Project: Verifying Bias of Machine Learning Algorithms using Adaptive Sampling and Online Learning

Philippe Preux and Debabrota Basu

philippe.preux@inria.fr and debabrota.basu@inria.fr

Where: Scool (previously SequeL), Inria Lille- Nord Europs, Villeneuve d'Ascq, France

Expected background: Master in CS, specialized in machine learning and mathematical statistics. This work will require the candidate to be comfortable with mathematical proofs.

Keywords: Online learning, Multi-armed bandits, Online batch classification, Regret bounds, Demographic fairness.

Objective: The problem of bias in the outcome of the machine learning algorithms is a concern of present times. Specially, the machine learning algorithms are growingly used for high-stake decision making, such as recidivism of accused persons, endowing an insurance, recruiting freelancers. Thus, it is imperative to assess the bias of the ML algorithms to certain demographics over the others, and also to mitigate it Chouldechova and Roth (2020); Mehrabi et al. (2019).

In this project, we extend the ongoing research on fairness verifiers Bastani et al. (2019); Ghosh et al. (2021, 2022). Fairness verifiers consider an ML algorithm as a black-box operating on a dataset $(X_1, \ldots, X_n) \subset \mathcal{X}$ obtained from a data-generating distribution $D(\mathcal{X})$, and producing a set of outputs $(Y_1, \ldots, Y_n) \subset \mathcal{Y}$.

The fairness verifiers operate on these input-output pairs and estimate the bias of the ML algorithm towards a specific part $u \subseteq \mathcal{X}$ with respect to all other $v \subseteq \mathcal{X}$. These subsets are often predefined depending on the sensitive features in input (e.g. gender, economic status etc.). But in applications like e-commerce, these sensitive attributes are often less clearly defined and identifying the u and v for which the unfairness occurs is an open problem. In this project, we aim to identify such subspaces of the input space for which the outcomes are discriminated the most. For this purpose, we aim to use adaptive sampling Bollapragada et al. (2018) and online learning methods (e.g. bandits Lattimore and Szepesvári (2020)) that can use the knowledge of the previously chosen samples to adaptively choose the next samples.

This will require developing theoretical and algorithmic tools as the present set of verifiers use the full input dataset and outputs to estimate the fairness of the ML algorithm, and also to specify the sensitive features beforehand.

The candidate is also expected to be comfortable with Python and coding. As the problem under investigation is practical, an open-source code demonstrating the usefulness of the developed algorithm is also desired.

This project is endorsed by the pilot project of Inria and French government, namely Regalia, to regulate the bias in the algorithms used for e-commerce and web platforms.

References

Bastani, O., Zhang, X., and Solar-Lezama, A. (2019). Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA):1–27.

Bollapragada, R., Byrd, R., and Nocedal, J. (2018). Adaptive sampling strategies for stochastic optimization. *SIAM Journal on Optimization*, 28(4):3312–3343.

Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89.

Ghosh, B., Basu, D., and Meel, K. S. (2021). Justicia: A stochastic sat approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7554–7563.

Ghosh, B., Basu, D., and Meel, K. S. (2022). Algorithmic fairness verification with graphical models.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.