Croucher Summer Course in Information Theory
CUHK, August 23 ∼ 27, 2021

August 23, 2021
Changho Suh (chsuh@kaist.ac.kr)

Lecture 1: Overview & a fair classifier using mutual information

AI is prevalent

This tutorial touches upon a role of information theory and statistics in the trending field of AI. As AI becomes prevalent in our daily lives, we anticipate AI can play a significant role in a widening array of domains ranging from emerging killer applications such as AI assistant and self driving, to sensitive human-right-concerned applications like job hiring, judgement and loan decision.

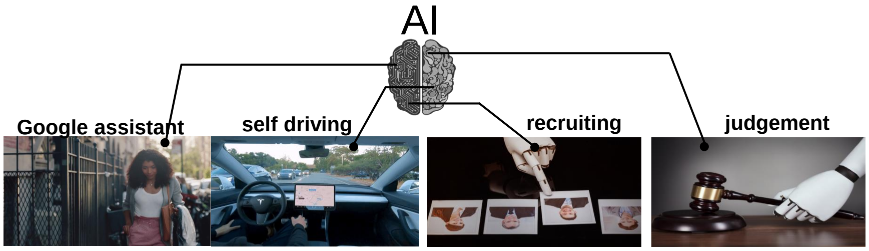


Figure 1: AI plays a powerful role in many applications.

Trustworthy AI

As AI becomes more and more powerful, one critical aspect that people wish to equip AI systems with is *trustworthiness*. To this end, major IT companies such as Google and IBM set out some promising directions towards trustworthy AI. Google targets *responsibility* for AI systems.

(Google): "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."

IBM pursues a new design paradigm centered around trustworthy AI.

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

There are five aspects that people take into account for enabling trustworthy AI. See Fig. 2. The first is *fairness*, which aims to design a model that does not discriminate among different demographics and/or individuals. The second is *robustness.* We desire to protect against noisy and possibly adversarial data. The third is *explainability.* A trained model should be explainable and interpretable so that people can readily be convinced by model's decision. The fourth is *value alignment*, meaning that a decision based on model's output should be aligned with actually what people want in reality. The last is *transparency.* A model should be developed in a transparent manner, being possibly be open to public. Obviously it is not that simple to satisfy all of these requirements. Recently, significant ongoing efforts have been made towards achieving the five

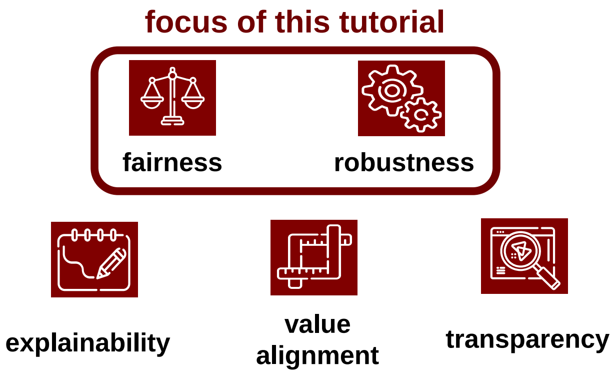


Figure 2: Five requirements for enabling trustworthy AI: (i) *fairness* across different demographics and/or individuals; (ii) *robustness* to data poisoning; (iii) *explanability* of trained models; (iv) *alignment* of model's output with actually what people want in reality; and (v) *transparency* of model development.

aspects. This tutorial targets only two: the *fairness* and *robustness* topics which we have made some recent progress on via some tools of information theory and statistics.

Outline of this tutorial

Specifically we will explore fairness issues in the context of one prominent machine learning model that concerns a supervised learning setup. That is, *fair classifiers* which intend to make unbiased decisions in light of different groups and/or individuals. Specifically what we are going to cover are three-folded. In today's lecture (Lecture 1), we will first figure out what it means by fairness in the context of classifiers. We will then study one fair classifier using arguably the most powerful information-theoretic notion: *mutual information.* In Wednesday's lecture (Lecture 2), we will next investigate another fair classifier that is built upon a very well-known statistical method named *Kernel Density Estimation (KDE).* We will also emphasize that it offers a better accuracy-fairness tradeoff performance. The fairness performance metric will be defined shortly. Lastly in Friday's lecture (Lecture 3), we will explore another fair classifier also being *robust* to data poisoning.

