IEEE East Asian School of Information Theory
KAIST, August 3 ∼ 6, 2021

August 3, 2021
Changho Suh (chsuh@kaist.ac.kr)

Lecture 1: Fair machine learning & overview

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.

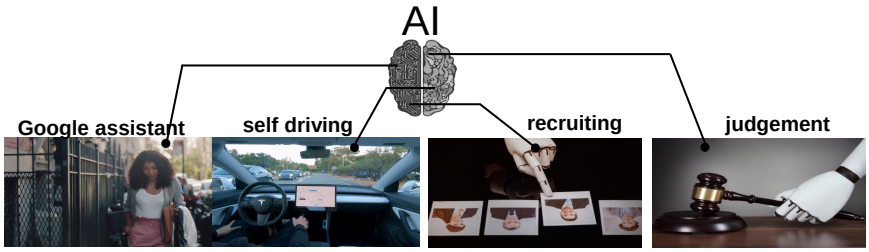

Figure 1: AI plays a powerful role in many applications.

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

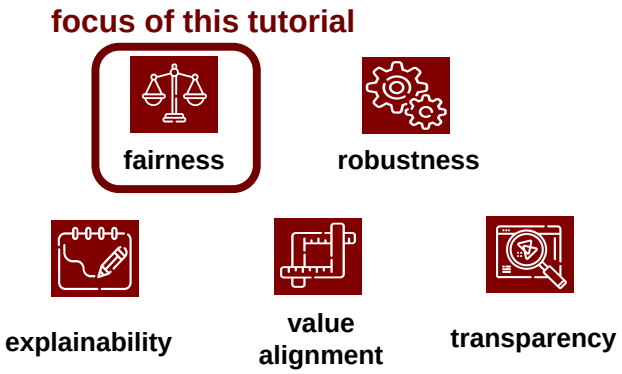

Figure 2: Five requirements for enabling trustworthy AI: (i) *fairness* across different demographics and/or individuals; (ii) *robustness* to data poisoning; (iii) *explanability* of trained models; (iv) *alignment* of model's output with actually what people want in reality; and (v) *transparency*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

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

Two fairness contexts of this tutorial's focus

Specifically we will explore fairness issues in the context of two prominent machine learning models. The first concerns a supervised learning setup. That is, *fair classifiers* which intend to make unbiased decisions in light of different groups and/or individuals. The second is about an unsupervised learning setup. That is, *fair generative models* wherein the goal is to synthesize fake samples that resemble real data while ensuring fairness.

Fairness in the context of classifiers

To figure out what we are going to study in detail, let me first explain what it means by *fairness* in the context of classifiers. There are many fairness concepts that people have considered in this context. One prominent concept of this tutorial's focus is the so called *group fairness*. 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? The reason is obvious. 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

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.

demographics.

Demographic Parity (DP)

One concrete fairness condition (in the realm of group fairness) that is very popular and simple, and therefore I would like to focus on in this tutorial is the so called Demographic Parity (DP) condition [1, 2]. Let me explain what the DP condition 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 output 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 simply 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; in this example, $\mathcal{Z} = \{0, 1\}$. There are many ways to quantify the DP condition. One natural way that we will take in this tutorial 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 $\mathsf{DDP} = 0$ 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 outcomes of the two groups are far apart 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 quite distinct 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*, 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 this condition may have nothing to do with such asymmetric case $\mathbb{P}(Y = 1 | Z = 1) \gg \mathbb{P}(Y = 1 | Z = 0)$. Hence, enforcing the EO condition may not necessarily degrade predication accuracy. 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):

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

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 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

[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.

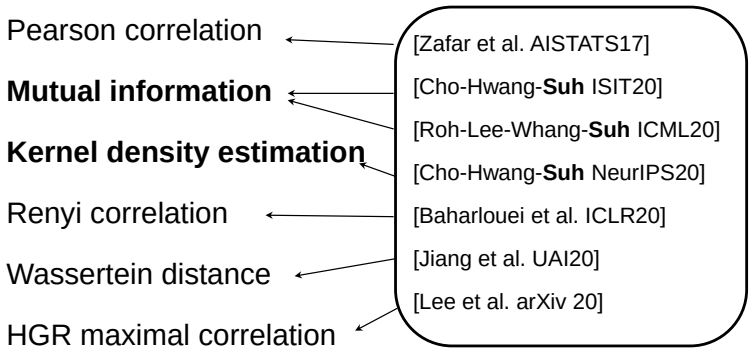

Figure 5: References taking approaches relevant to information theory.

(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 during the upcoming lectures. In Lecture 2, we will study an interesting connection between fairness measures (DDP and DEO) and mutual information (MI), and then will build upon the connection to investigate an MI-inspired fair classifier developed in [9]. In Lecture 3, we will explore the state of the art based on KDE [11].

Fairness in the context of generative models

From Lecture 4 onwards, we will switch-gear to explore another context: *fair generative models*. To figure out what we will study in detail, let me first explain what it means by fairness in this context. Similar to the prior context, there are also many fairness concepts. One important concept that this tutorial will focus on is the so called *fair representation*. It pursues class-balanced generated samples even when trained with size-biased real data across different demographics.

