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

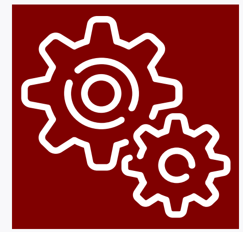
2. Studied one fair classifier based on mutual information.

3. Investigated another based on kernel density estimation.

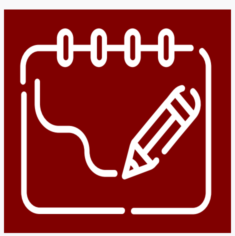
Revisit: Five aspects for trustworthy AI

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

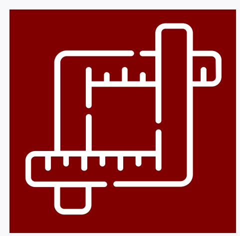
fairness



robustness



explainability



value alignment

transparency

Today’s lecture

Will explore the recent work on fairness & robustness, and discuss other contexts (beyond classifiers).

1. Figure out what it means by robustness in fair classifiers.
2. Study a fair & robust classifier.
3. Investigate experimental results.
4. Discuss other contexts such as fair recommender systems and fair ranking.
5. Conclude the tutorial.

Robustness in fair classifiers?

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

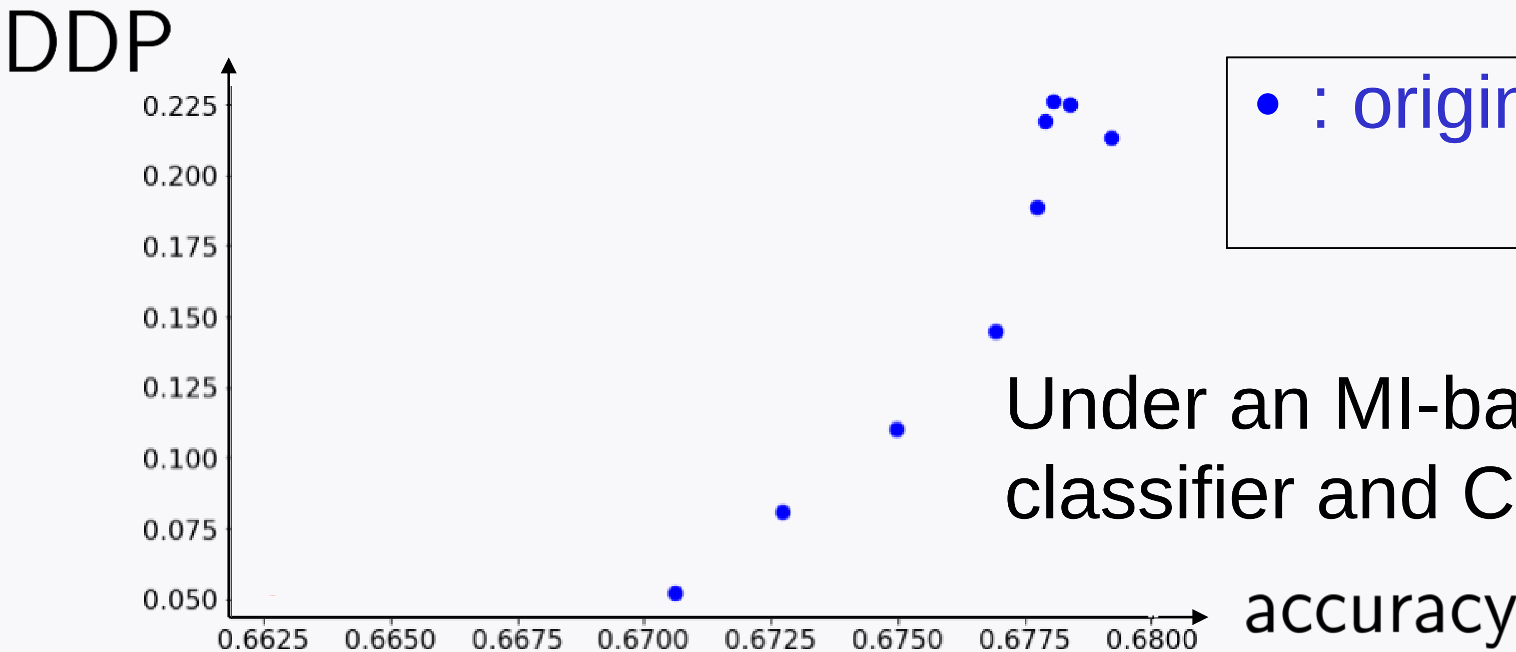
Performance metric: Accuracy-vs-fairness tradeoff

