Fair machine learning

Lecture 3

Changho Suh
EE, KAIST

Aug. 27, 2021

A fair & robust classifier, other fairness contexts

Reading: **TN3**

Summary so far

1. Explored two prominent fairness measures:
   DDP and DEO

2. Studied one fair classifier based on mutual information.

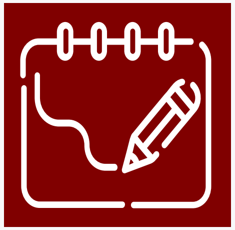
3. Investigated another based on kernel density estimation.

Revisit: Five aspects for trustworthy AI

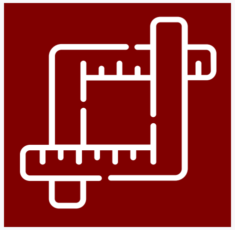
A recent progress: Roh-Lee-Whang-Suh, ICML20

fairness

robustness



explainability



value alignment

transparency

Today's lecture

Will explore the recent work on fairness & robustness, and discuss other contexts.

1. Introduce a robustness issue that arises in fair classifiers.

2. Study a recent technique that ensures fairness in the presence of data poisoning.

3. Discuss other contexts such as fair recommender systems and fair ranking.

4. Conclude the tutorial.

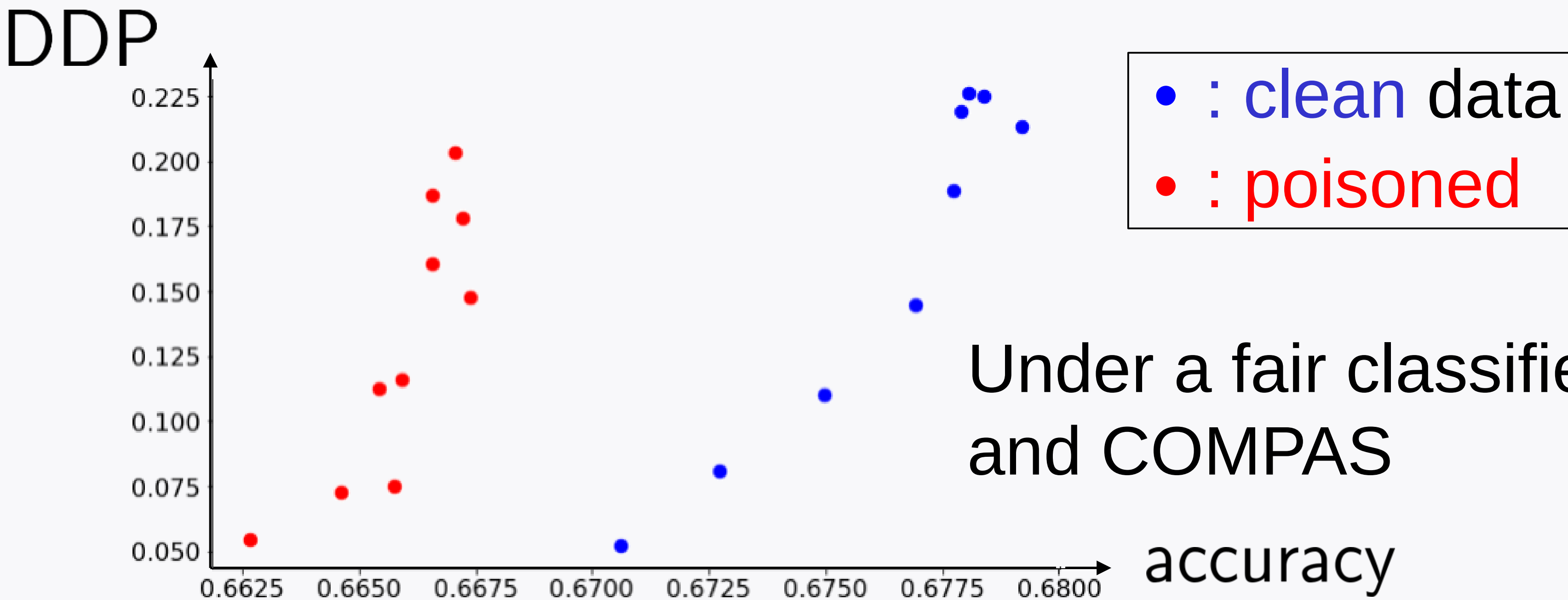
Robustness in fair classifiers?

