Fair machine learning

Lecture 2

Changho Suh
EE, KAIST

Aug. 25, 2021

A fair classifier using kernel density estimation

Reading: TN2

Recap: MI-based optimization

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

Mentioned: training instability.

Rationale behind training instability:

$I(Z; \hat{Y})$ = “max” optimization

“min-max” optimization often suffers from training instability.

Recap

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

Claimed: There is another fair classifier that addresses training instability while offering a better tradeoff.

Today's lecture

Will study the new fair classifier in depth.

1. Explore a way to directly compute the fairness measure DDP.

2. Introduce a trick that allows us to well approximate DDP:

   Kernel Density Estimation (KDE)

3. Develop a KDE-based optimization for a fair classifier.

4. Study how to solve the optimization.

Revisit: the fairness measure DDP

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

Let's try to compute this directly.

First focus on:

$$\mathbb{P}(\tilde{Y} = 1) = \mathbb{P}(\hat{Y} \geq \tau) = \int_{\tau}^{\infty} \underbrace{f_{\hat{Y}}(t)}_{\text{pdf}} dt \qquad \tilde{Y} := \mathbf{1}\{\hat{Y} \geq \tau\}$$

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) → h

a kernel function (e.g., Gaussian kernel) → f_{ker}

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

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

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

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

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

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

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

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

Approximated DDP

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

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:

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

This enables us to readily obtain gradient.

How to choose bandwidth *h*?

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

Turns out:

There is a sweet spot for h that miminizes the 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

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

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

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

m_{yz}: $|I_{yz}|$

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

Experiments

A benmark real dataset: **COMPAS**

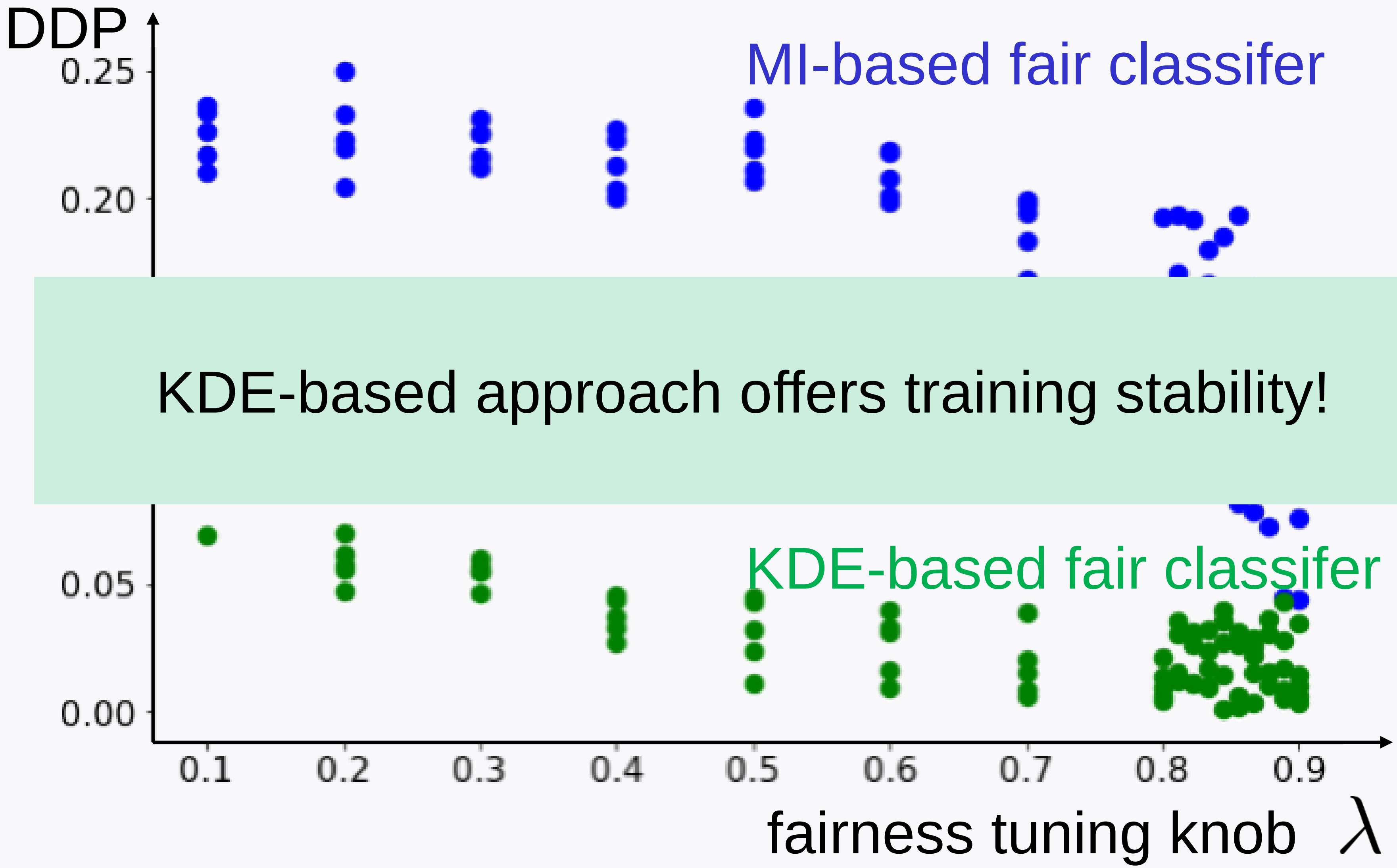
$$(x, z, y)$$

criminal records

black or white

reoffend or not in near future

Trainining instability?

