Fair machine learning

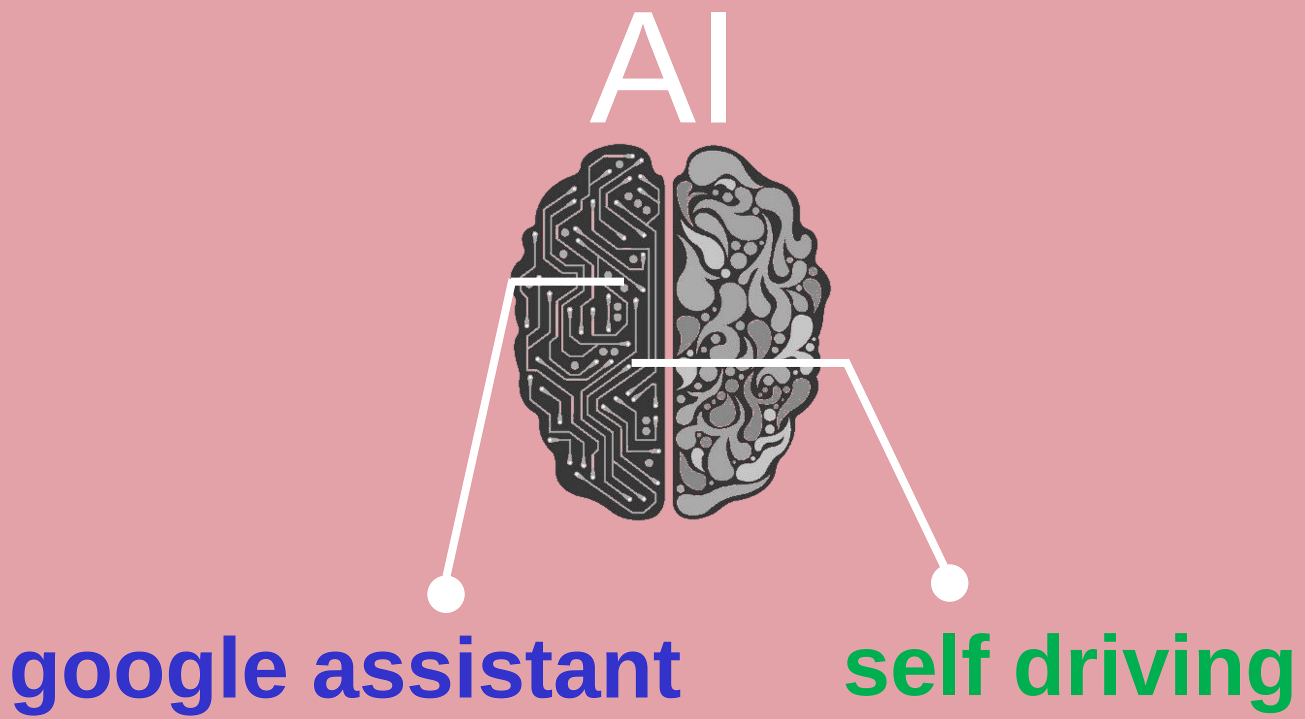
Lecture 1

Changho Suh
EE, KAIST

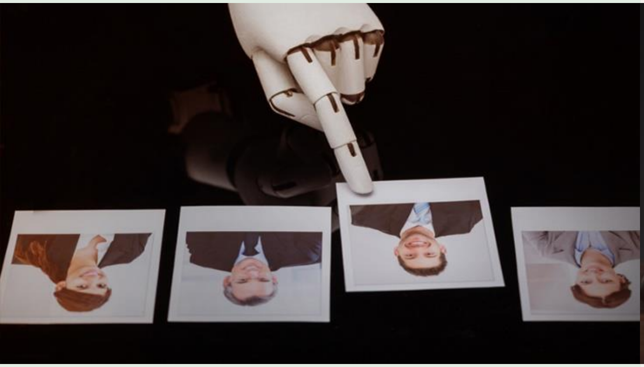
Aug. 23, 2021

Overview & a fair classifier using mutual information

Reading: **Tutorial Note (TN) 1**


AI
google assistant
self driving

recruiting **judgement** **loan decision**



Trustworthy AI

"AI has significant potential to help solve challenging problems, including by advancing medicine, understanding language, and fueling scientific discovery. ***To realize that potential****, it's critical that* ***AI*** *is used and developed* ***responsibly."***

