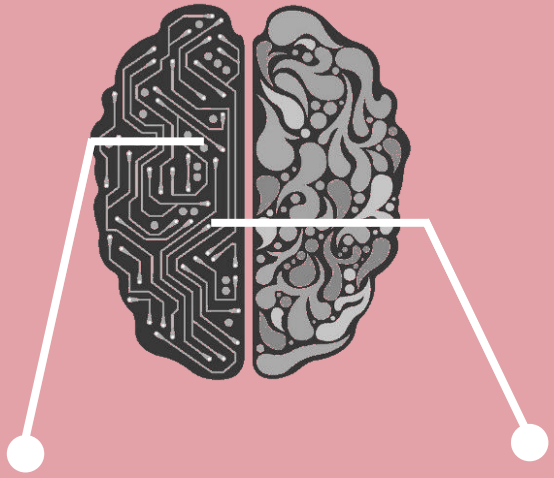
2021 KICS AI and Communications Workshop

Fair machine learning

Changho Suh
EE, KAIST

Aug. 13, 2021

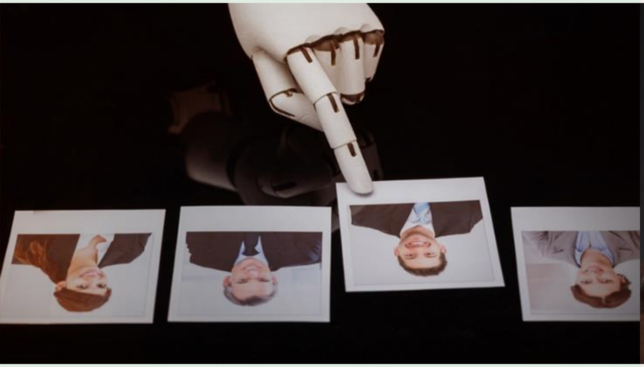
AI



google assistant

self driving

recruiting



judgement

loan decision

Trustworthy AI

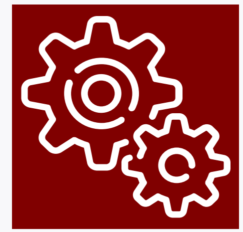
"AI has significant potential to help solve challenging problems, including by advancing medicine, understanding language, and fueling scientific discovery. ***To realize that potential****, it's critical that* ***AI*** *is used and developed* ***responsibly."***

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

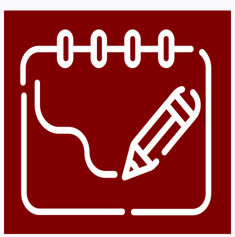
Five aspects of trustworthy AI

focus of this talk

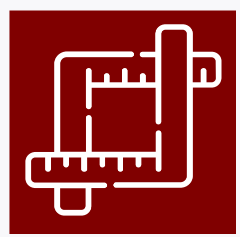
fairness



robustness



explainability



value alignment

transparency

A machine learning model of this talk's focus

Will explore fairness issues in the context of **classifiers.**

Role: Make fair prediction.

Fairness in the context of classifers?

There are many fairness concepts.

One important concept is **group fairness**:

Predictions should exhibit similar statistics regardless of sensitive attributes of groups

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

Applications of fair classifiers

job hiring

parole decision (가석방판결)

Applicants want no discrimination depending on race or sex.

A fair predictor for recidivism score (재범위험도) 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)

재범X 재범

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 distribution 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: $\tilde{Y} \perp Z \mid Y$

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

relevant to prediction accuracy

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

A quantified measure:

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

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]

focus of this talk

Why?