Fairness in the context of classifiers

Let me first explain what it means by *fairness* in the context of classifiers. There are many fairness concepts that people have considered for classifiers. One prominent concept of this tutorial's focus is the so called *group fairness* [1]. It is about prediction outcomes. The group fairness pursues predictions to exhibit similar statistics regardless of sensitive attributes of individuals such as race, gender, age and religion. Why do we care about this? It is because there are many applications concerning such sensitive attributes. Two applications are highlighted in Fig. 3: (i) job hiring; (ii) parole decision. In these applications, fair classifiers serve to ensure fairness among different demographics.

Demographic Parity (DP)

One concrete fairness condition (in the realm of group fairness) that is very popular and therefore

[1] There are two other concepts extensively explored in the literature: (i) individual fairness (pursuing fairness in the level of individuals); and (ii) causality-based fairness (exploring the causal relationship between sensitive attributes). But we will not touch upon these in this tutorial.

job hiring
parole decision (假釋放判決)

Figure 3: Two important applications of fair classifiers: (i) job hiring in which applicants want no discrimination depending on their race and/or sex; (ii) parole decision for which a fair predictor of recidivism (reoffending) score can play a crucial role.

I would like to focus on is the so called Demographic Parity (DP) condition [1, 2]. Let me explain what it is in the context of the recidivism score prediction. Let Z be a sensitive attribute, say 0 for black and 1 for white. Let $\tilde{Y}$ be a prediction made in hard decision, e.g., $\tilde{Y} = 1$ (reoffending in the near future, say within two years) or 0 (not reoffending). The DP condition means the *independence* between prediction and sensitive attribute, $\tilde{Y} \perp Z$, formally stated as:

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

where $\mathcal{Z}$ denotes the alphabet set of Z; $\mathcal{Z} = \{0, 1\}$ in this example. There are many ways to quantify how well the DP condition is satisfied. One natural way that we will take here is to quantify the degree of fairness via the Difference between two interested probabilities that arise in the DP condition (1) (DDP for short):

$$\mathsf{DDP} := \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|. \tag{2}$$

Notice that the independence implies DDP $= 0$ and vice versa. Hence, the smaller DDP, the prediction $\tilde{Y}$ is more independent of Z, thereby representing a fairer scenario.

Equalized Odds (EO)

The DP condition might not be desirable when the ground-truth label statistics of the two groups are by far distinct with each other, i.e., $\mathbb{P}(Y = 1 | Z = 1) \gg \mathbb{P}(Y = 1 | Z = 0)$ or vice versa. In this case, the DP condition is far from the ground-truth label distribution, and therefore enforcing the DP condition may aggravate prediction accuracy significantly. This shortcoming motivated the use of the following condition, named *Equalized Odds* (EO), which pursues the *conditional independence*: $\tilde{Y} \perp Z | Y$, i.e.,

$$\mathbb{P}(\tilde{Y} = 1 | Y = y, Z = z) = \mathbb{P}(\tilde{Y} = 1 | Y = y), \qquad \forall z \in \mathcal{Z}, \forall y \in \mathcal{Y}. \tag{3}$$

Notice that $\mathbb{P}(\tilde{Y} = 1 | Y = y)$ is closely coupled with prediction accuracy, e.g., it reads the probability of being correct when $y = 1$. The EO condition somehow promotes the equalized prediction accuracies, and enforcing the EO condition actually has little to do with reducing accuracy in such asymmetric case $\mathbb{P}(Y = 1 | Z = 1) \gg \mathbb{P}(Y = 1 | Z = 0)$. So the EO condition is much more preferably employed in practice, although it is not guaranteed to be always strictly better than the DP condition. Similar to DDP, the EO condition can be quantified via the Difference between the two interested probabilities in the EO condition (3) (DEO for short):

$$\mathsf{DEO} := \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Y = y, Z = z) - \mathbb{P}(\tilde{Y} = 1 | Y = y)|. \tag{4}$$

