Croucher Summer Course in Information Theory
CUHK, August 23 ∼ 27, 2021

August 27, 2021
Changho Suh (chsuh@kaist.ac.kr)

Lecture 3: A fair & robust classifier and other fairness contexts

Summary of Lectures 1 and 2

We have thus far focused on fair classifiers. We first explored two prominent fairness measures in the realm of group fairness that I declared to focus on in this tutorial: (i) DDP that captures the degree of independence between prediction and an interested sensitive attribute; (ii) DEO that quantifies the degree of such independence yet conditioned on true labels. We then studied one fair classifier inspired by the most prominent (and possibly most favorite) information-theoretic notion: *mutual information*. We also investigated another fair classifier (known as the state of the art) that performs well both in accuracy-vs-fairness tradeoff and training stability. It is based on a well-known statistical measure also prevalent in information theory: *kernel density estimation*.

Now what is next? To introduce the last content, let us first revisit the five aspects which I emphasized yesterday in Lecture 1 as the requirements for enabling trustworthy AI. See Fig. 1 again. We have four remaining aspects not explored yet: robustness; explainability; value alignment; and transparency. The last content is relevant to one recent progress that we made towards addressing both fairness and robustness issues [2].

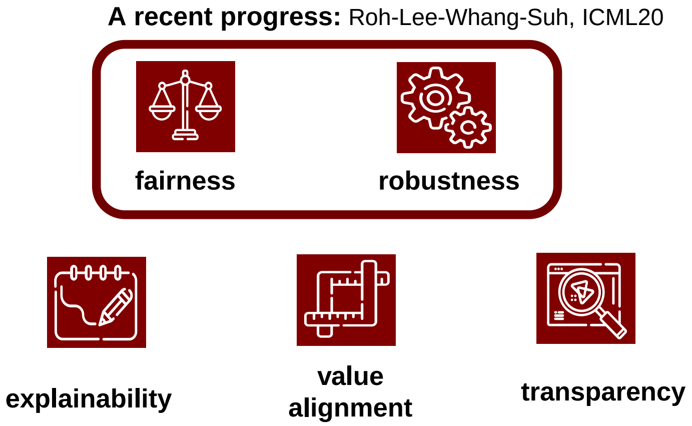

Figure 1: Revisit the five aspects for enabling trustworthy AI. A recent progress was made towards addressing both fair and robust training.

Today's lecture

In today's lecture, we will explore the recent work on fairness & robustness. This tutorial touches solely upon one context: fair classifiers. Obviously there are many others beyond this. So we will also discuss some other contexts. Specifically what we are going to cover are four folded. First we will introduce a robustness issue that arises in fair classifiers. We will then study a recent technique that ensures fairness in the presence of data poisoning. Next we will discuss other contexts such as fair recommender systems and fair ranking (yet in a very brief manner).

Lastly we will conclude the tutorial with a few remarks.

Robustness in fair classifiers

Let us start by figuring out what it means by *robustness*. It means that a trained model should guarantee a negligible performance degradation due to poisoned data. Here data poisoning refers to any negative action made on training data, such as adding noise or subjective (or possibly adversarial) perturbation.

A challenge

A challenge arises in the course of ensuring both fairness and robustness aspects. It turns out the accuracy-vs-fairness (e.g., DDP) tradeoff performance is significantly worsen in the presence of data poisoning if we naively applied a fair classifier without taking into account the robustness aspect. See Fig. 2 for instance. Here we plot accuracy-vs-DDP tradeoff performances for two different datasets (under COMPAS): (i) the original clean data; (ii) poisoned data in which 20% true labels are flipped in the focused binary classification setting along a direction of aggravating prediction accuracy. We employ the MI-based fair classifier [1]. Notice that the tradeoff curve (marked in red dots) under poisoned data is significantly shifted upward, implying a worsened tradeoff. This suggests that we need a non-trivial fair classifier in which accuracy-vs-fairness performance is well maintained even under poisoned data.

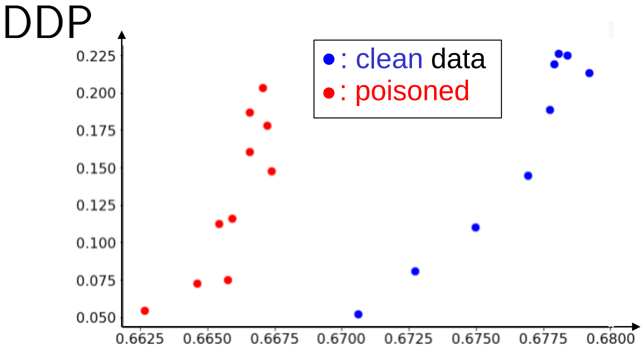

Figure 2: Accuracy-vs-DDP performance when applying a fair classifier [1] without consideration of the robustness aspect.