[Cho-Hwang-**Suh** NeurIPS20] focus of this talk

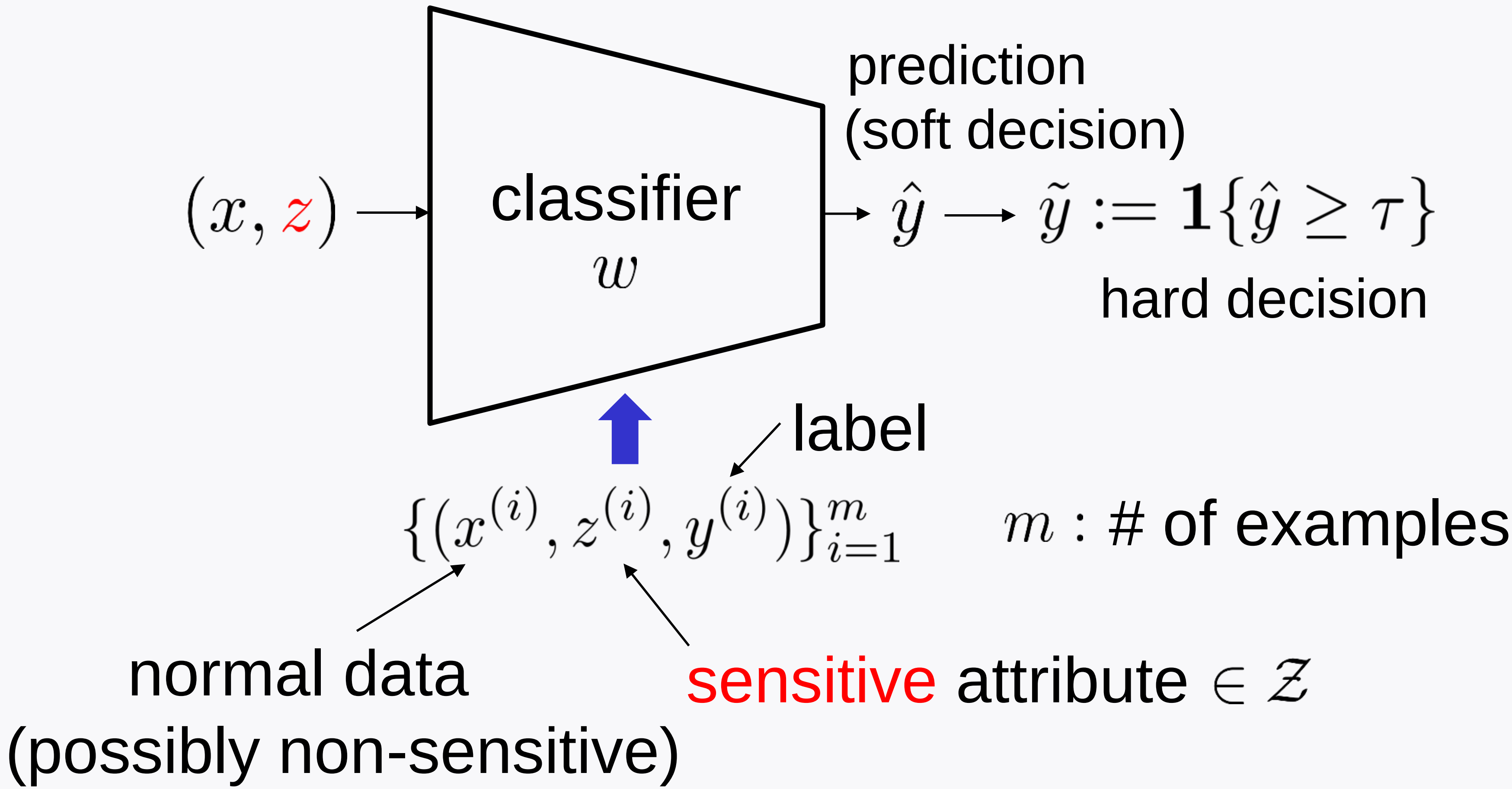
State of the art!

Employs a prominent statistical technique that is often employed in information theory and communication:

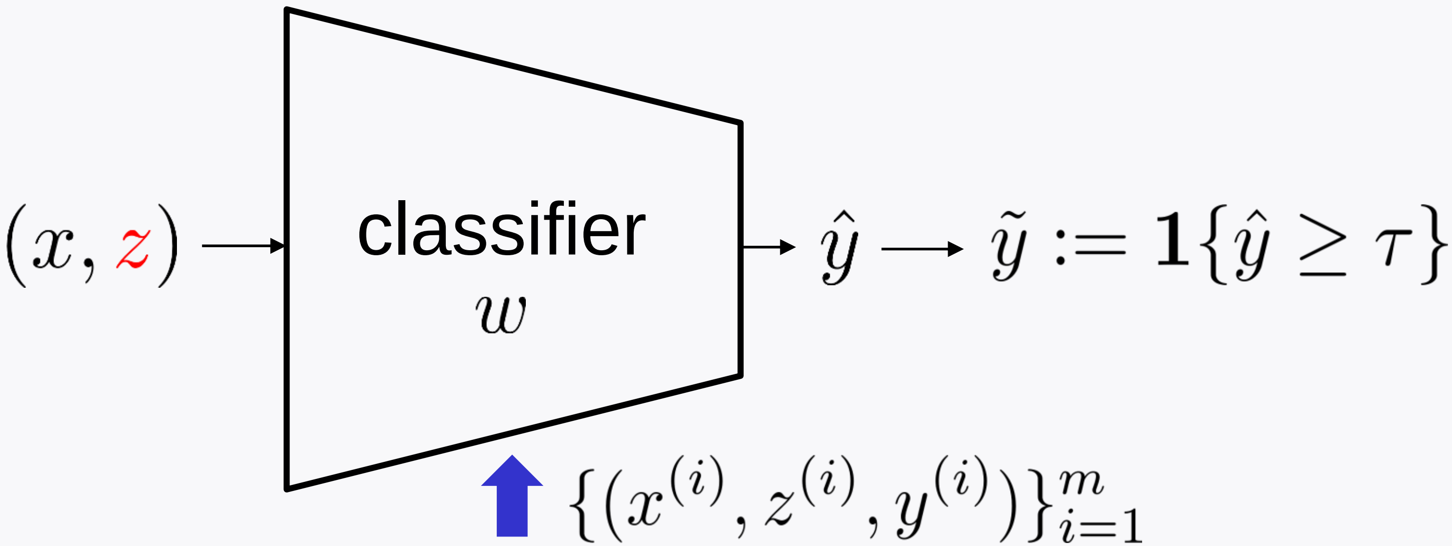
Kernel Density Estimation (KDE)

For the rest:
Will explore details on the KDE-based fair classifier.

Problem setting



Problem setting



For illustrative purpose, this talk 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.

Incorporating DDP as a regularization term

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

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

A challenge

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

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

$$\mathbb{P}(\tilde{Y} = 1) = \mathbb{P}(\hat{Y} \geq \tau) \qquad \tilde{Y} := \mathbf{1}\{\hat{Y} \geq \tau\}$$

$$= \int_{\tau}^{\infty} \underbrace{f_{\hat{Y}}(t)}_{\text{pdf}} dt$$

pdf uknown!

Instead: We are given samples $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(m)}\}$

Question: A way to infer the pdf from samples?

Kernel density estimation (KDE)

$$\mathbb{P}(\tilde{Y} = 1) = \int_{\tau}^{\infty} f_{\hat{Y}}(t)dt$$

Given samples $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(m)}\}$, KDE is defined as:

$$\widehat{f}_{\hat{Y}}(t) := \frac{1}{mh} \sum_{i=1}^{m} f_{\text{ker}}\left(\frac{t - \hat{y}^{(i)}}{h}\right)$$

a smoothing parameter (bandwidth)

a kernel function (e.g., Gaussian kernel)

Accuracy of KDE?

$$\mathbb{P}(\tilde{Y} = 1) = \int_{\tau}^{\infty} f_{\hat{Y}}(t)dt$$

Given samples $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(m)}\}$, KDE is defined as:

$$\widehat{f}_{\hat{Y}}(t) := \frac{1}{mh} \sum_{i=1}^{m} f_{\text{ker}} \left(\frac{t - \hat{y}^{(i)}}{h} \right)$$

Jiang ICML17: $|\widehat{f}(t) - f(t)|_{\infty} \lesssim \frac{1}{m^{\frac{1}{d}}}$ dim. of an interested r.v.

→ Yields an inaccurate estimate under **high-dim.** cases

Good news: In our setting, $d = 1$

Approximation via KDE

$$\mathbb{P}(\tilde{Y} = 1) = \int_{\tau}^{\infty} f_{\hat{Y}}(t)dt$$

$$\widehat{\mathbb{P}}(\tilde{Y} = 1) = \int_{\tau}^{\infty} \widehat{f}_{\hat{Y}}(t)dt$$

$$= \int_{\tau}^{\infty} \frac{1}{mh} \sum_{i=1}^{m} f_{\text{ker}} \left(\overbrace{\boxed{\frac{t - \hat{y}^{(i)}}{h}}}^{=: y} \right) dt$$

$$= \frac{1}{m} \sum_{i=1}^{m} \int_{\frac{\tau - \hat{y}^{(i)}}{h}}^{\infty} f_{\text{ker}}(y)dy$$

$$= \frac{1}{m} \sum_{i=1}^{m} Q\left(\frac{\tau - \hat{y}^{(i)}}{h} \right) \quad \text{(Gaussian kernel)}$$