On a side note: There is another fairness measure beyond DDP and DEO, which concerns precision quality and is defined below:

$$\sum_{\tilde{y}\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}|\mathbb{P}(Y=1|\tilde{Y}=\tilde{y},Z=z)-\mathbb{P}(Y=1|\tilde{Y}=\tilde{y})|. \tag{5}$$

The use of fairness measures depend on applications which might put an emphasis on accuracy or precision performance.

[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]

Figure 4: A partial list of references regarding fair classifiers.

Many recent works on fair classifiers

There has been a proliferation of fairness algorithms that intend to minimize DDP or DEO. Fig. 4 exhibits only a partial list of the relevant references. These are chronologically listed up, yet categorized into two columns. The references in the second column are the ones which are relevant to *information theory & statistics* of this audience's potential interest and hence I would like to put a particular emphasis on. Specifically Zafar et al. [2] employ a well-known statistical measure, called *Pearson correlation*, which also often arises in information theory. Baharlouei et al. [12] and Lee et al. [14] rely upon other prominent measures, Rényi correlation and HGR (Hirchfeld-Gebelein-Rényi) maximal correlation, respectively. Jiang et al. [13] employ the famous Wasserstein distance. Cho-Hwang-Suh [9] and Roh-Lee-Whang-Suh [10] employ arguably the most powerful and prominent information-theoretic measure, *mutual information.* There is another work [11] which exploits a well-known statistical method: *Kernel Density Estimation* (KDE).

Among these, we will focus on the following three works concerning mutual information and KDE: Cho-Hwang-Suh [9], Roh-Lee-Whang-Suh [10] and Cho-Hwang-Suh [11]. A couple of reasons why I made such a choice. The first and obvious reason is that I can teach them well, as I was involved in as a co-author. Second, the references [9, 10] concern the very famous *mutual information* that some of you guys are excited about and/or familiar with. Third, the last reference [11] proposes a simple yet powerful fair classifier which I believe is the state of the art.

Here are what we are going to cover in detail next. For the rest of this lecture, we will study an interesting connection between fairness measures (DDP and DEO) and mutual information (MI), and then will exploit the connection to investigate an MI-inspired fair classifier developed in [9]. In Lecture 2, we will explore the state of the art based on KDE [11]. In Lecture 3, we will study another fair classifier that is also robust to data poisoning [10].

Problem setting of a fair classifier

Fig. 5 illustrates the architecture of a conventional binary classifier. There are two types of

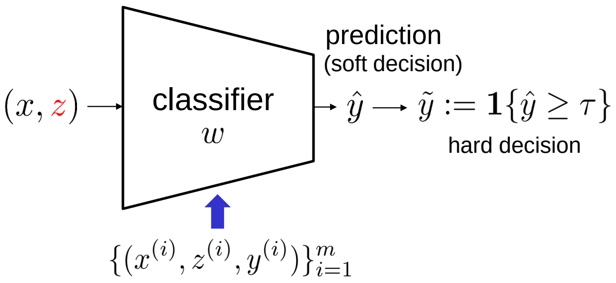

prediction
(soft decision)
(x, z)
classifier
w
$\hat{y}$
$\tilde{y} := \mathbf{1}\{\hat{y} \geq \tau\}$
hard decision
$\{(x^{(i)}, z^{(i)}, y^{(i)})\}_{i=1}^{m}$