It means: ensuring **negligible performance degradation** due to data poisoning.

Data poisoning refers to any negative action made on training data, such as adding noisy or subjective (or possibly adversarial) perturbation.

A challenge

Turns out: Accuracy-vs-fairness tradeoff is significantly worsen in the presence of data poisoning.



Under a fair classifier and COMPAS

Hence: Needs a **fair** classifier also being **robust** to data poisoning.

Insights from the prior work

Recall: MI-based optimization for a fair classifier

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

Turns out: *Mutual information* can also be instrumental in equipping the robustness aspect.

Idea for ensuring robustness

Impose a constraint on a classifier hard-decision $\tilde{Y}$:

$(X, Z, \tilde{Y})$ acts as a clean data.

This way: Can sanitize data *indirectly*.

Issue: Clean data may not be often available especially when we target data poisoning scenarios.

To address this issue, we employ an *additional clean* yet *small* validation dataset

↓

5-10% relative to the original real data

How to use clean validation set?

Impose a constraint on a classifier hard-decision $\tilde{Y}$:

$(X, Z, \tilde{Y})$ acts as a clean data.

Clean validation set: $\{(x_{\text{val}}^{(i)}, z_{\text{val}}^{(i)}, y_{\text{val}}^{(i)})\}_{i=1}^{m_{\text{val}}}$

Introduce a new random variable, say *V*, such that:

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

The constraint is then translated to: $I(V; \bar{X}, \bar{Z}, \bar{Y}) = 0$

Optimization for a fair and robust classifier

[Roh-Lee-Whang-Suh, ICML20]:

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

Question:

How to solve the optimization?

MI via function optimization

[Roh-Lee-Whang-Suh, ICML20]:

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

Remember:

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

Similarly:

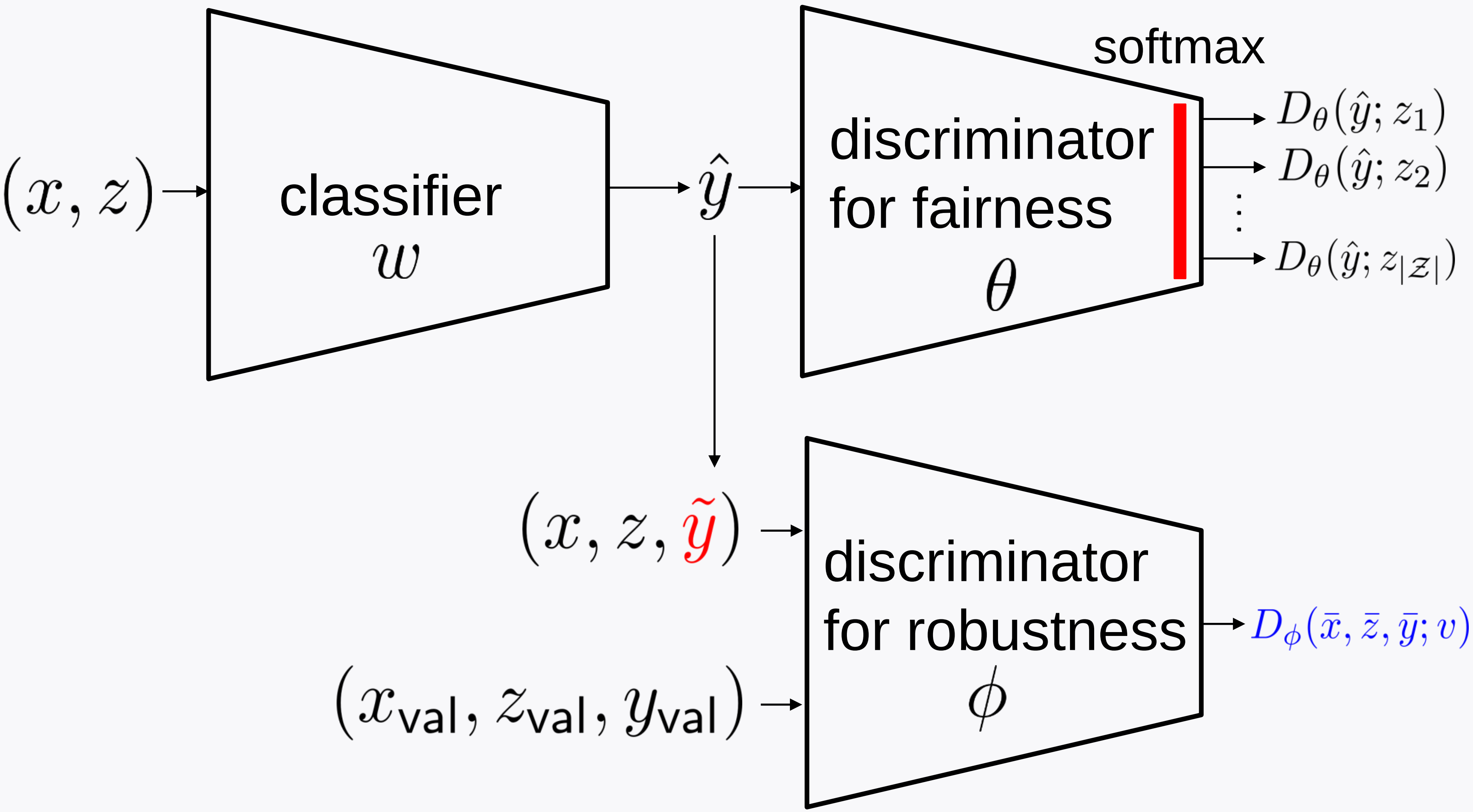
$$I(V;\bar{X},\bar{Z},\bar{Y}) \approx \max_{D(\bar{x},\bar{z},\bar{y};v):\sum_v D(\bar{x},\bar{z},\bar{y};v)=1} \sum_{i=1}^{m_{\mathsf{val}}}\frac{1}{m_{\mathsf{val}}}\log D(\bar{x}^{(i)},\bar{z}^{(i)},\bar{y}^{(i)};v^{(i)}) + H(V)$$

