2021 IEEE East Asian School of Information Theory

Fair machine learning

Lecture 2

Changho Suh
EE, KAIST

Aug. 3, 2021

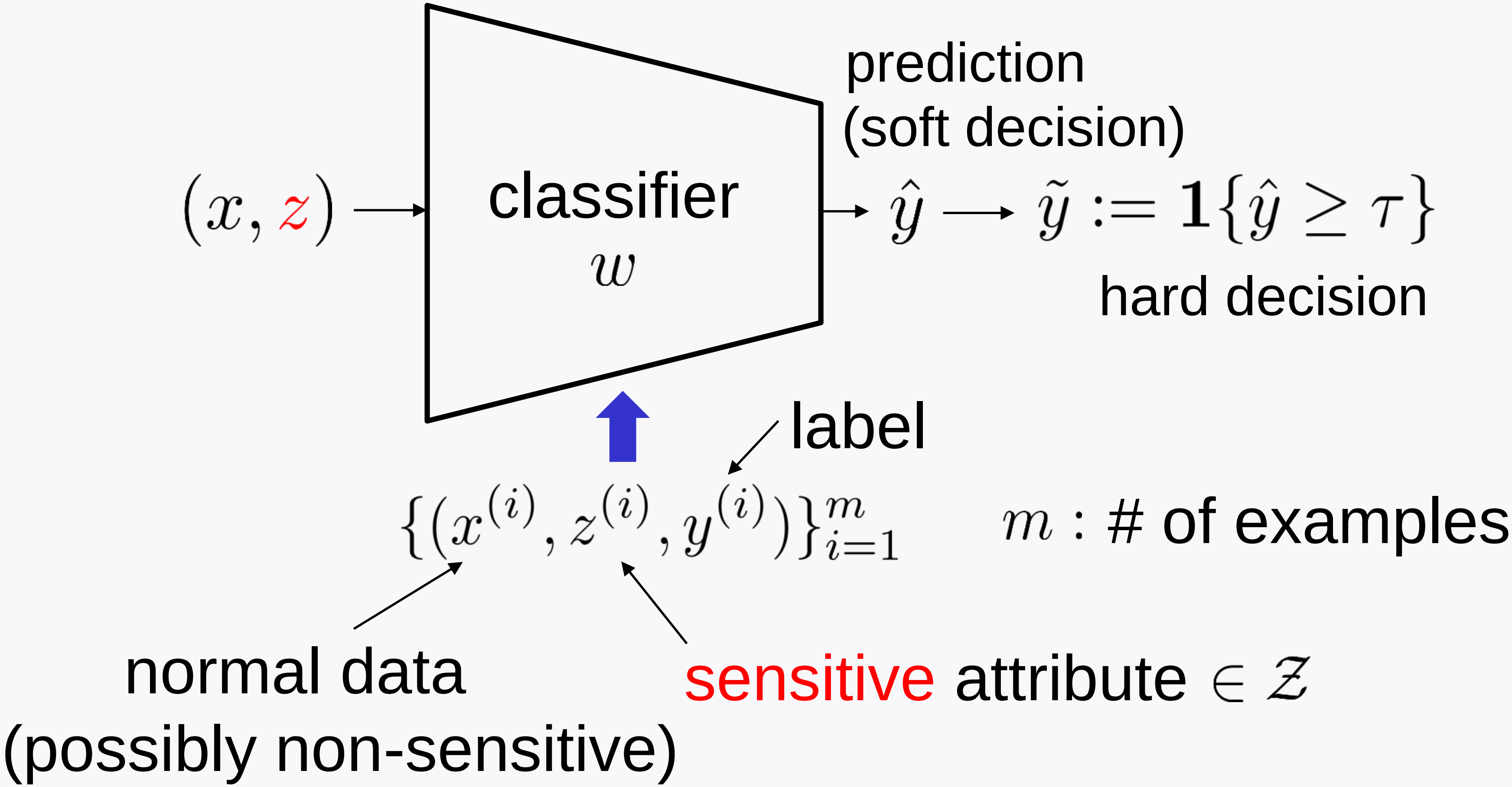
A fair classifier
using mutual information