Figure 5: A problem setting of a binary fair classifier. Here X denotes normal (possibly non-sensitive) data, $Z \in \mathcal{Z}$ indicates a sensitive attribute with arbitrary alphabet size, and Y is a binary label. Let $\hat{Y}$ be the prediction output that intends to learn the ground-truth conditional probability $\mathbb{P}(Y = 1|X = x, Z = z)$ and $\tilde{Y}$ be its hard-decision $\tilde{Y} := \mathbf{1}\{\hat{Y} \geq \tau\}$ where τ is a certain threshold. Here the classifier is parameterized by w.

data for input: (i) normal (possibly non-sensitive) data; (ii) sensitive attributes. We denote the normal data by X. In the case of recidivism score prediction, such X may refer to a collection of the number of prior criminal records and a criminal type, e.g., misdemeanour or felony. For sensitive data, we employ a different notation, say Z. In the above example, Z may indicate a race type among black ($Z = 0$) and white ($Z = 1$). In general, the alphabet size of Z is arbitrary. For instance, there are many race types such as Black, White, Asian, Hispanic, to name a few. Also there could be multiple sensitive attributes like gender and religion. In order to reflect such practically-relevant scenarios, we consider $Z \in \mathcal{Z}$ with an arbitrary alphabet size that can represent a collection of possibly many sensitive attributes. Let $\hat{Y}$ be the classifier output which aims to represent the ground-truth conditional distribution $\mathbb{P}(y|x, z) := \mathbb{P}(Y = y|X = x, Z = z)$. Here $Y \in \mathcal{Y}$ denotes the ground-truth label. In the recidivism score prediction, $Y = 1$ means reoffending in the near future, say within two years ($Y = 0$ otherwise), while $\hat{Y}$ indicates the probability of such event being occurred. Let $\tilde{Y}$ be its hard-decision $\tilde{Y} := \mathbf{1}\{\hat{Y} \geq \tau\}$ where τ is a certain threshold. Here the classifier is parameterized by w. We consider a supervised learning setup, so we are given m example triplets: $\{(x^{(i)}, z^{(i)}, y^{(i)})\}_{i=1}^{m}$.

For illustrative purpose, this tutorial focuses on the simple binary classification setting and one fairness measure DDP:

$$\mathsf{DDP} := \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1|Z = z) - \mathbb{P}(\tilde{Y} = 1)|. \tag{6}$$

Fairness-regularized optimization

A conventional classifier optimization often takes the following form:

$$\min \frac{1}{m} \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) \tag{7}$$

where $\ell_{\mathsf{CE}}(y, \hat{y})$ indicates cross entropy loss:

$$\ell_{\mathsf{CE}}(y, \hat{y}) := -y \log \hat{y} - (1 - y) \log(1 - \hat{y}). \tag{8}$$

How to incorporate the fairness measure DDP? Notice that the smaller DDP, the fairer situation.

Hence, one natural approach is to enforce fairness via regularization as below:

$$\min \frac{1-\lambda}{m} \sum_{i=1}^{m} \ell_{\sf CE}(y^{(i)}, \hat{y}^{(i)}) + \lambda \cdot {\sf DDP} \tag{9}$$

where λ denotes a regularization factor that lies in between 0 and 1. One can interpret λ as a fairness tuning knob. Here a challenge arises in solving the regularized optimization (9). Recalling the definition (6) of $\sf DDP$, we see that $\sf DDP$ is a complicated function of the optimization variable w. It turns out it is not that simple to express $\sf DDP$ in terms of w. One effort to address this challenge was made by Zafar et al. [2]. They introduce an easily-expressible *proxy* for the fairness measure. Specifically they employ a covariance function between $\hat{Y}$ and Z. However, this proxy serves only as a *weak* constraint because a small covariance does not necessarily imply the independence although the reverse always hold. In this tutorial, we will study another approach which introduces a different regularization term that can serve as a *strong* constraint for the independence.