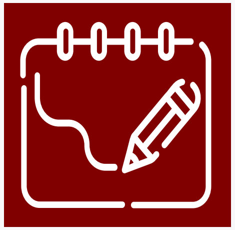
"Moving forward, "build for performance" will not suffice as an AI design paradigm. We must learn how to build, evaluate and monitor for ***trust****."*

Five aspects of trustworthy AI

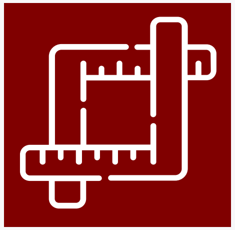
focus of this tutorial

fairness

robustness



explainability



value alignment

transparency

A ML model of this tutorial's focus

Classifier!

Will explore fairness & robustness issues that arise in **classifiers.**

Outline of this tutorial

Lecture 1 (Today):

Figure out what it means by fairness in classifiers.

Study one fair classifier using *mutual information.*

Lecture 2 (Wed):

Investigate another fair classifier that offers better performance.

It employs a statistical technique prevalent in information theory: *Kernel Density Estimation (KDE)*

Lecture 3 (Fri):

Explore another fair classifier also being *robust to data poisoning*.

A fair classifier using mutual information

Fairness in the context of classifiers?

There are many fairness concepts.

One important concept is **group fairness**:

Pursues predictions to exhibit similar statistics regardless of sensitive attributes of groups

e.g., race, gender, age, religion, etc.

Applications of fair classifiers

job hiring

Applicants want no discrimination depending on race or sex.

parole decision (假釋放判決)

A fair predictor for recidivism (再犯) score plays a crucial role.

A fairness measure

Feldman et al. SIGKDD15
Zafar et al. AISTATS17

Y: class $\in \{0, 1\}$ $\tilde{Y}$: prediction (hard decision)

no reoffend reoffend

black white

Z: sensitive attribute e.g., $\in \mathcal{Z} = \{0, 1\}$

Demographic Parity (DP) condition:

$$\tilde{Y} \perp Z : \quad \mathbb{P}(\tilde{Y} = 1 | Z = z) = \mathbb{P}(\tilde{Y} = 1), \forall z \in \mathcal{Z}$$

A quantifed measure: **D**ifference btw two interested probabilities in **DP** condition

$$\text{DDP} := \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|$$

Limitation of DP condition

Demographic Parity (DP) condition:

$$\tilde{Y} \perp Z: \quad \mathbb{P}(\tilde{Y} = 1 | Z = z) = \mathbb{P}(\tilde{Y} = 1), \forall z \in \mathcal{Z}$$

Suppose that the ground-truth label dist. respects:

$$\mathbb{P}(Y = 1 | Z = 1) \gg \mathbb{P}(Y = 1 | Z = 0)$$

Enforcing the DP condition may aggravate prediction accuracy significantly.

Another fairness notion

Hardt-Price-Srebo NeurIPS16

Equalized Odds (EO) condition: $\tilde{Y} \perp Z \mid Y$

$$\mathbb{P}(\tilde{Y} = 1 | Y = y, Z = z) = \underline{\mathbb{P}(\tilde{Y} = 1 | Y = y)} \;\; \forall z \in \mathcal{Z}, \forall y \in \mathcal{Y}$$

relevant to prediction accuracy

Enforcing the EO condition has little to do with reducing prediction accuracy.

A quantified measure:

$$\text{DEO} := \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \left| \mathbb{P}(\tilde{Y} = 1 | Y = y, Z = z) - \mathbb{P}(\tilde{Y} = 1 | Y = y) \right|$$

Many recent works on fair classifiers

Here is only a *partial* list:

[Feldman et al. SIGKDD15]

[Hardt-Price-Srebo NeurIPS16]

[Pleiss et al. NeurIPS17]

[Zhang et al. AIES18]

[Donini et al. NeurIPS18]

[Agarwal et al. ICML18]

[Roh-Lee-Whang-**Suh** ICLR 21]

[Zafar et al. AISTATS17]

[Cho-Hwang-**Suh** ISIT20]

[Roh-Lee-Whang-**Suh** ICML20]

[Cho-Hwang-**Suh** NeurIPS20]