Reading: TN2

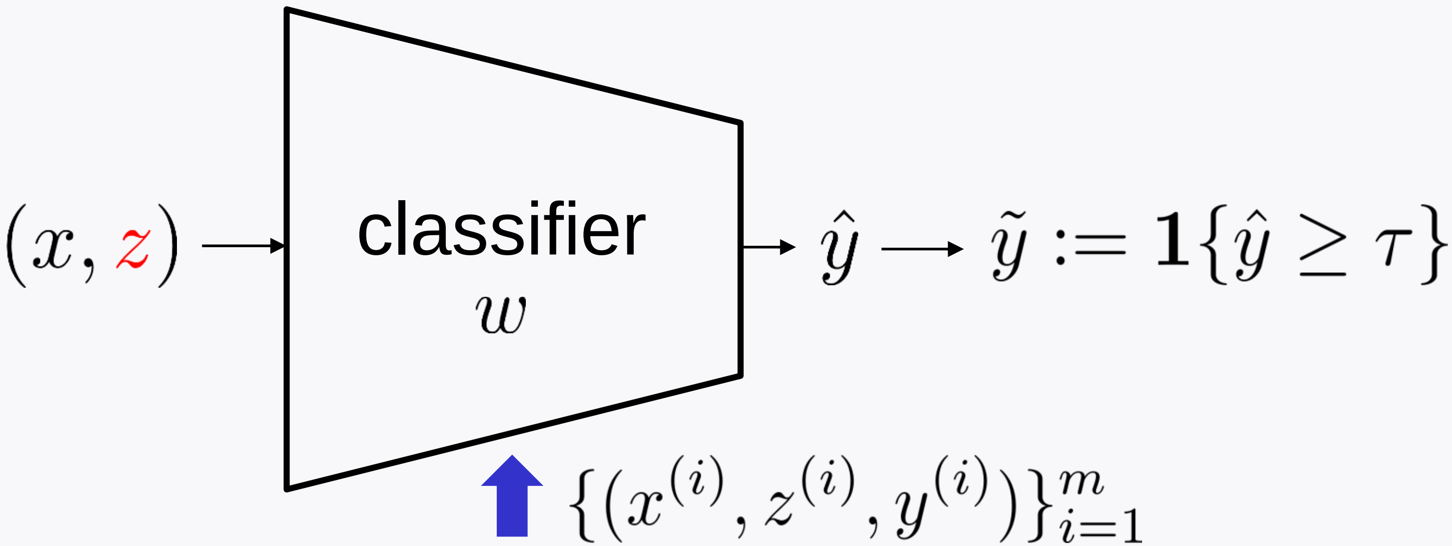
Outline of Lecture 2: Fair classifier #1

1. Introduce a problem setting.
2. Introduce an optimization framework.
3. Establish a connection btw fairness measure and mutual information (MI).
4. Develop an MI-based optimization for a fair classifier.
5. Traslate it into an implementable optimization, thereby coming up with a concrete way to solve the optimization.

Problem setting



Problem setting



For illustrative purpose, this tutorial focuses on:

(i) binary classifier &

(ii) one fairness measure:

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

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?

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

Observation: The smaller DDP, the more fair.

Enforcing fairness via regularization

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

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

Turns out: 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:

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

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

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

MI-based approach

Connection:

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

irrelevant of (θ, w)

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:

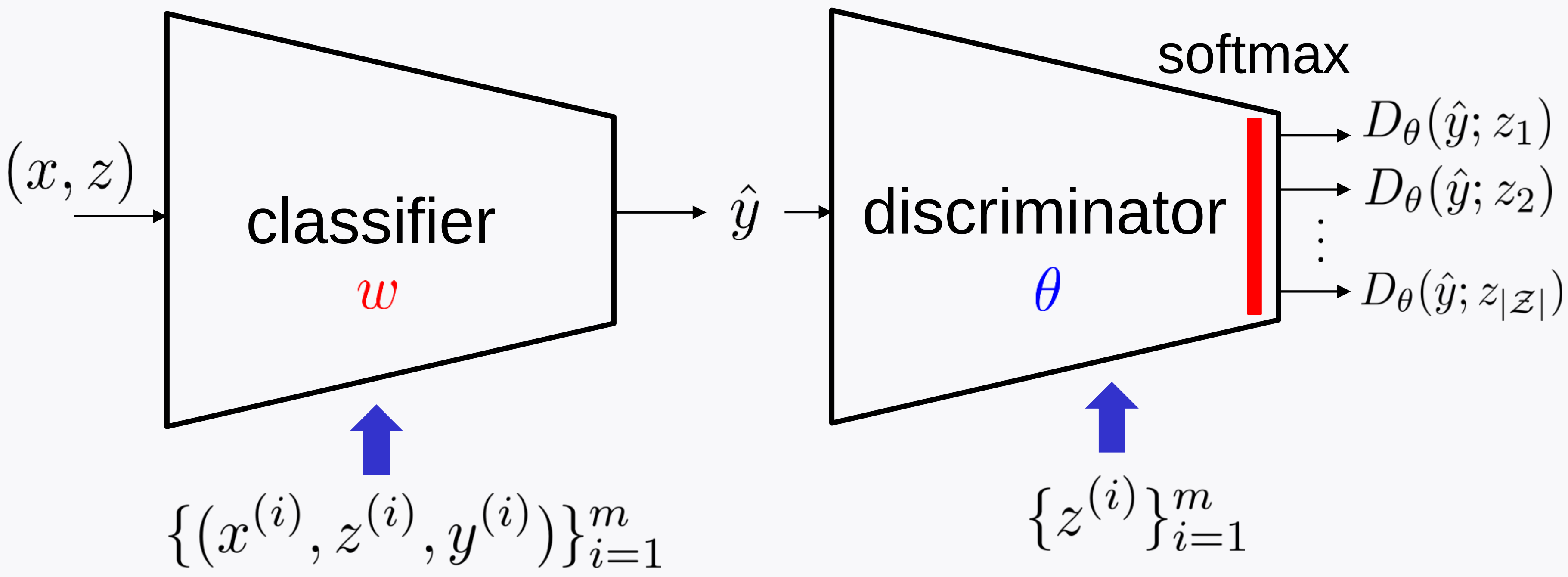
(i) Given w, update θ via the inner opt;

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

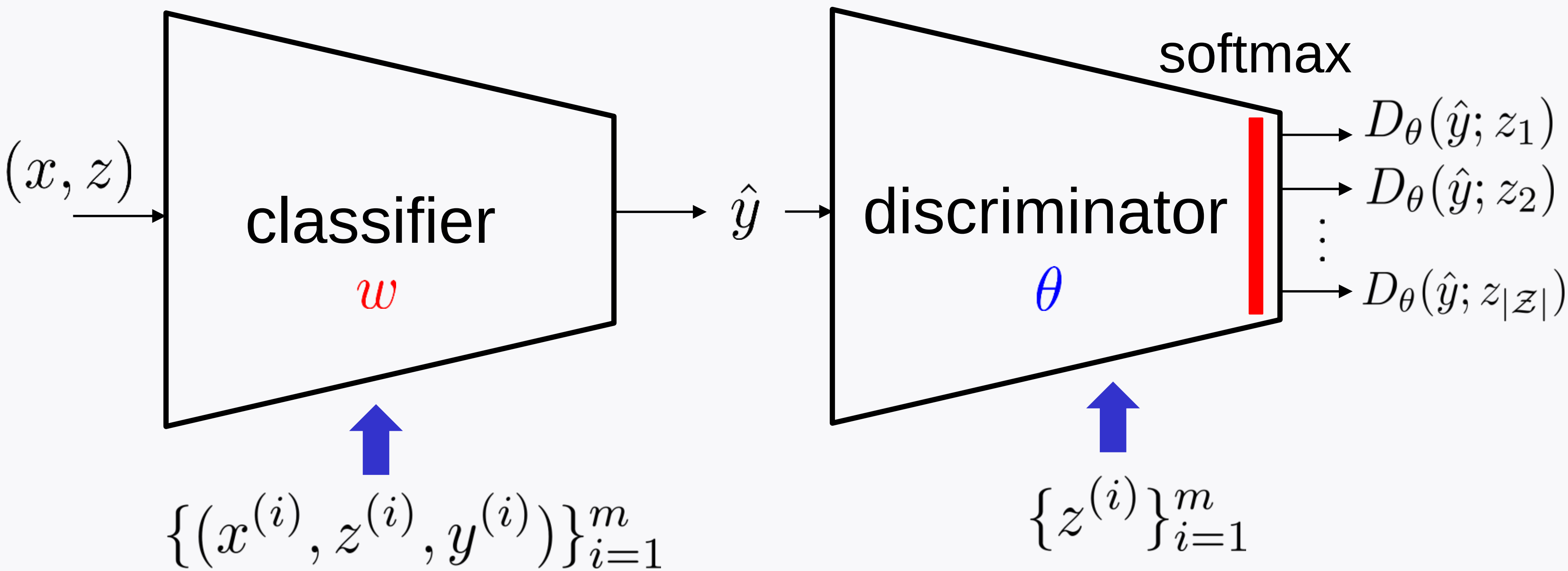
(iii) iterate this process until converge.