Idea for ensuring robustness [2]

To this end, we can also invoke a similar idea as employed in [1]. To figure out what it means, let us recall the MI-based fair optimization:

$$\min_{w} \frac{1-\lambda}{m} \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda \cdot I(Z; \hat{Y}). \tag{1}$$

Here we employ mutual information in an effort to enforce the Demographic Parity (DP) condition, reflected in $I(Z; \hat{Y})$. It turns out *mutual information* can also be instrumental in equipping with the robustness aspect.

For the robustness aspect, we impose a constraint on classifier hard-decision $\tilde{Y}$. The idea is to design $\tilde{Y}$ so that the classifier output together with input data (X, Z) looks like clean data. This way, we can somehow *sanitize data indirectly*. However, there is an issue in comparing to clean data. Since we target data poisoning scenarios, clean data may not be often available. To address the issue, we employ an additional *clean yet small validation dataset*. As for the degree of small, we consider, say a 5–10%-sized validation set, relative to the original real dataset, which can possibly be gathered with some efforts. Given the clean validation dataset, we then introduce a new random variable, say V, which indicates whether we take $(X, Z, \tilde{Y})$ (real and possibly poisoned) or $(X_{\mathsf{val}}, Z_{\mathsf{val}}, Y_{\mathsf{val}})$ (clean validation set):

$$(\bar{X}, \bar{Z}, \bar{Y}) = \begin{cases} (X, Z, \tilde{Y}) & \text{if } V = 1; \\ (X_{\mathsf{val}}, Z_{\mathsf{val}}, Y_{\mathsf{val}}) & \text{if } V = 0. \end{cases} \tag{2}$$

With these notations, the imposed constraint can then be translated to the *independence* between V and $(\bar{X}, \bar{Z}, \bar{Y})$, because the independence encourages $(X, Z, \tilde{Y})$ to indeed act as clean data:

$$I(V; \bar{X}, \bar{Z}, \bar{Y}) = 0. \tag{3}$$

Now what we want is to minimize $I(V; \bar{X}, \bar{Z}, \bar{Y})$. This naturally motivates us to consider the following regularized optimization:

$$\min_{w} \frac{1 - \lambda_1 - \lambda_2}{m} \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda_1 \cdot I(Z; \hat{Y}) + \lambda_2 \cdot I(V; \bar{X}, \bar{Z}, \bar{Y}) \tag{4}$$

where $0 \leq \lambda_1 \leq 1$ and $0 \leq \lambda_2 \leq 1 - \lambda_1$ indicate fairness and robustness regularization factors, respectively. A similar question arises. How to express $\lambda_2 \cdot I(V; \bar{X}, \bar{Z}, \bar{Y})$ in terms of the optimization variable w?

MI via function optimization

To this end, we can apply exactly the same trick that we saw in Lecture 1. Remember in Lecture 1 that $I(Z; \hat{Y})$ was approximated as the following implementable max-optimization:

$$I(Z; \hat{Y}) \approx \max_{D(\hat{y};z): \sum_z D(\hat{y};z)=1} \sum_{i=1}^{m} \frac{1}{m} \log D(\hat{y}^{(i)}; z^{(i)}) + H(Z). \tag{5}$$

Similarly for $I(V; \bar{X}, \bar{Z}, \bar{Y})$, one can readily obtain:

$$I(V; \bar{X}, \bar{Z}, \bar{Y}) \approx \max_{D(\bar{x},\bar{z},\bar{y};v): \sum_v D(\bar{x},\bar{z},\bar{y};v)=1} \sum_{i=1}^{m_{\mathsf{val}}} \frac{1}{m_{\mathsf{val}}} \log D(\bar{x}^{(i)}, \bar{z}^{(i)}, \bar{y}^{(i)}; v^{(i)}) + H(V) \tag{6}$$

where m_{val} is the number of examples in the validation set. Here the only distinctions are: we read v (instead of z), $(\bar{x}, \bar{z}, \bar{y})$ (instead of $\hat{y}$) and m_{val} (instead of m).

