A pragmatic take on fair machine learning

Tatiana Lozano Ortega[1]
Alfredo Lozano Ortega[2]

Abstract

Machine Learning is becoming more and more accessible for developers to implement in automatic decision making which may involve tasks that can lead to systematic discrimination. Several studies have revealed the ease in which machine learning algorithms can learn to replicate biases from human values when trained on data that contains signal about such biases for a specific task (Boulbaski et al. 2016, Larson et al. 2016). In this work we will divulge specific algorithms and settings for algorithmic fairness while emphasizing on the limitations of the approach taken in the state of the art. We will also contribute with an ethical overview of the concept offered for fairness in the field. Then we will identify key points for future work on fair AI.

References

Bellamy, Rachel K.E., et al. (2018). "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias." *arXiv preprint arXiv:1810.01943v1*

Bolukbasi, T., Chang, K.W., Zou, J. Y., Saligrama, V. & Kalai, A. T. (2016). "'Man is to computer programmer as woman is to homemaker? Debiasing word embeddings', Advances in Neural Information Processing Systems", pp. 4349–4357.

Calmon, Flávio du Pin, et al. (1970). "Optimized Pre-Processing for Discrimination Prevention - Semantic Scholar."

Hardt, Moritz, Eric Price, and Nati Srebro. (2016). "Equality of opportunity in supervised learning." *Advances in neural information processing systems*.

Hu, Lily, and Yiling Chen. (2018). "Welfare and distributional impacts of fair classification." *arXiv preprint arXiv:1807.01134*

Kamiran, Faisal, and Toon Calders. (2011). "Data Preprocessing Techniques for Classification without Discrimination." *SpringerLink*, Springer-Verlag.

Kamiran, Faisal, Asim Karim, and Xiangliang Zhang. (2012). "Decision theory for discrimination-aware classification." *2012 IEEE 12th International Conference on Data Mining*. IEEE.

Kamishima, Toshihiro, et al. (2012). "Fairness-aware classifier with prejudice remover regularizer." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). "How we analyzed the COMPAS recidivism algorithm".

Pleiss, Geoff, et al. (2017). "On fairness and calibration." *Advances in Neural Information Processing Systems*.

Zemel, Richard S., et al. (1970). "Learning Fair Representations - Semantic Scholar."

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. (2018). "Mitigating unwanted biases with adversarial learning." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

[1] tatiana@tueleccion.org, Universidad Panamericana
[2] Instituto Tecnológico Autónomo de México