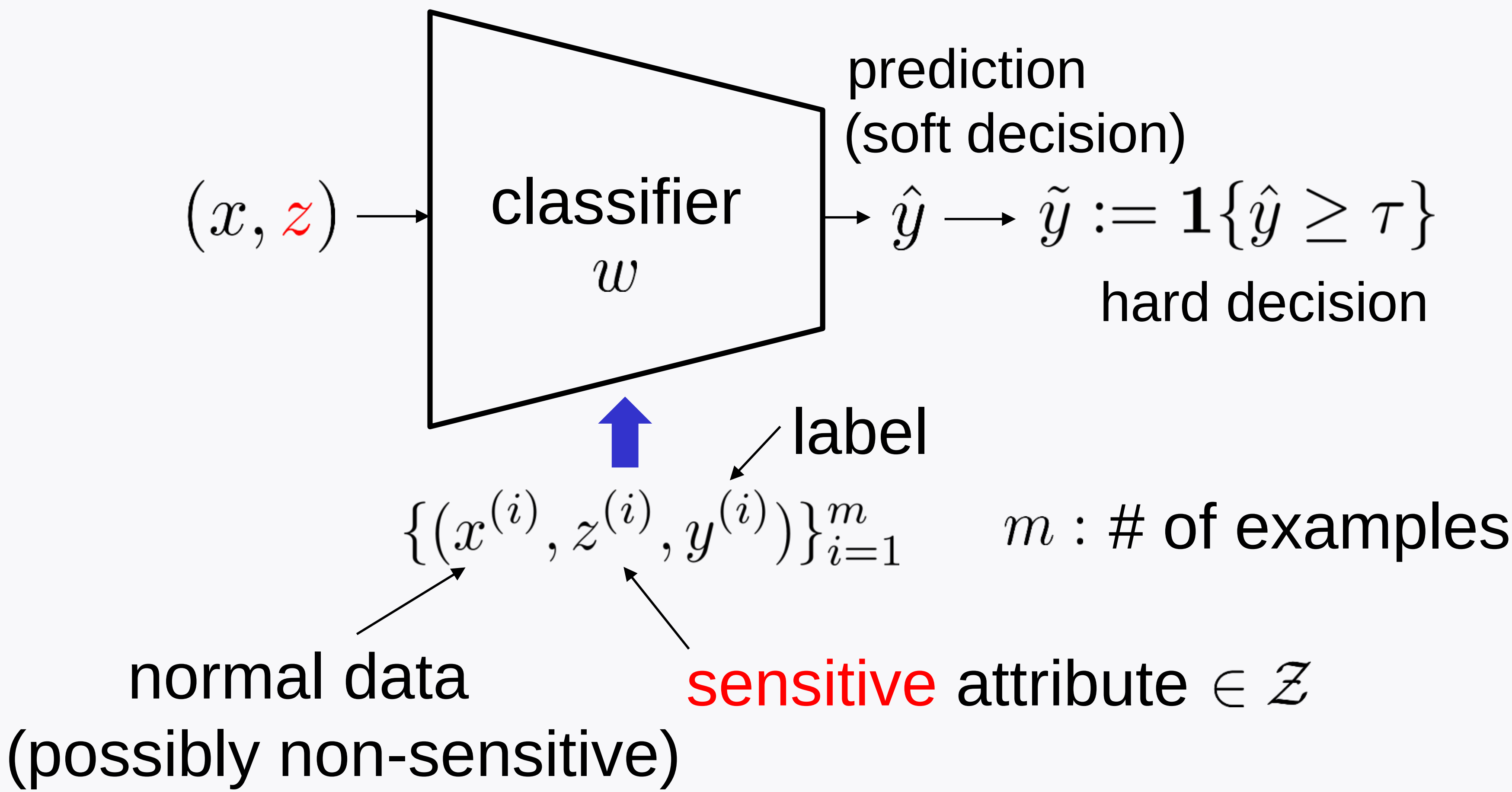
[Baharlouei et al. ICLR20]

[Jiang et al. UAI20]

