and provide different trade-offs with respect to accuracy / utility, it is neither legal nor ethical to enforce any business 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 satifies a given fairness criteria irrespective of the particular algorithm / dataset used.

2 Model

We first check whether an algorithm is fair against a family of possible distributions. In particular, we consider distributions which are weighted empirical distributions and weightes are chosen so that the new weighted distribution is close to the original training distribution.

2.1 Setup

Suppose, we have access to a "protected" attribute $A \in \{0, 1\}$, and a qualification attribute $Y \in \{0, 1\}$. Let us assume that $X \in \mathcal{X}$ denote the set of remaining attributes which are used as input to a classifier f. For a distirbution P over the space of attributes $\mathcal{X}$, we consider the following two fairness criteria:

- Demographic Parity (**DP**):

$$\left|\mathbb{E}_P[f(X, a)|A = a] - \mathbb{E}_P[f(X, a')|A = a']\right| \leq \epsilon$$

  for all a and a'.

- Equalized Odds (**EO**):

$$\left|\mathbb{E}_P[f(X, a)|Y = y, A = a] - \mathbb{E}_P[f(X, a')|Y = y, A = a']\right| \leq \epsilon$$

  for all y, a, and a'.

We assume that we have data (X_i, Y_i, A_i) for $i = 1, \ldots, n$ and P can be represented as a weighted empirical distribution i.e. for any $(x, y, a) \in \mathcal{X} \times \{0, 1\} \times \{0, 1\}$ we have

$$P(x, y, a) = \sum_{i=1}^{n} w_i \mathbf{1}_{(X_i, Y_i, A_i) = (x, y, a)}.$$

2.2 Checking Demographic Parity (DP)

As a start, we assume that the weighted empirical distributions are such that the marginal distributions over the protected attributes are preserved. In particular, we consider weights such that $\sum_{i=1}^{n} w_i \mathbf{1}_{A_i=a} = \pi_a$ for all a. Here the π_a are the proportions for the protected attributes, which we assume are known. We use the following linear programs to check if the classifier f fails on some allowable weighted empirical distribution. For each a, a':

$$\begin{aligned} \max_{w} \quad & \frac{1}{\pi_a} \sum_{i=1}^{n} w_i f(X_i) \mathbf{1}_{A_i=a} - \frac{1}{\pi_{a'}} \sum_{i=1}^{n} w_i f(X_i) \mathbf{1}_{A_i=a'} \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i \mathbf{1}_{A_i=a} = \pi_a \\ & \sum_{i=1}^{n} w_i \mathbf{1}_{A_i=a'} = \pi_{a'} \\ & w_i \geq 0 \quad \forall i \in [n] \\ & \sum_{i=1}^{n} w_i = 1 \end{aligned} \tag{1}$$

If there exists a pair of protected attributes a, a' such that the optimal value of the linear program is more than ϵ, then we have found a violation of DP.

2.3 Checking Equalized Odds (EO)

As in the previous section, we assume that we only care about weighted empirical distributions such that $\sum_{i=1}^{n} w_i \mathbf{1}_{A_i=a, Y_i=y} = \pi_{a,y}$ for all a and y where $\pi_{a,y}$ are known proportions for the protected and qualification attributes. We again use the following linear programs to check if f fails on some weighted empirical distribution.

$$\begin{aligned} \max_{w} \quad & \frac{1}{\pi_{a,y}} \sum_{i=1}^{n} w_i f(X_i) \mathbf{1}_{A_i=a, Y_i=y} - \frac{1}{\pi_{a',y}} \sum_{i=1}^{n} w_i f(X_i) \mathbf{1}_{A_i=a', Y_i=y} \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i \mathbf{1}_{A_i=a, Y_i=y} = \pi_{a,y} \\ & \sum_{i=1}^{n} w_i \mathbf{1}_{A_i=a', Y_i=y} = \pi_{a',y} \\ & w_i \geq 0 \quad \forall i \in [n] \\ & \sum_{i=1}^{n} w_i = 1 \end{aligned} \tag{2}$$

3 Result

4 Conclusion

References

[Aga+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A Reductions Approach to Fair Classification". In: *arXiv preprint arXiv:1803.02453* (2018) (cit. on p. 1).

[BG18] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Conference on fairness, accountability and transparency*. 2018, pp. 77–91 (cit. on p. 1).

[Bol+16] 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*. 2016, pp. 4349–4357 (cit. on p. 1).

[HPS+16] Moritz Hardt, Eric Price, Nati Srebro, et al. "Equality of Opportunity in Supervised Learning". In: *Advances in neural information processing systems*. 2016, pp. 3315–3323 (cit. on p. 1).

[Zem+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. "Learning Fair Representations". In: *International Conference on Machine Learning*. 2013, pp. 325–333 (cit. on p. 1).