Exploring Alternative Definitions of Fairness in Machine Learning using Sensitive Networks

François Queyroi*, Hoel Le Capitaine†
LS2N – Nantes Université

Scientific Context

Many studies have shown that learning models can lead to inequality of treatment and unfair decisions [1]. A decision algorithm is said to be "unfair" if it's outcome depends (even indirectly) on some *protected* attribute (*e.g.* race, gender, *etc.*). Many efforts have been made to improve fairness in machine learning. However, there are many ways to define the dependency (or lack thereof) between the attributes and the decision making process, and therefore many ways to define fairness. One form of "fairness" may be in conflict with another form of fairness (*e.g.* individual and group fairness) or with the accuracy of the model. For these reasons, "it is nearly impossible to understand how one fairness solution would fare under a different definition of fairness"[1].

In much of the literature, however, the *protected* attributes are mostly discrete, encoding the fact that an individual belongs (or does not) belong to one or more groups. The definition of these groups may be based on legal definitions or simply rely on what is available in the datasets. A challenge in this context is to take into account the *intersectionality* of possible discriminations faced by individuals [2]. This leads to several problems *e.g.* underrepresentation and the need to adjust the definition of fairness [3]. But, even with *intersectionality* in mind, the sensitive attribute mainly relies on a set of discrete attributes.

Objectives

The aim of this project is to explore alternatives to the use of discrete variables to encode *sensitive attributes*. One possible way is to use a graph (the *sensitive network*) to encode proximity/relationship between individuals. In this context, *fairness* could be defined as the lack of correlation between the existence of relationships and the decision/score. An intuitive example of an "unfair decision" is hiring only people who know the same people in the network.

There is a "fair" amount of literature on fairness in machine learning on graphs [4]. In our case, however, the decision process could take any kind of data as input, but fairness would be access according to a *sensitive network*. Note that the latter can also encode simple discrete attributes. For example, isolated cliques of people could be used to encode the fact that the individuals belong to a single group.

*françois.queyroi@univ-nantes.fr
†hoel.lecapitaine@univ-nantes.fr

The objectives of this internship are to

1. Develop a state-of-the-art on alternative notions of algorithmic fairness in the context of *intersectionality*.
2. Reformulate well-known definitions of group fairness in the context of simple *sensitive networks*.
3. Find potential case studies and datasets in order to start a benchmark.
4. Implement measures of *network fairness* and evaluate them on the datasets.

Keywords

machine learning, fairness, graphs

Requirements

We are looking for a M2 mathematics/computer science student (or equivalent) with an interest and skills in data analysis, graph mining and fairness in machine learning. A background in the humanities (sociology, philosophy, *etc.*) is a big plus.

- With an interest for academic research. Able to work both in a team and independently
- With good python skills
- With good writing skills
- A good English level

References

[1] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[2] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[3] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349, 2022.

[4] Charlotte Laclau, Christine Largeron, and Manvi Choudhary. A survey on fairness for machine learning on graphs. *arXiv preprint arXiv:2205.05396*, 2022.