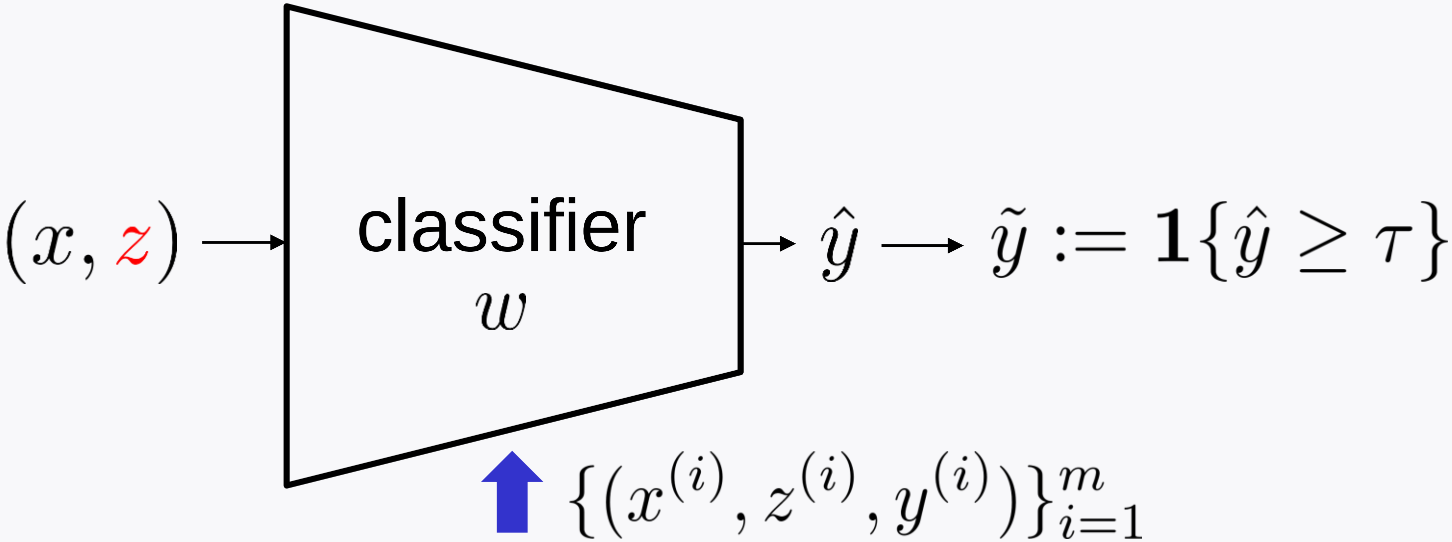
[Lee et al. arXiv 20]

employ mutual information

Problem setting



Problem setting



For illustrative purpose, this tutorial focuses on:

(i) binary classifier &

(ii) one fairness measure:

$$\text{DDP} := \sum_{z \in \mathcal{Z}} \left| \mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1) \right|$$

Optimization

Conventional optimization for classifiers:

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} \underbrace{\ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)})}_{\text{cross entropy loss}}$$

$$-y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

How to incorporate the fairness measure DDP?

$$\mathsf{DDP} := \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|$$

Observation: The smaller DDP, the more fair.

Enforcing fairness via regularization

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

where $\mathsf{DDP} := \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|$

Challenge: DDP is a complicated function of the optimization variable *w*.

Will study another approach which employs a different regularization term.

It is based on a connection between DDP and mutual information.

Connection btw DDP & mutual information

Observation:

$$\mathsf{DDP} = 0 : \tilde{Y} \perp Z \quad \Longleftrightarrow \quad I(Z; \tilde{Y}) = 0$$

$$\Uparrow \; \tilde{Y} := \mathbf{1}\{\hat{Y} \geq \tau\}$$

$$I(Z; \hat{Y}) = 0$$

MI-based approach

Cho-Hwang-Suh ISIT20

Connection:

$$\text{DDP} = 0 : \tilde{Y} \perp Z \Longleftarrow I(Z; \hat{Y}) = 0$$

Idea: Employ $\lambda \cdot I(Z; \hat{Y})$ (instead of $\lambda \cdot$ DDP)

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

How to express it with *w*?

A careful look at mutual information

$$I(Z;\hat{Y}) = H(Z) - H(Z|\hat{Y}) = H(Z) - (H(\hat{Y},Z) - H(\hat{Y}))$$

$$= H(Z) + \mathbb{E}\left[\log\frac{1}{\mathbb{P}_{\hat{Y}}(\hat{Y})}\right] - \mathbb{E}\left[\log\frac{1}{\mathbb{P}_{\hat{Y},Z}(\hat{Y},Z)}\right]$$

$$= H(Z) + \sum_{\hat{y},z}\mathbb{P}_{\hat{Y},Z}(\hat{y},z)\log\underbrace{\frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})}}_{=:\, D^*(\hat{y};z)}$$

$$\sum_{z} D^*(\hat{y};z) = 1 \quad \forall \hat{y}$$

MI via function optimization

$$I(Z;\hat{Y}) = H(Z) + \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log \underbrace{\frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})}}_{=:\, D^*(\hat{y};z)}$$

$$\sum_{z} D^*(\hat{y};z) = 1 \ \ \forall \hat{y}$$

Theorem:

$$I(Z;\hat{Y}) = H(Z) + \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z)$$