To better understand the concept, let us consider a scenario in which fake samples are generated via a partial set of CelebA [23] which exhibits an asymmetric ratio of female to male samples $\sim 85:15$. Fig. 6 shows generated fake samples trained with the size-biased real data. We see

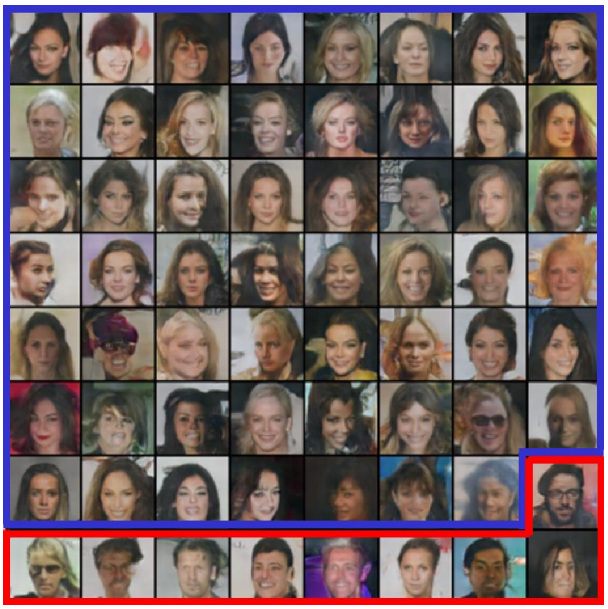

Figure 6: Generated fake samples trained with biased real data taking the female-to-male ratio as $\sim \frac{85}{15}$.

that the generated samples are indeed biased as we expect. The female samples are dominant. A natural goal in this context is to develop a *fair generative model* that promotes fair representation even under such imbalanced real data.

A measure for fair representation

In an effort to quantify the degree of fair representation, a measure has been introduced, named Fairness Discrepancy (FD) [20]. It is defined as the ℓ_2-norm distance between the distribution $\mathbb{P}_G(z)$ of generated samples w.r.t. sensitive attribute $z \in \mathcal{Z}$ and the desired distribution $\mathbb{P}_{\text{desired}}(z)$, e.g., uniform:

$$\text{FD} := \left\| \begin{bmatrix} \mathbb{P}_G(z_1) \\ \vdots \\ \mathbb{P}_G(z_{|\mathcal{Z}|}) \end{bmatrix} - \begin{bmatrix} \mathbb{P}_{\text{desired}}(z_1) \\ \vdots \\ \mathbb{P}_{\text{desired}}(z_{|\mathcal{Z}|}) \end{bmatrix} \right\|_2 = \sqrt{\sum_{z \in \mathcal{Z}} (\mathbb{P}_G(z) - \mathbb{P}_{\text{desired}}(z))^2}. \tag{5}$$

[Xu et al. BigData18]
[Xu et al. BigData19]
[Sattigeri et al. IBM-Journal19]
[Jalal et al. ICML21]
[Yu et al. ECCV20]
[Choi et al. ICML20]
[Tan et al. arXiv20]
[Um-**Suh** '21]

Figure 7: References regarding fair generative models.

Recent works on fair generative models

There are some recent works along this research direction. Fig. 7 shows a list of the associated references. The references in the second column concern *fair representation* of this tutorial's focus. These are further categorized into two groups. One group exploits demographic labels to

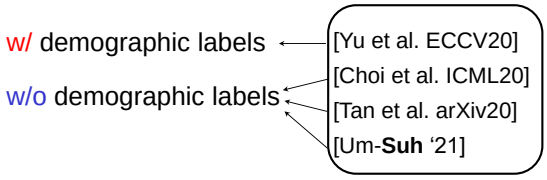

Figure 8: Reference for fair representation.

design a generative model that promotes fair representation. The paper by Yu et al. [19] belongs to this. The other group does not require the knowledge of such demographic labels [20, 21, 22].

Focus of this tutorial

This tutorial puts an emphasis on one recent work [22] of the second category. The reasons are similar as before. I can teach it well since I am a co-author of the paper. Also, it targets challenging yet more realistic scenarios where demographic labels are often unavailable in practice due to privacy. In addition, it employs a well-known statistical measure that often arises in information theory and hence that you may be interested in: *total variation distance* (TVD). Further, it is the state of the art, performing well both in FD (fairness) and the quality of samples.

But there is one thing we need to be prepared to understand the technique in [22]. Actually it is built upon a famous GAN (Generative Adversarial Network) that some people often call *hinge GAN* [24, 25]. So in Lecture 4, we will first study hinge GAN, as well as make an interesting with TVD. Building upon the connection, we will then explore the TVD-based fair generative model in Lecture 5. In Lecture 6, we will discuss a couple of other issues. Specifically we will investigate some of our recent works that address the *robustness aspect* in addition to fairness. We will also discuss other fairness contexts beyond fair classifiers and fair generative models.

Look ahead

Next lecture, we will embark on fair classifiers and explore a connection between fairness measures and mutual information.

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] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. *IEEE International Conference on Big Data (Big Data)*, 2018.

[16] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. *IEEE International Conference on Big Data (Big Data)*, 2019.

[17] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 2019.

[18] A. Jalal, S. Karmalkar, J. Hoffmann, A. G. Dimakis, and E. Price. Fairness for Image Generation with Uncertain Sensitive Attributes. *In Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[19] N. Yu, K. Li, P. Zhou, J. Malik, L. Davis, and M. Fritz. Inclusive gan: Improving data and minority coverage in generative models. *European Conference on Computer Vision (ECCV)*, 2020.

[20] K. Choi, A. Grover, T. Singh, R. Shu, and S. Ermon. Fair generative modeling via weak supervision. *In Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[21] S. Tan, Y. Shen, and B. Zhou. Improving the fairness of deep generative models without retraining. *arXiv:2012.04842*, 2020.

[22] S. Um and C. Suh. A fair generative model using total variation distance. submitted to *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision (ICCV)*, 2015.

[24] J. H. Lim and J. C. Ye. Geometric GAN. *arXiv:1705.02894*, 2017.

[25] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2019.