Connection between $\sf DDP$ and mutual information

The approach is based on the popular information-theoretic measure: mutual information. To see its relevancy clearly, let us make a concrete connection. The connection is made via the following observation:

$${\sf DDP} = 0 : \tilde{Y} \perp Z \iff I(Z; \tilde{Y}) = 0. \tag{10}$$

This is because $I(Z; \tilde{Y}) = 0$ is the sufficient and necessary condition for the independence between Z and $\tilde{Y}$. The connection can also be made via the soft-decision prediction $\hat{Y}$. Notice that

$$I(Z; \tilde{Y}) \leq I(Z; \tilde{Y}, \hat{Y}) = I(Z; \hat{Y}) \tag{11}$$

where the 1st inequality comes from the chain rule $I(Z; \tilde{Y}, \hat{Y}) = I(Z; \tilde{Y}) + I(Z; \hat{Y}|\tilde{Y})$ and the non-negativity of mutual information; and the 2nd equality is due to the fact that $\tilde{Y}$ is a function of $\hat{Y}$ ($\tilde{Y} := \mathbf{1}\{\hat{Y} \geq \tau\}$) and hence $I(Z; \tilde{Y}|\hat{Y}) = 0$. This together with (10) yields:

$${\sf DDP} = 0 : \tilde{Y} \perp Z \impliedby I(Z; \hat{Y}) = 0. \tag{12}$$

We see that $I(Z; \hat{Y}) = 0$ can serve as a *strong* constraint for the independence.

MI-based approach [9]

The connection (12) naturally motivates us to employ $\lambda \cdot I(Z; \hat{Y})$ as a regularization term in (9) instead of $\lambda \cdot {\sf DDP}$:

$$\min_{w} \frac{1-\lambda}{m} \sum_{i=1}^{m} \ell_{\sf CE}(y^{(i)}, \hat{y}^{(i)}) + \lambda \cdot I(Z; \hat{Y}). \tag{13}$$

Now a question of interest is: How to express $I(Z; \hat{Y})$ in terms of the optimization variable w? It turns out there is an interesting way to do this. To figure out the way, let us massage $I(Z; \hat{Y})$ to arrive at the expression.

A careful look at mutual information

Starting with the definition of mutual information, we get:

$$\begin{aligned}
I(Z;\hat{Y}) &= H(Z) - H(Z|\hat{Y}) \\
&\overset{(a)}{=} H(Z) - (H(\hat{Y},Z) - H(\hat{Y})) \\
&\overset{(b)}{=} H(Z) - \mathbb{E}\left[\log \frac{1}{\mathbb{P}_{\hat{Y},Z}(\hat{Y},Z)}\right] + \mathbb{E}\left[\log \frac{1}{\mathbb{P}_{\hat{Y}}(\hat{Y})}\right] \\
&= H(Z) + \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})}
\end{aligned} \tag{14}$$

where (a) comes from the chain rule $H(\hat{Y},Z) = H(\hat{Y}) + H(Z|\hat{Y})$; and (b) is due to the definitions of entropy and joint entropy. Define the term placed in the last line marked in blue as:

$$D^*(\hat{y};z) := \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})}. \tag{15}$$

Due to the total probability law, $D^*(\hat{y};z)$ should respect the sum-up-to-one constraint w.r.t. z:

$$\sum_z D^*(\hat{y};z) = 1 \quad \forall \hat{y}. \tag{16}$$

Mutual information via function optimization

Instead of $D^*(\hat{y};z)$, one can think about another function, say $D(\hat{y};z)$, which respects only the sum-up-to-one constraint (16). It turns out $D^*(\hat{y};z)$ is the optimal choice among such $D(\hat{y};z)$ in a sense of maximizing:

$$\sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z), \tag{17}$$

and this gives insights into expressing $I(Z;\hat{Y})$ in terms of w. To see this clearly, let me formally state that $D^*(\hat{y};z)$ is indeed the maximizer via the following theorem.

