Measuring and Eliminating Bias in Machine Learning Algorithms with Multi-armed Bandits

Debabrota Basu and Philippe Preux

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

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

Expected background: Master in CS, specialised in machine learning and mathematical statistics. This work will require the candidate to be comfortable with mathematical proofs. Knowledge in Python is a plus.

Keywords: Fair ML, Demographic fairness, Fairness verification, Fairness explanation, Online learning, Multi-armed bandits, Regret bounds.

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 increasingly 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).

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 $(Y_1, \ldots, Y_n) \subset \mathcal{Y}$. It also assumes that the input dataset is divided in two parts: sensitive features A and non-sensitive features Z, such that $\mathcal{X} = A \cup Z$. For example, in an algorithm designed to admit students in universities, gender, ethnicity, and economic status are sensitive attributes and other qualifications, like grades, are non-sensitive features. The bias of the algorithm is quantified as the difference in probability of classification for these groups or some function of it, e.g. Statistical Parity (Dwork et al., 2012, SP)

$$SP \triangleq |P[Y = 1|Z = z, A = \text{male}] - P[Y = 1|Z = z, A = \text{female}]|.$$

Fairness verifiers (Bastani et al., 2019; Ghosh et al., 2021, 2022) aim to measure this bias (e.g. SP) given a classifier M and data-generating distribution $D(\mathcal{X})$. In this project, we extend our ongoing research on fairness verifiers (Ghosh et al., 2021, 2022).

Existing fairness verifiers assume knowledge of the sensitive attributes and try to compute the bias due to them. But in applications like e-commerce, these sensitive attributes A are less clearly defined. It is rather more interesting in these problems to identifying the subset of features $A \subset \mathcal{X}$ for which the ML classifier gets more biased. Understanding this phenomena and designing efficient bandit algorithms to discover this subspace 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 multi-armed bandit algorithms (Lattimore and Szepesvári, 2020) that can use the knowledge of the previously chosen samples to adaptively choose the next samples. This shall also make the existing verifiers sample efficient as they do not perform adaptive sampling rather uses the whole dataset to measure the bias.

Progressing with this novel problem 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. This is also closely related to the recent progress on identifying the partitions using multi-armed bandits and also designing adaptive samplers using bandits (Juneja and Krishnasamy, 2019). As pointed by these works, developing such methods pose interesting statistical questions which leverage the structure of practical ML applications. We are anticipating to explore these research avenues.

If we are successful in designing such adaptive fairness verifiers using multi-armed bandits, the next goal will be to develop ML classifiers which can use this feedback to eliminate the bias. Realisation of this part of the project will be decided depending on the developments in the project.

Project Outcome: In this project, we follow the philosophy of open science. We aim to publish the research results generated in this process in open-access platforms (arXiv, hal) and as well as in premier machine learning (AAAI, AISTATS, ICML, IJCAI, NeurIPS, COLT) venues with open-access proceedings.

Along with the research papers, we expect that the different phases of the project will lead to open-source software. As the problem under investigation is practical, an open-source code demonstrating the usefulness of the developed algorithm is desired. All the softwares will be developed and available through the gitlab platform hosted by Inria.

Special Remark: 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. Thus, the candidate will collaborate actively with researchers and engineers connected to Regalia project.

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.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

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.

Juneja, S. and Krishnasamy, S. (2019). Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory*, pages 1824–1852. PMLR.

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*.