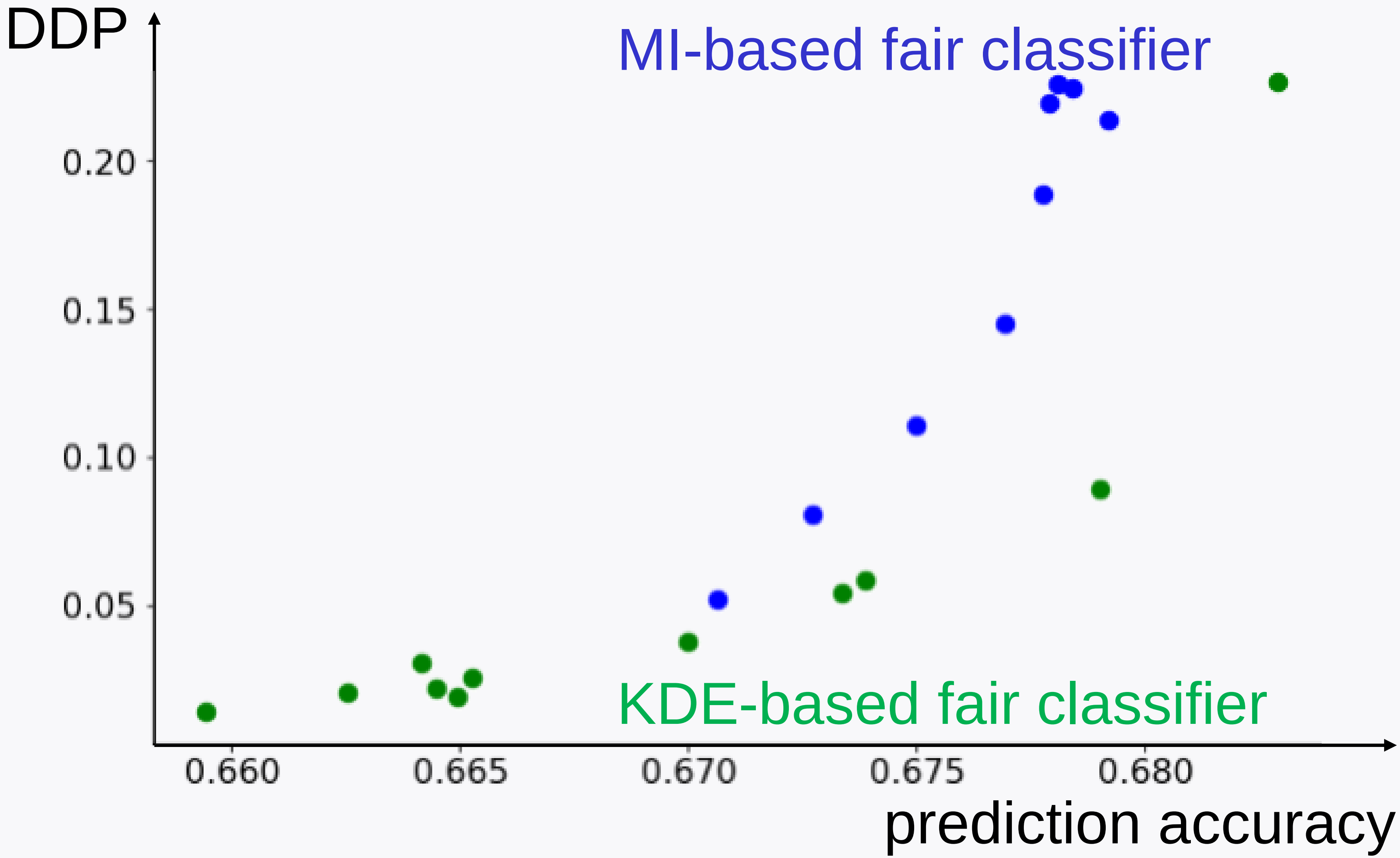


KDE-based approach offers training stability!

Accuracy vs DDP tradeoff

	Accuracy	**DDP**
Non-fair classifier	68.29 ± 0.44	0.2263 ± 0.0087
MI-based fair classifier	67.07 ± 0.85	0.0522 ± 0.0373
KDE-based fair classifier	67.00 ± 0.45	0.0374 ± 0.0079

Accuracy vs DDP tradeoff



Summary of Lectures 1 and 2

1. Explore fairness measures in fair classifiers.

2. Study an MI-based fair classifier which yields a good tradeoff while suffering from training instability.

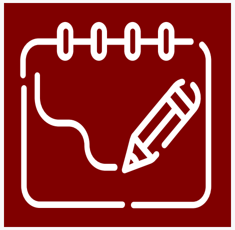
3. Investigate another fair classifer based on KDE, which addresses the training instability issue.

Revisit: Five aspects for trustworthy AI

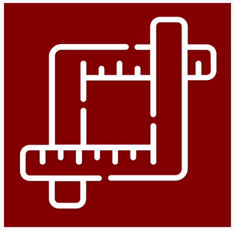
A recent progress: Roh-Lee-Whang-Suh, ICML20

fairness

robustness



explainability



value alignment

transparency

Look ahead

Will explore the recent work on fairness & robustness, and discuss relative issues.

Reference

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

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

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

[4] 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.