Master thesis announcement

Title: Inductive Adversarial Reweighting for Fairness-Aware Classification with Partial Demographic

Description

The pervasive integration of machine learning in people's lives, from social media recommendations to jurisdictional decisions, has created fairness concerns when automated decisions replicate and even amplify real-world biases against protected groups of data samples, such as people of certain genres or ethnicities or even create new types of biases. It has been found [1,2] that making classifiers fair (such as by achieving equalized odds[3] between protected and non-protected sample positive labels) can be achieved by skewing the weights of training examples based on their misclassification error and whether they belong to the protected group. However, protected group members may not be known. To address this problem, recent research [4] has proposed using adversarial learning to find training weights that penalize highly misclassified samples that are likely to belong to the protected group. However, if this group's samples tend to be misclassified more often, classification algorithms could overfit on those and hence introduce inverse discrimination, where the protected group is favored more.

In this master thesis we aim to produce adversarial weighting approaches that do not suffer from the above shortcoming by leveraging knowledge of some but not all protected group members. For example, jurisdictional decisions could take into account prior lawsuit outcomes. In detail, we want to enrich adversarial architectures with regularization towards achieving fairness for the known group members, so as to prevent both the original and inverse bias. Given that different types of fairness (e.g. positive label or misclassification odds) may be considered more important, we also plan to explore the efficacy of different types of weighting architectures and regularization schemes in letting adversarial training optimize different fairness-aware measures.

An ideal candidate should be:
- a self-motivated and independent learner
- knowledgeable about machine learning (indicated by good grades in related courses)
- experienced with Python or Java

The thesis will be co-supervised by Dr. Symeon Papadopoulos (papadop@iti.gr) and PhD candidate Emmanouil Krasanakis (maniospas@iti.gr) from the Information Technologies Institute, Greece.

References

[1] Krasanakis, Emmanouil, et al. "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification." Proceedings of the 2018 World Wide Web Conference. 2018.

[2] Iosifidis, Vasileios, and Eirini Ntoutsi. "AdaFair: Cumulative fairness adaptive boosting." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.

[3] Verma, S., & Rubin, J. (2018). Fairness definitions explained. In FairWare@ICSE (pp. 1–7). ACM.

[4] Lahoti, Preethi, et al. "Fairness without demographics through adversarially reweighted learning." arXiv preprint arXiv:2006.13114 (2020).