Implementable optimization

$$\min_{w} \max_{\theta: \sum_z D_\theta(\hat{y}; z)=1} \max_{\phi: \sum_v D_\phi(\bar{x}, \bar{z}, \bar{y}; v)=1} \frac{1-\lambda_1-\lambda_2}{m} \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)})$$

$$+ \frac{\lambda_1}{m} \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)}; z^{(i)}) + \frac{\lambda_2}{m_{\mathsf{val}}} \sum_{i=1}^{m_{\mathsf{val}}} \log D_\phi(\bar{x}^{(i)}, \bar{z}^{(i)}, \bar{y}^{(i)}; v^{(i)})$$

Architecture



Experiments

A benmark real dataset: **COMPAS**

$$(x, z, y)$$

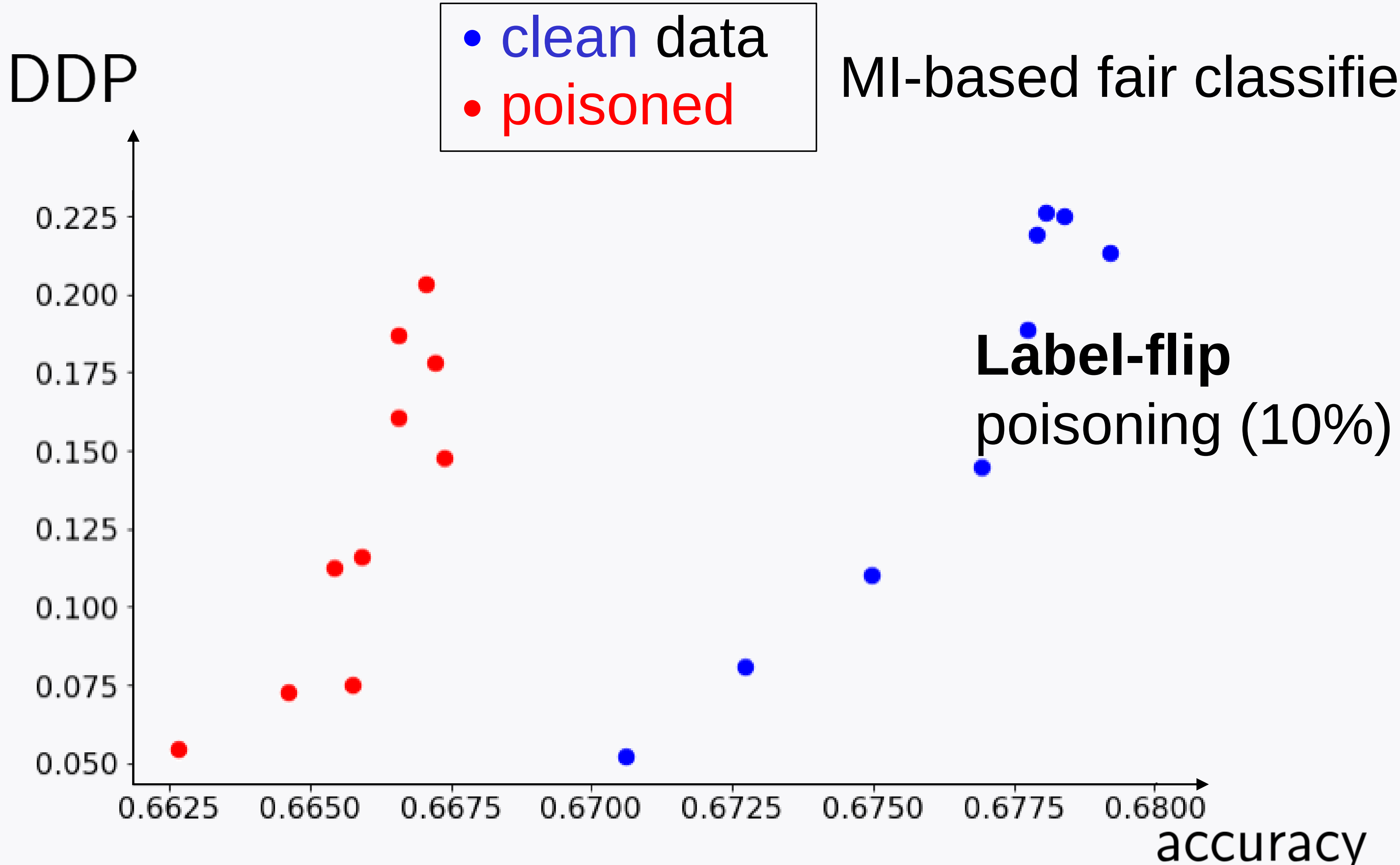
criminal records

black or white

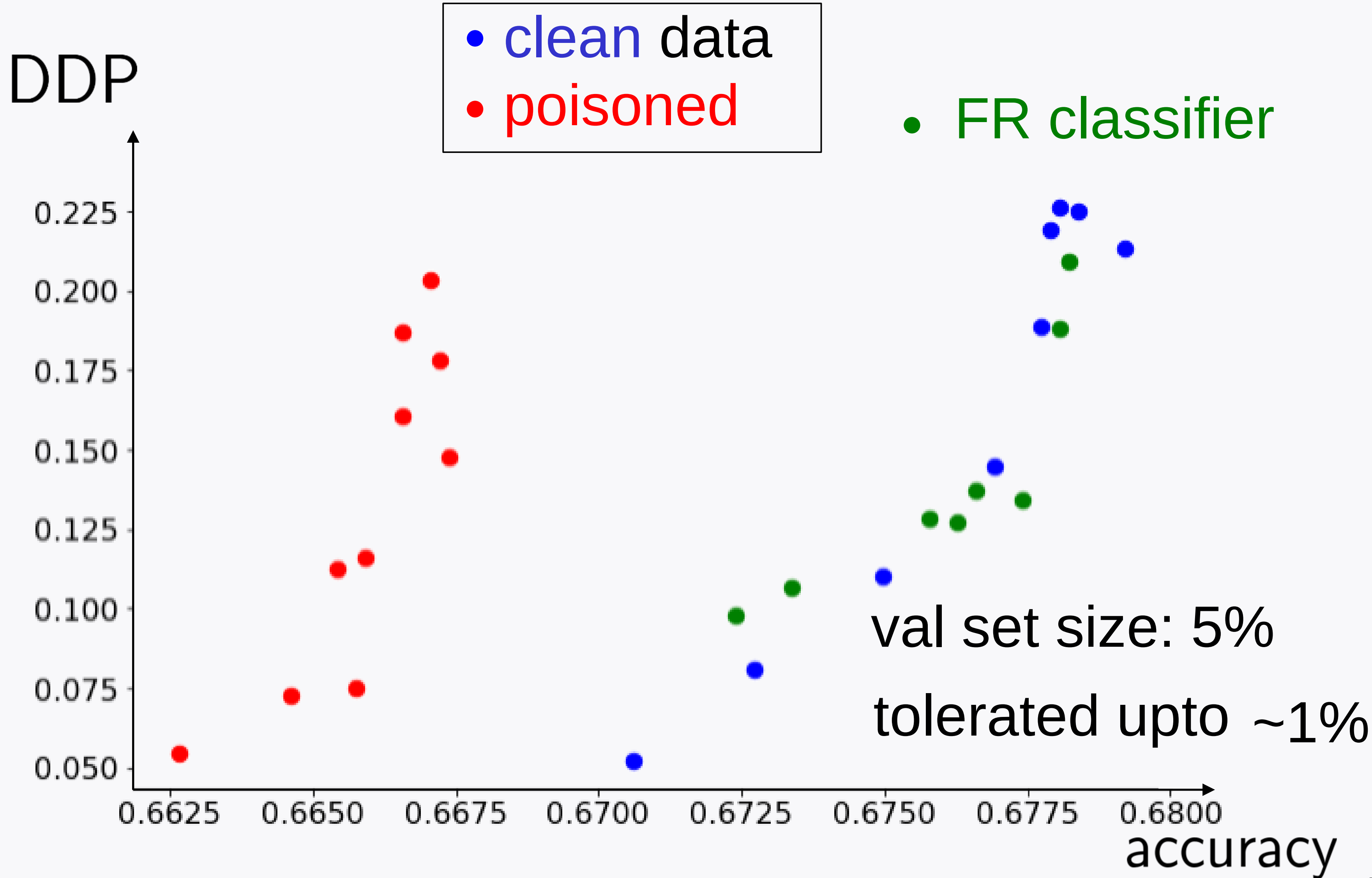
reoffend or not