Theorem: The mutual information $I(Z;\hat{Y})$, expressed as in the last line of (14), can be represented as the following *function optimization*:

$$I(Z;\hat{Y}) = H(Z) + \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z). \tag{18}$$

The proof of this is simple. Notice that the optimization (18) is *convex* in $D(\cdot,\cdot)$, since the log function is concave and the convexity preserves under addition. Hence, by checking the KKT condition (the optimality condition for convex optimization), one can prove that the optimal $D(\cdot,\cdot)$ indeed respects (15) and (16). Here is detail. Consider the Lagrange function:

$$\mathcal{L}(D(\hat{y};z),\nu(\hat{y})) = \sum_{\hat{y},z} \mathbb{P}_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z) + \sum_{\hat{y}} \nu(\hat{y}) \left(1 - \sum_z D(\hat{y};z)\right) \tag{19}$$

where $\nu(\hat{y})$'s indicate Lagrange multipliers w.r.t. the equality constraints. Consider the KKT conditions:

$$\left.\frac{d\mathcal{L}(D(\hat{y};z),\nu(\hat{y}))}{dD(\hat{y};z)}\right|_{D=D_{\mathsf{opt}},\nu=\nu_{\mathsf{opt}}} = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{D_{\mathsf{opt}}(\hat{y};z)} - \nu_{\mathsf{opt}}(\hat{y}) = 0; \tag{20}$$

$$\sum_z D_{\mathsf{opt}}(\hat{y};z) = 1. \tag{21}$$

So we get $D_{\mathsf{opt}}(\hat{y};z) = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\nu_{\mathsf{opt}}(\hat{y})}$. Plugging this into (21), we obtain:

$$\sum_z D_{\mathsf{opt}}(\hat{y};z) = \frac{\sum_z \mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\nu_{\mathsf{opt}}(\hat{y})} = 1, \tag{22}$$

which together with total probability law yields:

$$\nu_{\mathsf{opt}}(\hat{y}) = \sum_z \mathbb{P}_{\hat{Y},Z}(\hat{y},z) = \mathbb{P}_{\hat{Y}}(\hat{y}). \tag{23}$$

This together with (20) then gives:

$$D_{\mathsf{opt}}(\hat{y};z) = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\nu_{\mathsf{opt}}(\hat{y})} = \frac{\mathbb{P}_{\hat{Y},Z}(\hat{y},z)}{\mathbb{P}_{\hat{Y}}(\hat{y})} = D^*(\hat{y};z). \tag{24}$$

This completes the proof of the theorem.

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

Are we done with expressing $I(Z;\hat{Y})$ in terms of w? No. This is because $P_{\hat{Y},Z}(\hat{y},z)$ that appears in (18) is not available. To resolve this issue, we rely upon the empirical distribution instead:

$$\mathbb{Q}_{\hat{Y},Z}(\hat{y}^{(i)},z^{(i)}) = \frac{1}{m} \qquad \forall i \in \{1,\ldots,m\}.$$

In practice, the empirical distribution is very likely to be uniform, since $\hat{y}^{(i)}$ is real-valued and hence the pair $(\hat{y}^{(i)},z^{(i)})$ is unique with high probability. Now by parametrizing the function $D(\cdot,\cdot)$ with another, say θ, we can approximate $I(Z;\hat{Y})$ as:

$$I(Z;\hat{Y}) \approx H(Z) + \max_{\theta:\sum_z D_\theta(\hat{y};z)=1} \sum_{i=1}^{m} \frac{1}{m} \log D_\theta(\hat{y}^{(i)};z^{(i)}). \tag{25}$$

From the above parameterization building upon the function optimization (18), we can now approximately express $I(Z;\hat{Y})$ in terms of (w,θ).