Proof of Theorem

$$D^*(\hat{y};z) := \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})} \qquad \sum_z D^*(\hat{y};z) = 1 \;\; \forall \hat{y}$$

Theorem: concave in *D*

$$I(Z;\hat{Y}) = H(Z) + \underbrace{\max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1}} \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z)$$

Lagrange function:

$$\mathcal{L}(D(\hat{y};z),\nu(\hat{y})) = \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z) + \sum_{\hat{y}} \nu(\hat{y}) \left(1 - \sum_z D(\hat{y};z)\right)$$

KKT condition:

$$\left.\frac{d\mathcal{L}(D(\hat{y};z),\nu(\hat{y}))}{dD(\hat{y};z)}\right|_{D=D_{\text{opt}},\nu=\nu_{\text{opt}}} = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{D_{\text{opt}}(\hat{y};z)} - \nu_{\text{opt}}(\hat{y}) = 0 \qquad \forall \hat{y}, z$$

$$\sum_z D_{\text{opt}}(\hat{y};z) = 1 \qquad \forall \hat{y}$$

Proof of Theorem

$$D^*(\hat{y};z) := \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})} \qquad \sum_z D^*(\hat{y};z) = 1 \ \ \forall \hat{y}$$

Theorem:

$$I(Z;\hat{Y}) = H(Z) + \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z)$$

KKT condition:

$$\left.\frac{d\mathcal{L}(D(\hat{y};z),\nu(\hat{y}))}{dD(\hat{y};z)}\right|_{D=D_{\text{opt}},\nu=\nu_{\text{opt}}} = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{D_{\text{opt}}(\hat{y};z)} - \nu_{\text{opt}}(\hat{y}) = 0 \quad \forall \hat{y},z$$

$$\sum_z D_{\text{opt}}(\hat{y};z) = 1 \quad \forall \hat{y} \qquad \rightarrow D_{\text{opt}}(\hat{y};z) = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\nu_{\text{opt}}(\hat{y})}$$

$$\frac{\sum_z \mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\nu_{\text{opt}}(\hat{y})} = 1 \quad \rightarrow \nu_{\text{opt}}(\hat{y}) = \mathbb{P}_{\hat{Y}}(\hat{y}) \rightarrow D_{\text{opt}}(\hat{y};z) = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})} = D^*(\hat{y};z)$$

How to express $I(Z;\hat{Y})$ in terms of *w*?

$$\boxed{I(Z;\hat{Y}) = H(Z) + \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z)}$$

$P_{\hat{Y},Z}(\hat{y},z)$ not available!

Rely on **empirical** distributions: $\mathbb{Q}_{\hat{Y},Z}(\hat{y}^{(i)}, z^{(i)}) = \frac{1}{m}$

$$I(Z;\hat{Y}) \approx \underbrace{H(Z)}_{\text{irrelevant of } (\theta, w)} + \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{i=1}^{m} \frac{1}{m} \log D(\hat{y}^{(i)}; z^{(i)})$$

Parameterize $D(\cdot;\cdot)$ with θ

Implementable optimization

$$\min_{w} \max_{\theta:\sum_z D_\theta(\hat{y};z)=1} \frac{1}{m}\left\{\sum_{i=1}^{m}(1-\lambda)\ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) + \lambda\sum_{i=1}^{m}\log D_\theta(\hat{y}^{(i)};z^{(i)})\right\}$$

How to solve?