Data poisoning: Any negative action applied to training data.

> **Example:** Adding noisy perturbation either to label or to sensitive attribute

A challenge

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

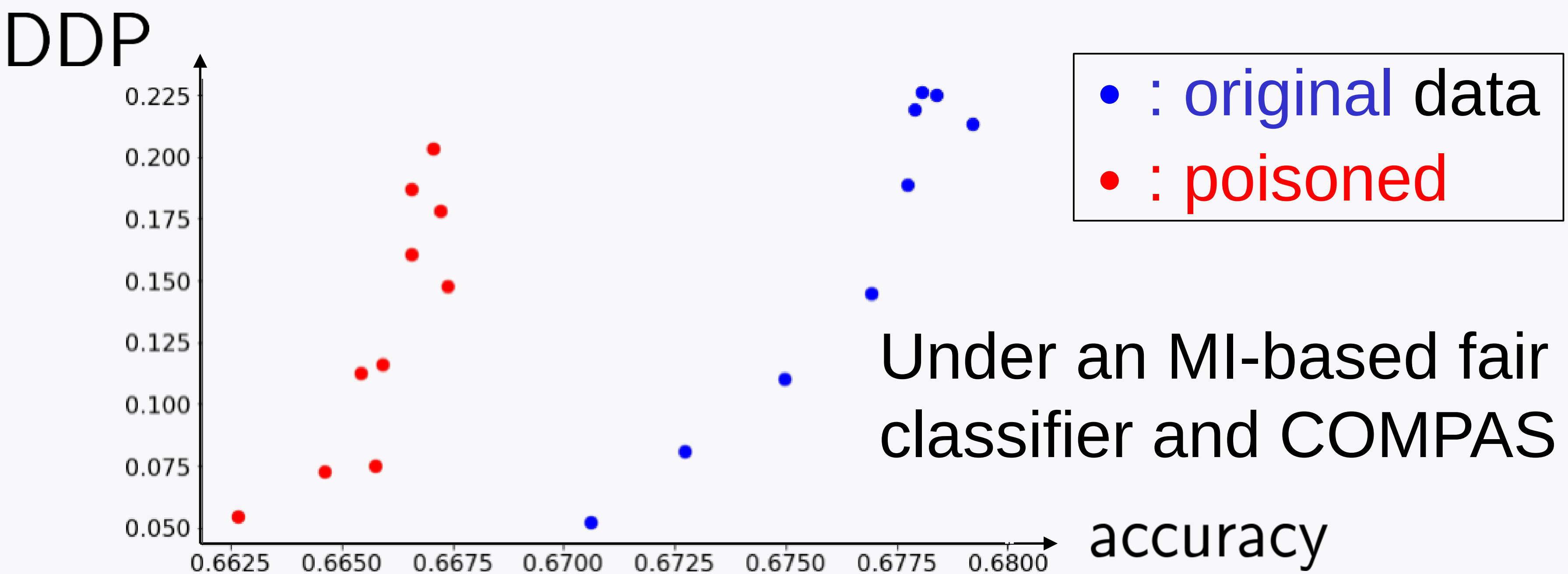


Under an MI-based fair classifier and COMPAS

Consider 10% label flipping.