Approximation via KDE

$$\widehat{\mathbb{P}}(\tilde{Y} = 1) = \frac{1}{m} \sum_{i=1}^{m} Q\left(\frac{\tau - \hat{y}^{(i)}}{h}\right)$$

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

Similarly, one can obtain:

$$\widehat{\mathbb{P}}(\tilde{Y} = 1 | Z = z) = \frac{1}{m_z} \sum_{i \in I_z} Q\left(\frac{\tau - \hat{y}^{(i)}}{h}\right)$$

$$|I_z| \qquad\qquad \{i : z^{(i)} = z\}$$

Approximated DDP

$$
\begin{aligned}
\text{DDP} &:= \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)| \\
&\approx \sum_{z \in \mathcal{Z}} |\widehat{\mathbb{P}}(\tilde{Y} = 1 | Z = z) - \widehat{\mathbb{P}}(\tilde{Y} = 1)| \\
&= \sum_{z \in \mathcal{Z}} \left| \frac{1}{m_z} \sum_{i \in I_z} Q\left(\frac{\tau - \hat{y}^{(i)}}{h}\right) - \frac{1}{m} \sum_{i=1}^{m} Q\left(\frac{\tau - \hat{y}^{(i)}}{h}\right) \right| \\
&\approx \sum_{z \in \mathcal{Z}} \left| \frac{1}{m_z} \sum_{i \in I_z} e^{-\frac{(\tau - \hat{y}^{(i)})^2}{2h^2}} - \frac{1}{m} \sum_{i=1}^{m} e^{-\frac{(\tau - \hat{y}^{(i)})^2}{2h^2}} \right|
\end{aligned}
$$

Can express DDP in terms of samples (thus *w*)

KDE-based optimization

Cho-Hwang-Suh NeurIPS20

$$\min_{w} \frac{1-\lambda}{m}\sum_{i=1}^{m}\ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) + \frac{\lambda}{m}\cdot\sum_{z\in\mathcal{Z}}\left|\frac{m}{m_z}\sum_{i\in I_z} e^{-\frac{(\tau-\hat{y}^{(i)})^2}{2h^2}} - \sum_{i=1}^{m} e^{-\frac{(\tau-\hat{y}^{(i)})^2}{2h^2}}\right|$$

Algorithm: Gradient descent

Issues: How to deal with the absolute function?

How to choose bandwidth *h*?

How to deal with the absolution func?

$$\min_{w} \frac{1-\lambda}{m} \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \frac{\lambda}{m} \cdot \sum_{z \in \mathcal{Z}} \left| \frac{m}{m_z} \sum_{i \in I_z} e^{-\frac{(\tau - \hat{y}^{(i)})^2}{2h^2}} - \sum_{i=1}^{m} e^{-\frac{(\tau - \hat{y}^{(i)})^2}{2h^2}} \right|$$

Instead, one can employ Huber loss:

$$H_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le \delta \\ \delta\left(|x| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$$

This enables us to readily obtain gradient.

How to choose bandwidth *h*?

$$\min_{w} \frac{1-\lambda}{m}\sum_{i=1}^{m}\ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) + \frac{\lambda}{m}\cdot\sum_{z\in\mathcal{Z}} H_{\delta}\left(\frac{m}{m_z}\sum_{i\in I_z} e^{-\frac{(\tau-\hat{y}^{(i)})^2}{2h^2}} - \sum_{i=1}^{m} e^{-\frac{(\tau-\hat{y}^{(i)})^2}{2h^2}}\right)$$

Turns out:

There is a sweet spot for h w.r.t. mean square error of KDE estimate.

Advise us to find h^* that minimizes the MSE.

See [Cho-Hwang-Suh NeurIPS20] for details.

Extension to another fairness measure DEO

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

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

$$\approx \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} \left| \frac{1}{m_{yz}} \sum_{i\in I_{yz}} e^{-\frac{(\tau-\hat{y}^{(i)})^2}{2h^2}} - \frac{1}{m_y} \sum_{i\in I_y} e^{-\frac{(\tau-\hat{y}^{(i)})^2}{2h^2}} \right|$$