Algorithm: *Alternating* gradient descent:

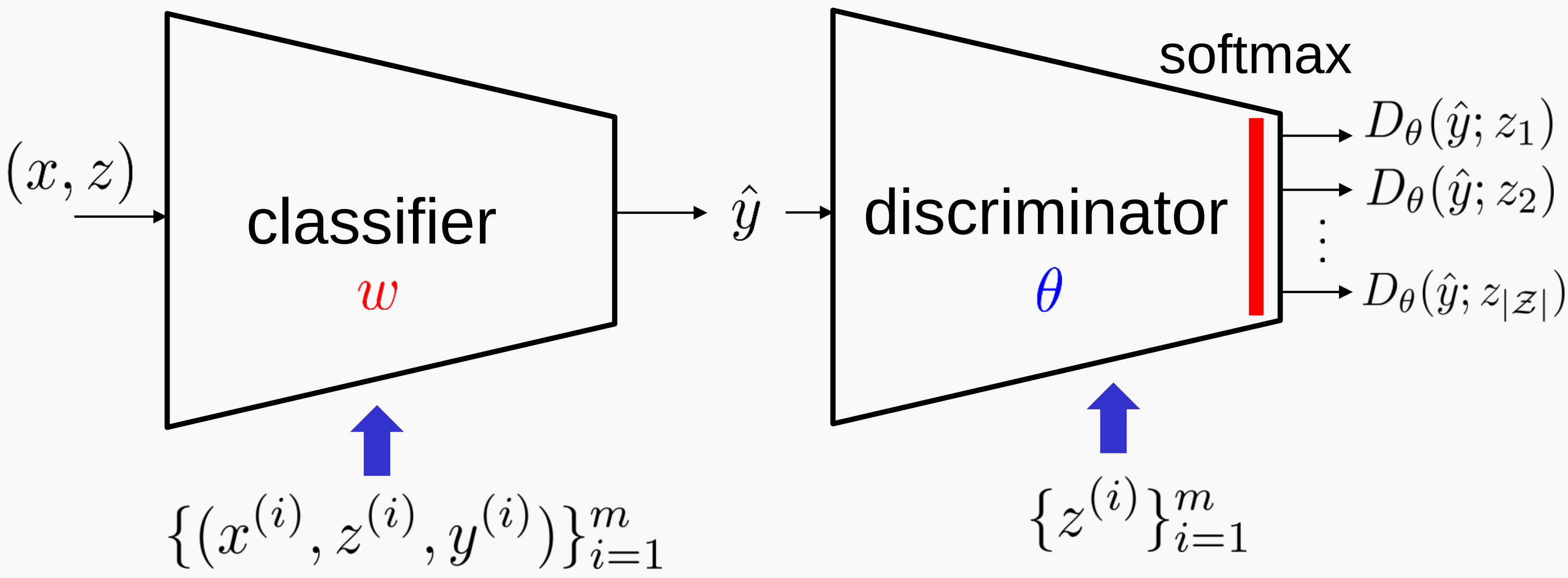
(i) Given w, update θ via the inner opt;

(ii) Given the updated θ, update w via the outer opt;

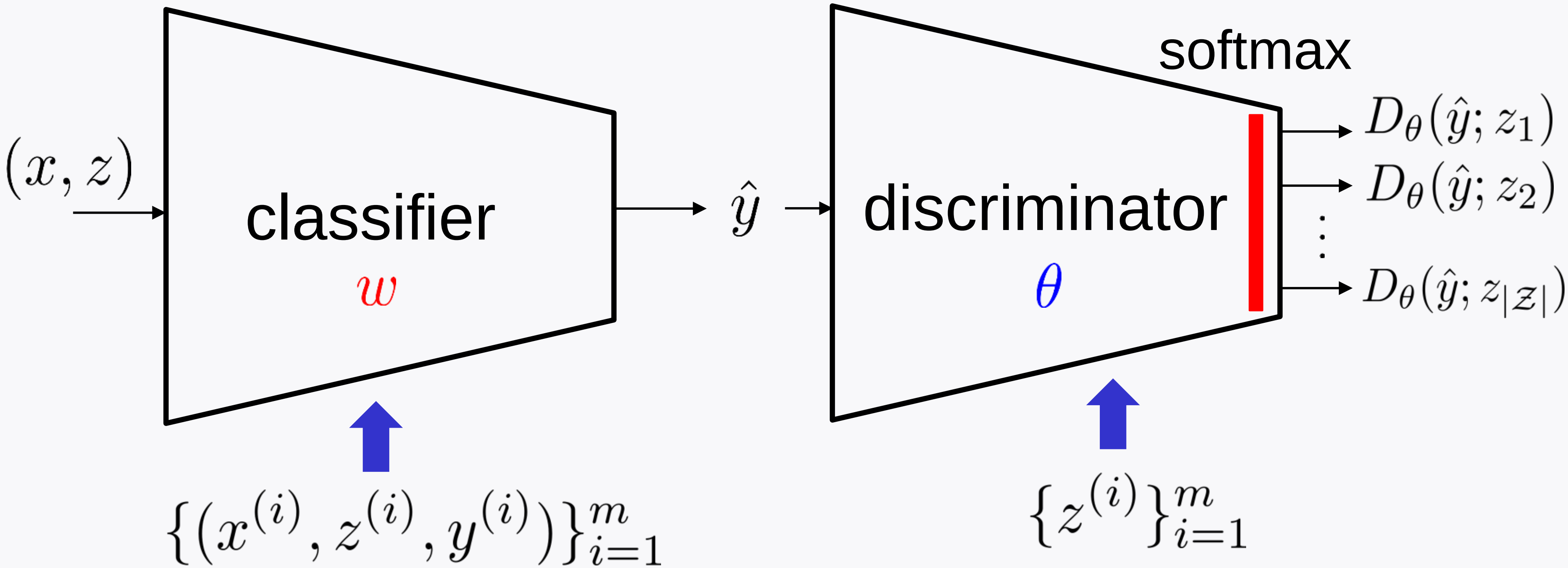
(iii) iterate this process until converge.