Implementable optimization

Applying (5) and (6) into (4) together with neural-net-based parameterizations (via θ and ϕ), we obtain the following implementable optimization:

$$\begin{aligned} \min_{w} \max_{\theta: \sum_z D_\theta(\hat{y};z)=1} \max_{\phi: \sum_v D_\phi(\bar{x},\bar{z},\bar{y};v)=1} & \frac{1 - \lambda_1 - \lambda_2}{m} \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) \\ & + \frac{\lambda_1}{m} \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)}; z^{(i)}) + \frac{\lambda_2}{m_{\mathsf{val}}} \sum_{i=1}^{m_{\mathsf{val}}} \log D_\phi(\bar{x}^{(i)}, \bar{z}^{(i)}, \bar{y}^{(i)}; v^{(i)}). \end{aligned} \tag{7}$$

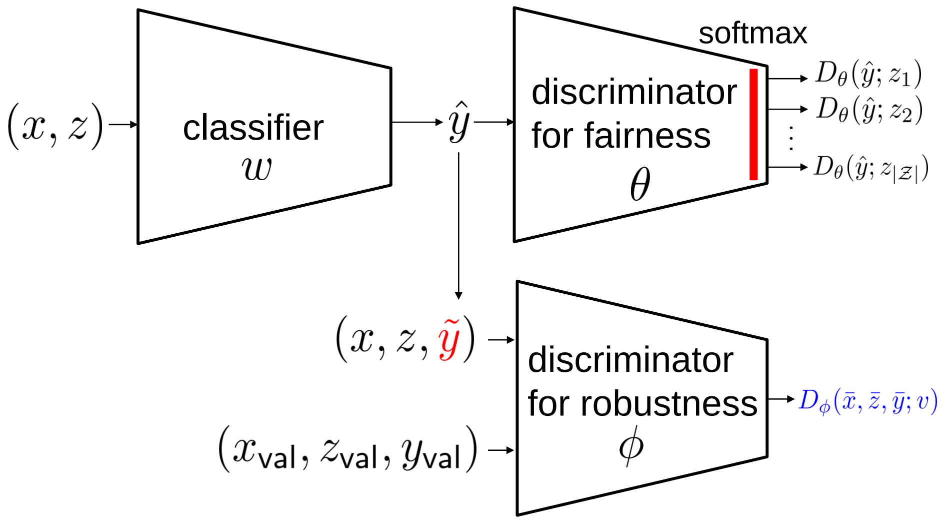

Figure 3: Architecture of the MI-based fair & robust classifier [2].

Here we have another max optimization w.r.t. D_ϕ, so this naturally motivates us to introduce another discriminator as illustrated in Fig. 3.

Architecture

See Fig. 3 for the architectural detail. We now have two discriminators. One is for fairness which we already visited in Lecture 1. The second is for robustness. The discriminator for robustness outputs another D, which indicates whether the input is either fake clean data $(x, z, \tilde{y})$ or real clean data $(x_{\mathsf{val}}, z_{\mathsf{val}}, y_{\mathsf{val}})$. It turns out for a $5 - 10\%$-sized clean validation set, this classifier exhibits almost no decrease in tradeoff performance under data poisoning. Please see the next section for detail.

Experiments

We consider the same benchmark real dataset COMPAS. Fig. 4 shows accuracy-vs-DDP performances for three cases: (i) applying the MI-based fair classifier with clean data; (ii) applying the same yet with 20%-label flipped poisoned data; and (iii) employing the fair and robust (FR) classifier under the poisoned data. Here we employ 5% validation set size relative to the original data. We see that the FR classifier yields a negligible performance degradation due to poisoning. We found this trend holds all the way down to $\sim 1\%$ validation set size. See [2] for detail.

Fairness in other contexts

As I promised earlier, before concluding this tutorial, I will leave a few words about other fairness contexts beyond fair classifiers that have been focused thus far. Two contexts. One is *fair recommender systems*. Here fairness means that a model ensures similar recommendation accuracies across different demographics. Or it means the model should recommend a diverse (non-biased) set of recommended items for every group. For instance, we may want to ban the situation in which science and math subjects are not recommended to woman group, which may often occur due to the inherent stereotype reflected in training data. The second is *fair ranking*. Here fairness means that top-ranked users should come from *diverse* groups (not from a single group). Or comparison data which is often employed for many ranking algorithms should encompass the broad scope covering the entire groups in a balanced manner. See Fig. 5.

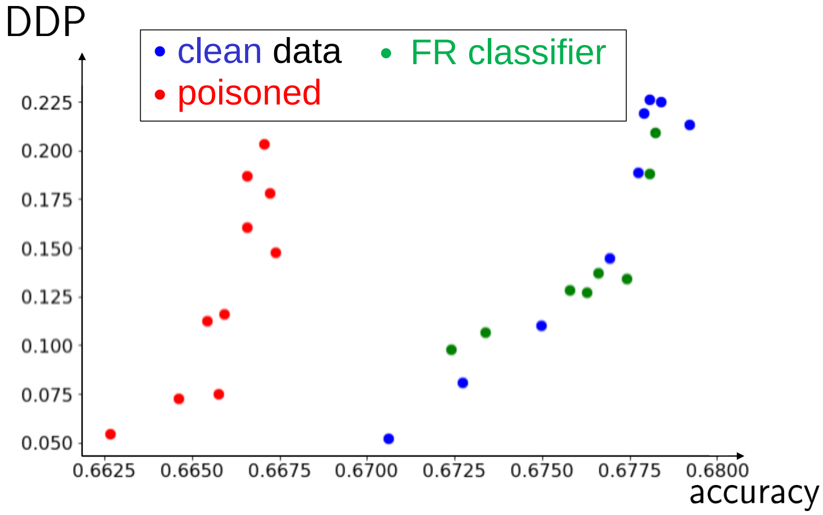

Figure 4: Accuracy-vs-DDP performance when applying a fair and robust (FR) classifier [2] (marked in green). Here we employ 5% validate set size. We also found that this robust performance maintains all the way down to $\sim 1\%$ validation set size.