$|I_{yz}|$ $\qquad$ $\{i : y^{(i)} = y, z^{(i)} = z\}$

Experiments

A benmark real dataset: **COMPAS** Angwin et al. ‘15

$$(x, z, y)$$

criminal records

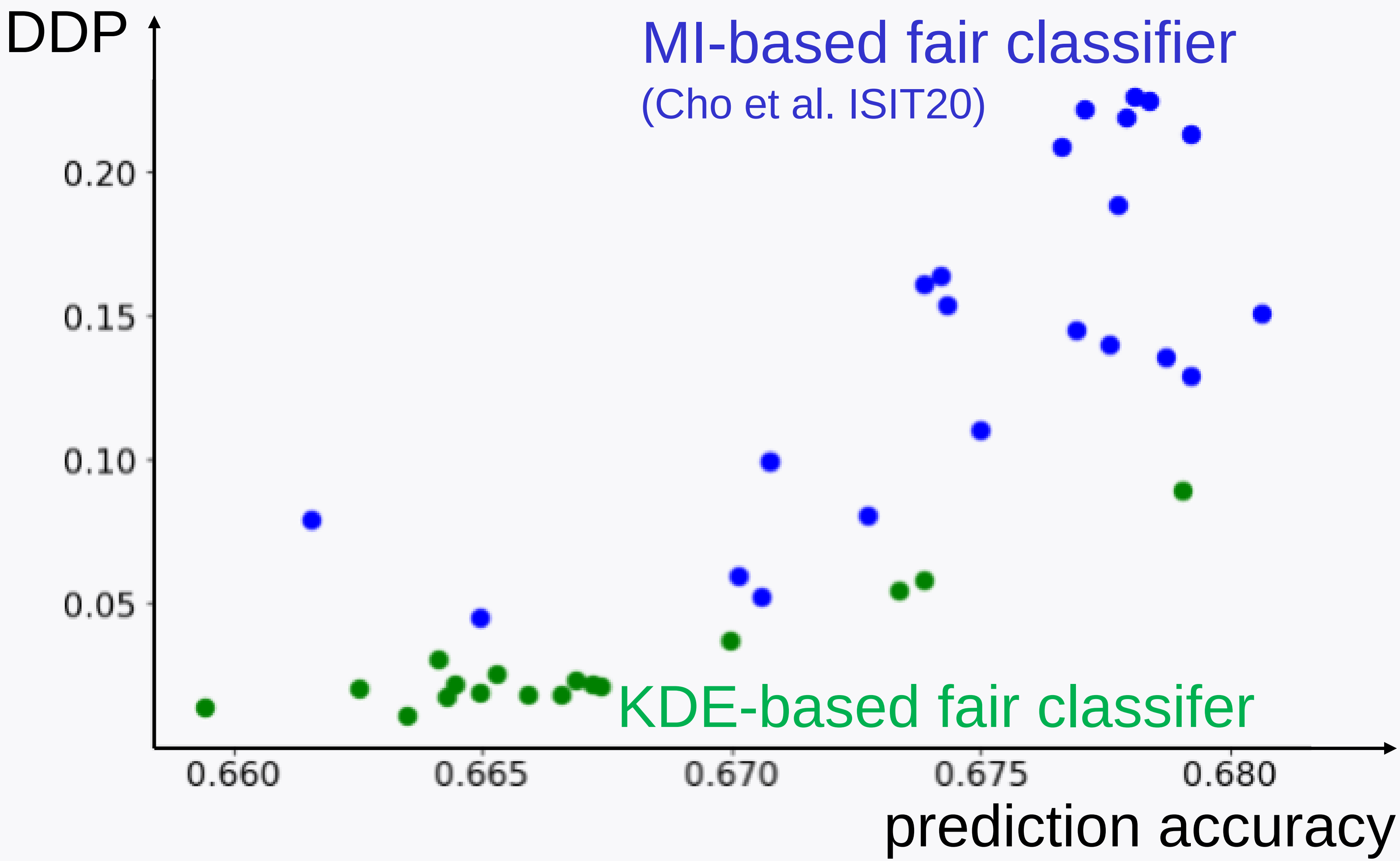
black or white

reoffend or not
in near future

Accuracy vs DDP tradeoff

	Accuracy	DDP
Non-fair classifier	68.29 ± 0.44	0.2263 ± 0.0087
KDE-based *fair* classifier	67.00 ± 0.45	0.0374 ± 0.0079
Mutual-Information-based fair classifier (Cho et al. ISIT20)	67.07 ± 0.47	0.0997 ± 0.0426

Accuracy vs DDP tradeoff



Conclusion

1. Explore fairness measures in fair classifiers.

2. Investigate the state-of-the-art fair classifer based on KDE, which performs well both in accuracy and fairness.

3. **Future direction:**
   Explore other aspects of trustworthy AI:
   robustness, explainability
   value alignment, transparency

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.