Implementable optimization

Notice in (25) that $H(Z)$ is irrelevant to the introduced optimization variables (w,θ). Hence, the MI-based optimization (13) can be (approximately) translated into:

$$\min_w \max_{\theta:\sum_z D_\theta(\hat{y};z)=1} \frac{1}{m}\left\{\sum_{i=1}^{m}(1-\lambda)\ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) + \lambda \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)};z^{(i)})\right\}. \tag{26}$$

The objective function is a function of (w,θ) and hence it is implementable, for instance, via famous neural networks. Many of the neural-net-based optimizations can readily be solved via a family of gradient descent algorithms. But here we see "min max". Hence, we can apply a slight variant of gradient descent that people often call *alternating gradient descent*, in which given w, θ is updated via the inner optimization and then given the updated θ, w is newly updated via the outer optimization, and this process iterates until it converges.

The architecture of the MI-based optimization (26) is illustrated in Fig. 6. On top of a classifier, we introduce a new entity, called *discriminator*, which corresponds to the inner optimization. In discriminator, we wish to find θ^* that maximizes $\frac{1}{m}\sum_{i=1}^{m}\log D_\theta(\hat{y}^{(i)};z^{(i)})$. On the other hand, the classifier wants to *minimize* such term. Hence, $D_\theta(\hat{y};z)$ can be viewed as the ability to

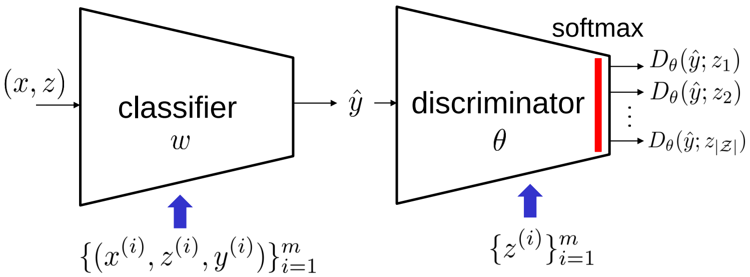
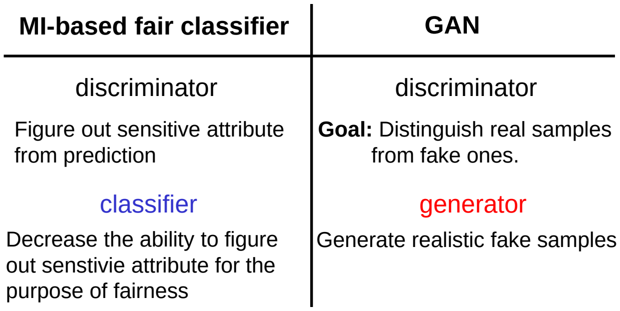


Figure 6: The architecture of the MI-based fair classifier (26). The prediction output $\hat{y}$ is fed into the discriminator wherein the goal is to figure out sensitive attribute z from $\hat{y}$. The discriminator output $D_\theta(\hat{y};z)$ can be interpreted as the probability that $\hat{y}$ belongs to the attribute z. Here the softmax function is applied to ensure the sum-up-to-one constraint (16).

figure out z from prediction $\hat{y}$. Notice that the classifier wishes to minimize such ability for the purpose of fairness, while the discriminator has the opposite goal. So one natural interpretation that can be made on $D_\theta(\hat{y};z)$ is that it captures the probability that z is indeed the ground-truth sensitive attribute for $\hat{y}$. Here the softmax function is applied to ensure the sum-up-to-one constraint (16).

Analogy with GAN [15]

Since the classifier and the discriminator are competing, one can make an analogy with a famous generative model: Generative Adversarial Networks (GANs), in which the generator and the discriminator also compete as in a two-player game. While the fair classifier and GANs bear strong similarity in their nature, these two are distinct in their roles. See Fig. 7 for the detailed distinctions.



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

Figure 7: MI-based fair classifier vs. GAN. Both bear similarity in structure (as illustrated in Fig. 6), yet distinctions in their roles.