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 fiture 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 → x

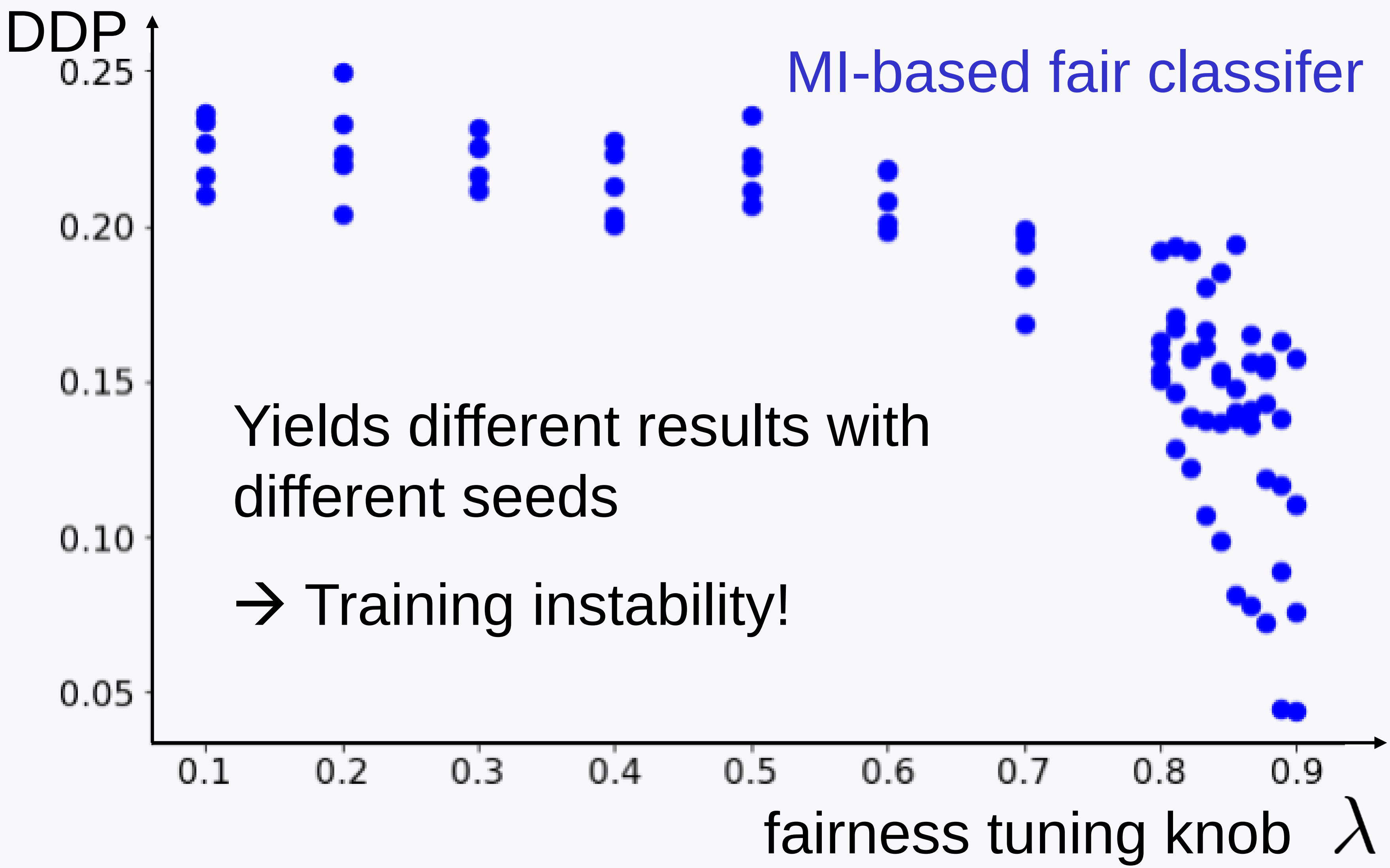
black or white → z

reoffend or not in near future → y

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

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

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

Kernel Density Estimation (KDE)

Look ahead

Explore the KDE-based fair classifier.

Reference

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

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

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *In Advances in Neural Information Processing Systems 27 (NeurIPS)*, 2014.

[5] 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.