A challenge

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



Under an MI-based 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

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

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

Idea for ensuring robustness

Sanitize data $(X, Z, \tilde{Y})$ *indirectly*:

By perturbing $\tilde{Y}$ while not changing (X, Z)

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

Issue in implementing the idea

Idea: Sanitize data $(X, Z, \tilde{Y})$ *indirectly*:

By perturbing $\tilde{Y}$ while not changing (X, Z)

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

Issue: We need *clean validation data* to compare with.

But clean data may be difficult to obtain especially when we target data poisoning scenarios.

Desired properties of validation dataset

Idea: Sanitize data $(X, Z, \tilde{Y})$ *indirectly*:

By perturbing $\tilde{Y}$ while not changing (X, Z)

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

1. Clean

2. Small e.g., 5-10% relative to the original real data

How to use clean validation set? $\{(x_{\text{val}}^{(i)}, z_{\text{val}}^{(i)}, y_{\text{val}}^{(i)})\}_{i=1}^{m_{\text{val}}}$

Idea: Sanitize data $(X, Z, \tilde{Y})$ *indirectly*:

By perturbing $\tilde{Y}$ while not changing (X, Z) so that $(X, Z, \tilde{Y})$ acts as a clean data.

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

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

Want to make poisoned data indistinguishable from clean validation data.

How to use clean validation set? $\{(x_{\text{val}}^{(i)}, z_{\text{val}}^{(i)}, y_{\text{val}}^{(i)})\}_{i=1}^{m_{\text{val}}}$

Idea: Sanitize data $(X, Z, \tilde{Y})$ *indirectly*:

By perturbing $\tilde{Y}$ while not changing (X, Z)

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

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

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

→ Can be 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 implement?

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:

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

paraterize w/ θ

Similarly:

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

paraterize w/ ϕ

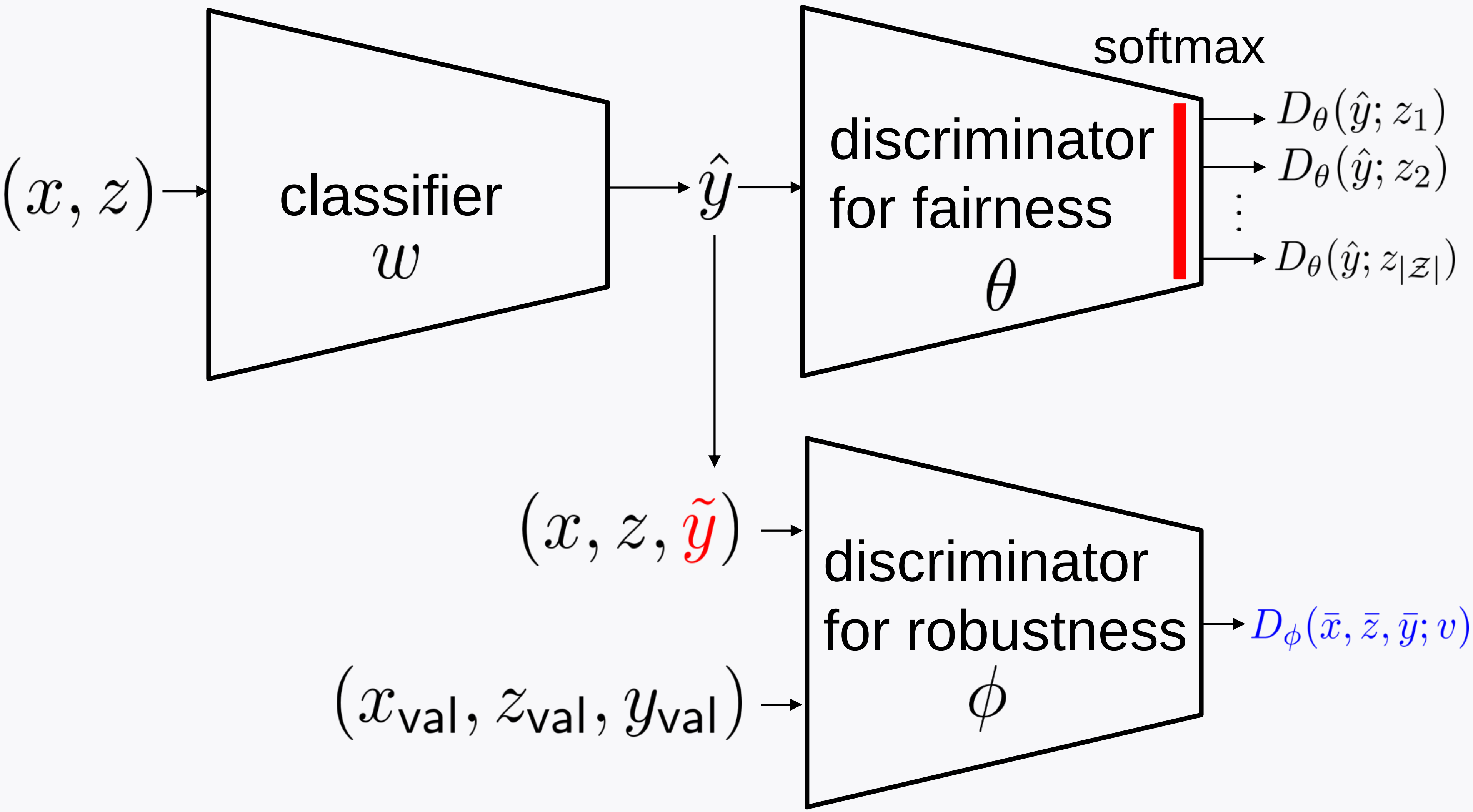
Implementable optimization

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

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

Algorithm: Alternating gradient descent

Architecture



Experiments

A benmark real dataset: **COMPAS**

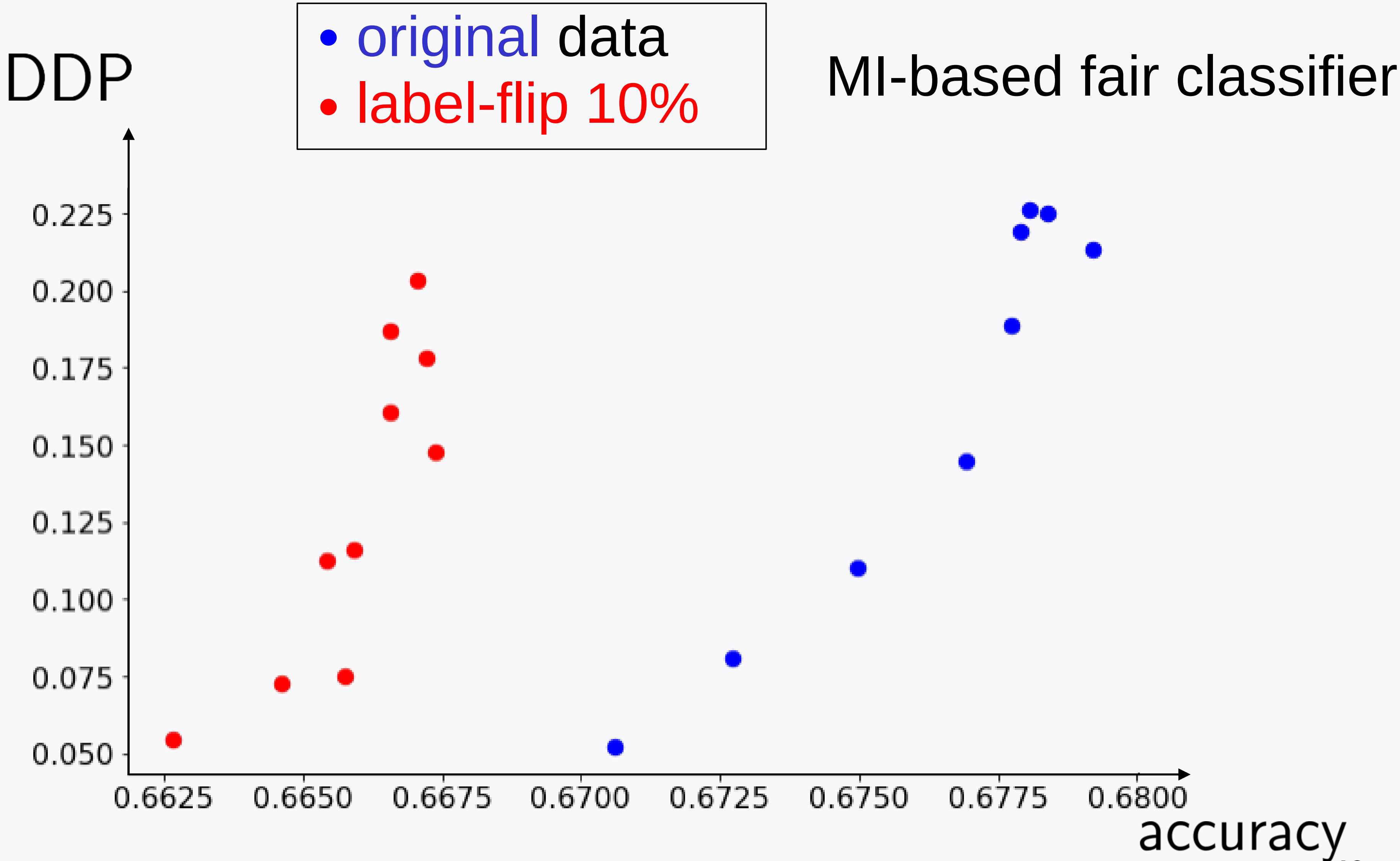
$$(x, z, y)$$

criminal records

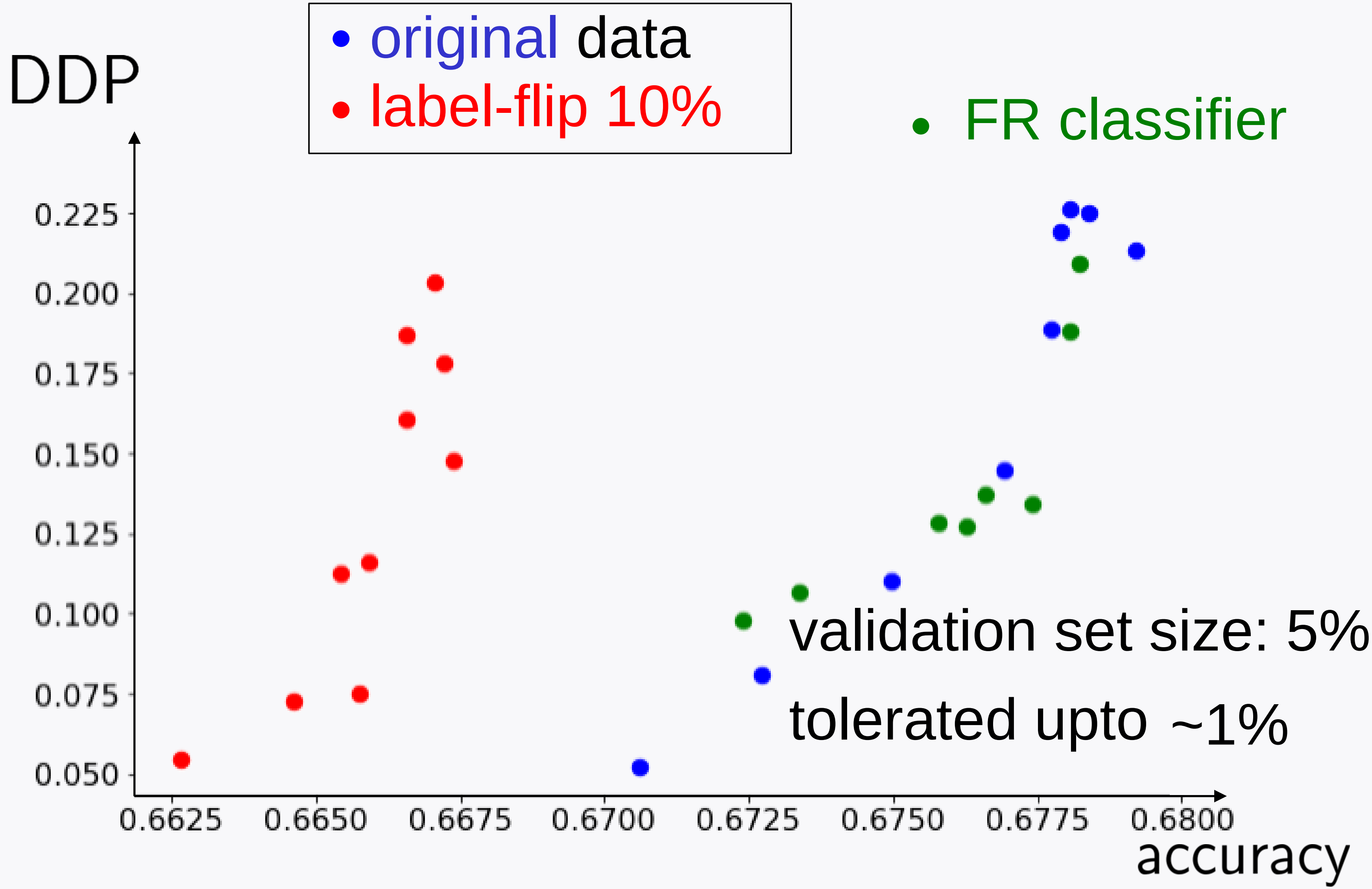
black or white

reoffend or not
in near future

Recall: Worsen tradeoff due to poisoning



Fair and Robust (FR) classifier



Other fairness contexts

Fair recommender systems

$$\tilde{Y} = 1 \text{ (like) or } 0 \text{ (dislike)}$$

Fairness means: Recommendation statistics is irrelevant to sensitive attributes of groups.

An example in which fairness issue arises:

Subject (course) recommendation

Consider: STEM courses for women

→ No or low rating (unfair)

How to address such unfairness?