Extension to another fairness measure DEO

So far we have focused on one fairness measure DDP. One can also apply almost the same trick to another measure DEO:

$$\mathsf{DEO} := \sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} |\mathbb{P}(\tilde{Y}=1|Y=y,Z=z) - \mathbb{P}(\tilde{Y}=1|Y=y)|. \tag{27}$$

Specifically one can make a similar connection like:

$$\mathsf{DEO} = 0 : \tilde{Y} \perp Z | Y \quad \Longleftarrow \quad I(Z; \hat{Y}|Y) = 0. \tag{28}$$

This then leads to an implementable optimization:

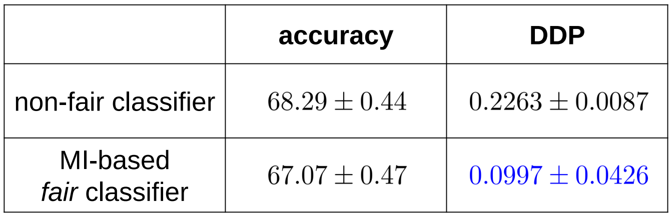
$$\min_{w} \max_{\theta : \sum_z D_\theta(\hat{y};z,y)=1} \frac{1}{m} \left\{ \sum_{i=1}^{m} (1-\lambda)\ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)}; z^{(i)}, y^{(i)}) \right\}. \tag{29}$$

Here the only distinction is that we read $D_\theta(\hat{y}; z, y)$ instead of $D_\theta(\hat{y}; z)$.

Experiments

We provide experimental results to demonstrate that the MI-based fair classifier offers a good fairness performance. For illustrative purpose, we focus on a single yet popular benchmark real data: COMPAS [16]. Also we consider only one baseline: a non-fair classifier which does not incorporate any fairness-regularized term. For a sensitive attribute, we consider a race type (white vs. black), so Z is binary. In COMPAS, X contains prior criminal records, e.g., felony or misdemeanour and Y denotes whether or not an associated individual reoffends within two years.

Fig. 8 exhibits accuracy-vs-DDP tradeoff performances for the non-fair and MI-based fair classifiers. Notice that the fair classifier yields a significant fairness performance (reflected in a small DDP) with a negligible performance degradation in prediction accuracy.



	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

Figure 8: Accuracy-vs-DDP tradeoff. The MI-based fair classifier improves DDP significantly with a marginal degradation of accuracy.

A challenge

While it offers a great tradeoff performance, it comes with a challenge. The challenge is that the min max structure in the MI-based optimization (26) may lead to *training instability*. The training instability problem indeed occurs. The problem is particularly significant when λ is around 1. See Fig. 9. Here each point represents a performance evaluated on a single seed in training. We see different points spread over a wide range of DDP, implying an unstable training performance.

Look ahead

There has been a recent work [11] that addresses the training instability while offering a better tradeoff. It is based on a prominent statistical method often employed by information theorists: *kernel density estimation (KDE)*. Next lecture, we will explore the KDE-based fair classifier.

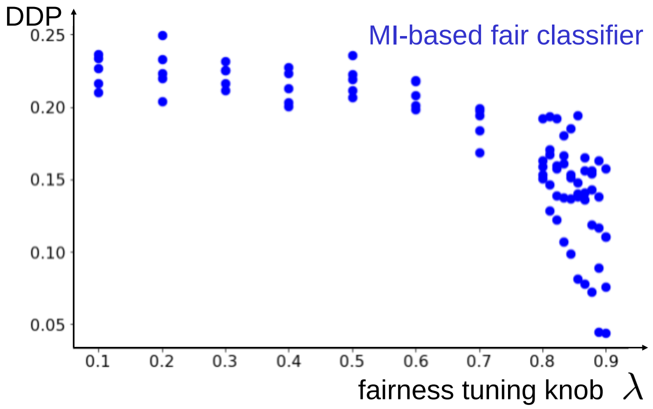


Figure 9: DDP as a function of the fairness tuning knob λ. Each blue dot corresponds to a single result w.r.t. one particular seed for training. The spreadness of the blue dots in particular near $\lambda \approx 1$ implies that the $\min\max$ optimization framework (26) yields different results with distinct seeds, thereby incurring training instability.

References

[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.

[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.

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

[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.