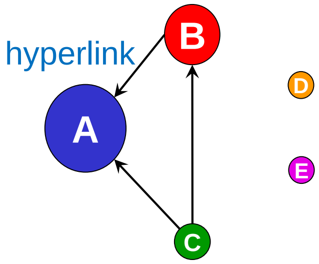

Figure 5: Pairwise comparison data employed for Google's PageRank. The websites D and E isolated can be unfavourably treated.

Due to the recent surge of interest in the fairness topic, there has been a proliferation of many recent works on fair recommender systems and fair ranking. Fig. 6 showcases only a partial list of them. Due to the interest of time, we will not cover them. But I believe these are different directions on fairness that you may want to pursue. In such a case, these references may give you some guideline.

Fair recommender systems

[Yao-Huang NeurIPS2017]

[Beutel et al. SIGKDD2019]

[Mehrotra et al. CIKM2018]

[Xiao et al. RecSys2017]

[Burke arXiv17]

Fair ranking

[Narasimhan et al. AAAI2020]

[Zehlike et al. CIKM2017]

[Singh et al. SIGKDD2018]

[Yadav et al. arXiv19]

Figure 6: References regarding fair recommender systems (left) and fair ranking (right).

A concluding remark

Now let me conclude. One last remark that I would like to leave. I believe fairness is a very important issue that often arises in a widening array of current & future applications. In this tutorial, I provided only a few instances where tools of information theory and statistics that you guys may be interested in are instrumental in addressing fairness issues that arise in the context of fair classifiers. I believe there would be many more such contexts. I hope you can exploit the tools of your interest and possibly your expertise to address many interesting fairness-relevant issues.

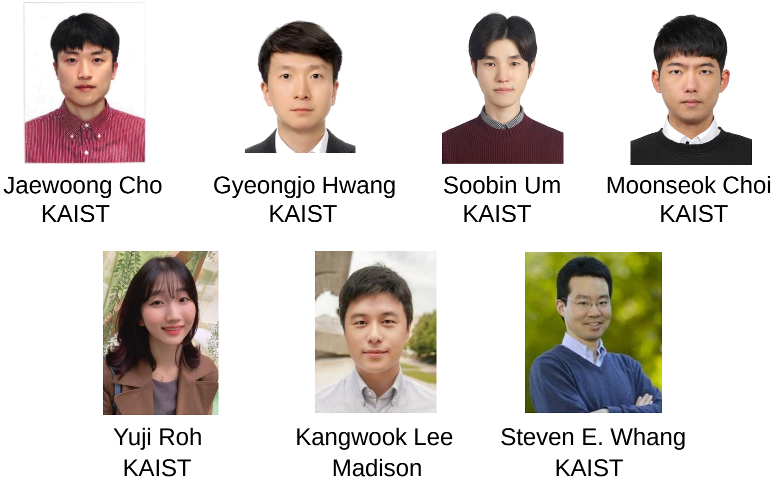

Figure 7: Collaborators of the works presented in this tutorial.

References

[1] J. Cho, G. Hwang and C. Suh. A fair classifier using mutual information. *IEEE International Syposium on Inofrmation Theory (ISIT)*, 2020.

[2] Y. Roh, K. Lee, S. E. Whang and C. Suh. FR-Train: A mutual information-based approach to fair and robust training. *In Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[3] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[4] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, et al. Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[5] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. *Proceedings of the 27th ACM international conference on information and knowledge management (CIKM)*, 2018.

[6] H. Narasimhan, A. Cotter, M. Gupta, and S. Wang. Pairwise fairness for ranking and regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[7] R. Burke. Multisided fairness for recommendation. *arXiv:1707.00093*, 2017.

[8] L. Xiao, Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, and M. Shaoping. Fairness-aware group recommendation with pareto-efficiency. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017.

[9] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017.

[10] Singh, Ashudeep, and Thorsten Joachims. Fairness of exposure in rankings. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[11] Yadav, Himank, Zhengxiao Du, and Thorsten Joachims. Fair learning-to-rank from implicit feedback. *arXiv:1911.08054*, 2019.