Recent works on fair recommender systems

Pursue: $\tilde{Y} \perp Z_{\text{item}}$

[Yao-Huang NeurIPS2017]

[Beutel et al. SIGKDD2019]

[Mehrotra et al. CIKM2018]

[Xiao et al. RecSys2017]

[Burke arXiv2017]

[Kamishima-Akaho RecSys2017]

[Li et al. arXiv2021]

Pursue: $\tilde{Y} \perp Z_{\text{user}}$

Proposed particular ways to promote such independence.

If you are interested, you may want to try different ways to promote.

Fair ranking

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

Example: Poster prizes

Suppose: Winners come only from a certain group

→ Perhaps considered to be unfair

Recent works on fair ranking

[Narasimhan et al. AAAI2020]

[Zehlike et al. CIKM2017]

[Singh et al. SIGKDD2018]

[Yadav et al. arXiv19]

[Konstantinov et al. arXiv21]

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] T. Kamishima and S. Akaho. Considerations on recommendation independence for a good-items task. *RecSys* 2017.

[10] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang. User-oriented Fairness in Recommendation. *arXiv:2104.10671*, 2021.

References

[11] 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.

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

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

[14] N. Konstantinov and C. H. Lampert. Fairness through regularization for learning to rank. *arXiv:2102.025996,* 2021.

backup

FR classifier based on KDE?

Recall the new variable *V*:

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

Instead of $I(V; \bar{X}, \bar{Z}, \bar{Y})$, one may want to minimize:

$$\sum_x \sum_z \sum_y |\mathbb{P}(\bar{X} = x, \bar{Z} = z, \bar{Y} = y | V = 1) - \mathbb{P}(\bar{X} = x, \bar{Z} = z, \bar{Y} = y | V = 0)|$$

Issue: KDE of $\mathbb{P}(\bar{X} = x, \bar{Z} = z, \bar{Y} = y | V = 1)$ may not be accurate for moderate *m*.

Reason: Dimension of $(\bar{X}, \bar{Z}, \bar{Y})$ is large!