Architecture

$$\min_{w} \max_{\theta:\sum_z D_\theta(\hat{y};z)=1} \frac{1}{m}\left\{\sum_{i=1}^{m}(1-\lambda)\ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) + \lambda\sum_{i=1}^{m}\log D_\theta(\hat{y}^{(i)};z^{(i)})\right\}$$



Interpretation on $D_\theta(\hat{y}; z)$



Observe: Discriminator wishes to maximize $D_\theta(\hat{y}^{(i)}; z^{(i)})$, while classifier wishes to minimize.

Can interpret $D_\theta(\hat{y}; z)$ as the ability to figure out z from $\hat{y}$.

Analogy with GAN

Goodfellow et al. NeurIPS14

MI-based fair classifier	**GAN**
discriminator	discriminator
Figure out sensitive attribute from prediction	**Goal:** Distinguish real samples from fake ones.
classifier	generator
Decrease the ability to figure out senstivie attribute for the purpose of fairness.	Generate realistic fake samples

Extension to another fairness measure DEO

Connection:

$$\text{DEO} = 0 : \tilde{Y} \perp Z \,|Y \quad \Longleftarrow \quad I(Z; \hat{Y}|Y) = 0$$

Implementable optimization:

$$\min_{w} \max_{\theta : \sum_z D_\theta(\hat{y}; z, y) = 1} \frac{1}{m} \left\{ \sum_{i=1}^{m} (1-\lambda) \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)}; z^{(i)}, y^{(i)}) \right\}$$