in near future

Recall: Worsen tradeoff due to poisoning

MI-based fair classifier



Fair and Robust (FR) classifier



Other fairness contexts

Fair recommender systems

Fairness means: *Similar* recommendation *accuracies* across different demographics

Or it means: A diverse set of items should be recommended for every group.

- Example: STEM courses for women

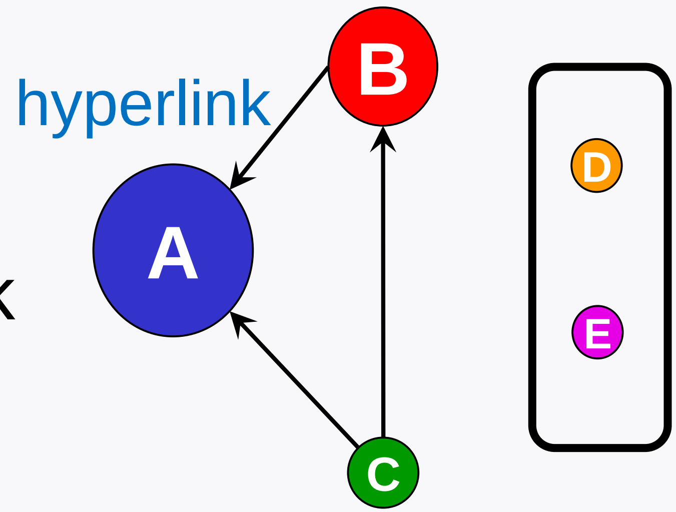
Fair ranking

Fairness means: Top-ranked users from *diverse* groups

Or it means: Data employed for ranking should not be biased.

Example: Localized comparison data

Google's PageRank



these isolated websites can be unfavorably treated

Recent works

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]

If you pursue these research directions, the references might give you some guideline.

A concluding remark

Fairness becomes more crucial in many current & future applications.

Expect: Information-theoretic tools explored in this tutorial would help address many fairness-relevant issues.

Acknowledgement

Jaewoong Cho
KAIST

Gyeongjo Hwang
KAIST

Soobin Um
KAIST

Moonseok Choi
KAIST

Yuji Roh
KAIST

Kangwook Lee
Madison

Steven E. Whang
KAIST

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. *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.

References

[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.