Experiments

A benchmark real dataset: **COMPAS** Angwin et al. '15

$$(x, z, y)$$

criminal records

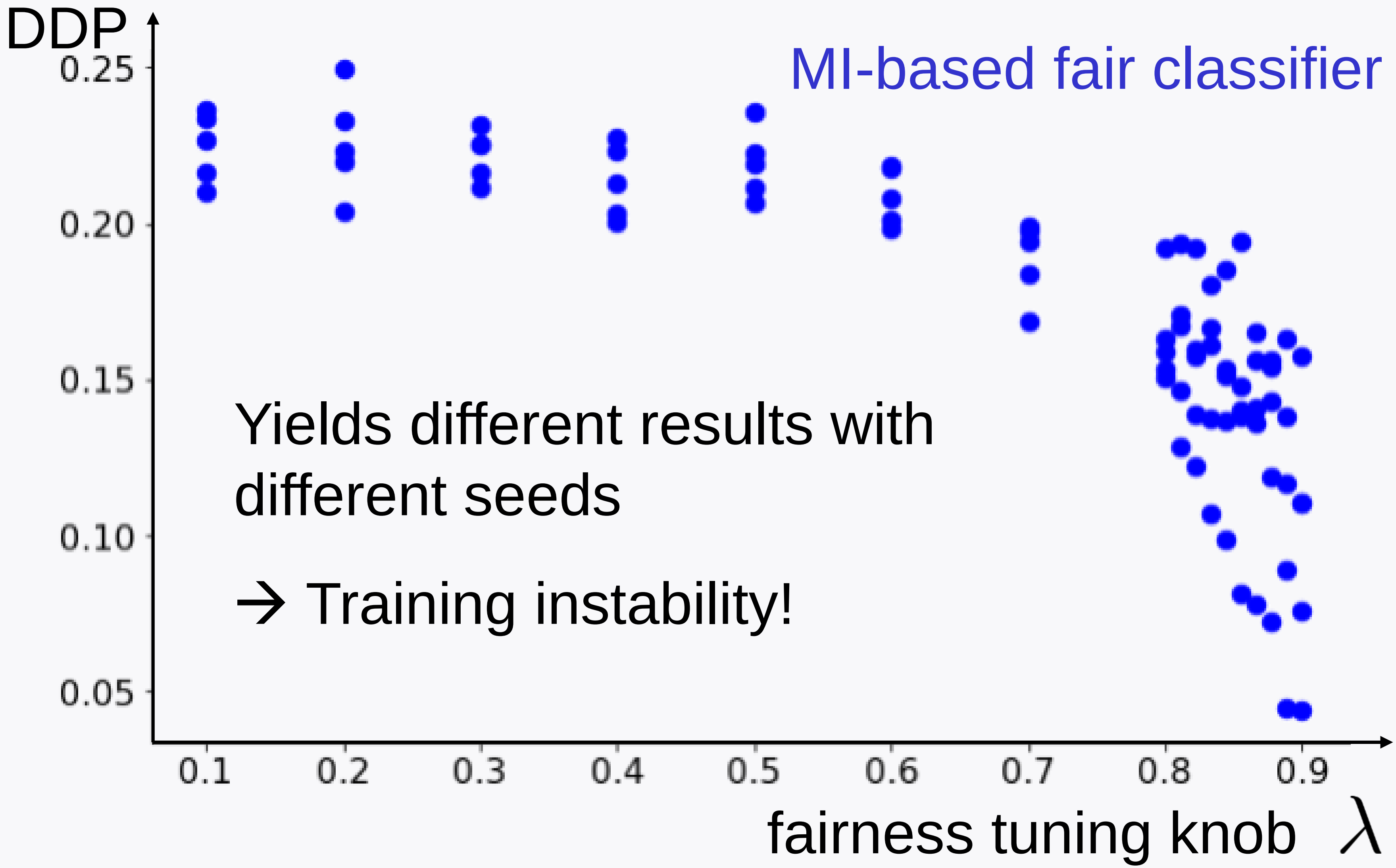
black or white

reoffend or not
in near future

Accuracy vs DDP tradeoff

	Accuracy	**DDP**
Non-fair classifier	68.29 ± 0.44	0.2263 ± 0.0087
MI-based *fair* classifier	67.07 ± 0.47	0.0997 ± 0.0426

A challenge



Yields different results with different seeds

→ Training instability!

Recent work

Cho-Hwang-Suh NeurIPS20

Another fair classifier *resolves the training instability* while offering a better tradeoff.

It is based on a well-known statistical method that often arises in information theory:

Kernel Density Estimation (KDE)

Look ahead

Explore the KDE-based fair classifier.

Reference

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[2] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics Conference (AISTATS)*, 2017.

[3] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *In Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

[4] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *In Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[5] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2018.

Reference

[6] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *In Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

[7] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification*. In Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[8] Y. Roh, K. Lee, S. E. Whang and C. Suh. FairBatch: Batch selection for model fairness. *International Conference on Learning Representations* (*ICLR),* 2020.

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

[10] 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.

Reference

[11] J. Cho, G. Hwang and C. Suh. A fair classifier using kernel density estimation. *In Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[12] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn. Renyi fair inference. *International Conference on Learning Representations* (*ICLR),* 2020.

[13] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein Fair Classification. *In Proceedings of the 35th Uncertainty in Artificial Intelligence Conference* (*UAI),* 2020.

[14] J. Lee, Y. Bu, P. Sattigeri, R. Panda, G. Wornell, L. Karlinsky, and R. Feris. A maximal correlation approach to imposing fairness in machine learning. *arXiv:2012.15259*, 2020.

[15] H. Jiang. Uniform convergence rates for kernel density estimation. *International Conference on Machine Learning (ICML)*, 2017.

[16] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to 272 predict future criminals. And it's biased against blacks. *https://www.propublica.org/article/machine-bias-risk-assessments-incriminal-sentencing*, 2015.