Learning Adversarially Fair and Transferrable Representations

$\underbrace{\text{David Madras}^{\natural\flat}}_{\text{me}}$ Elliot Creager[♮♭] Toniann Pitassi[♮♭] Richard Zemel[♮♭]

[♮]University of Toronto
[♭]Vector Institute

July 13, 2018

- Classification: a tale of two parties
- Example: **targeted advertising**: owner $\rightarrow$ vendor $\rightarrow$ prediction

Data owner

Prediction vendor

[Dwork et al., 2012]

Why Fairness?

- Want to minimize unfair targeting of disadvantaged groups by vendors
    - e.g. showing ads for worse lines of credit, lower paying jobs
- We want **fair predictions**

Data owner

Prediction vendor

Why Fair Representations?

- Previous work emphasized the role of the vendor
- Can we trust the vendor?
- How can the **owner** ensure fairness?

Data owner

Prediction vendor

The Data Owner

- How should the data be represented?
  - Feature selection? Measurement?
- How can we choose a data representation that ensures fair classifications downstream?
- Let's *learn* a fair representation!

Data owner→*Representation learner*

[Zemel et al., 2013]

Background: Fair Classification

Assume: data $X \in \mathbb{R}^d$, label $Y \in 0, 1$, sensitive attribute $A \in 0, 1$
Goal: predict $\hat{Y}$ fairly with respect to A

- Demographic parity

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

- Equalized odds

$$P(\hat{Y} \neq Y|A = 0, Y = y) = P(\hat{Y} \neq Y|A = 1, Y = y) \; \forall y \in \{0, 1\}$$

- Equal opportunity: equalized odds with only $Y = 1$

$$P(\hat{Y} \neq Y|A = 0, Y = 1) = P(\hat{Y} \neq Y|A = 1, Y = 1)$$

[Dwork et al., 2012] [Hardt et al., 2016]

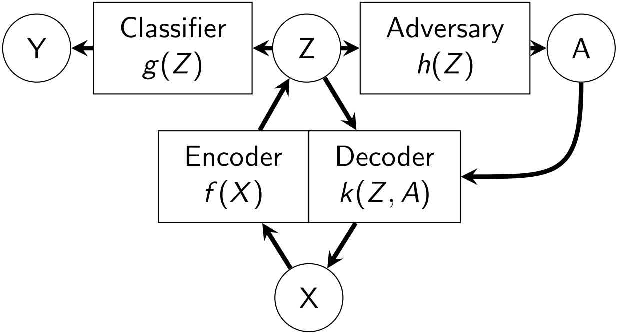
Goals of Fair Representation Learning

- Fair classification: learn $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$
  - encoder f, classifier g
- Fair representation: learn $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$
- $Z = f(X)$ should:
  - Maintain **useful information** in X
  - **Yield fair downstream classification** for vendors g

Types of unfair vendors

- Consider two types of unfair vendors
  - The **indifferent** vendor: doesn't care about fairness, only maximizes utility
  - The **malicious** vendor: doesn't care about utility, discriminates maximally
- This suggests an adversarial learning scheme

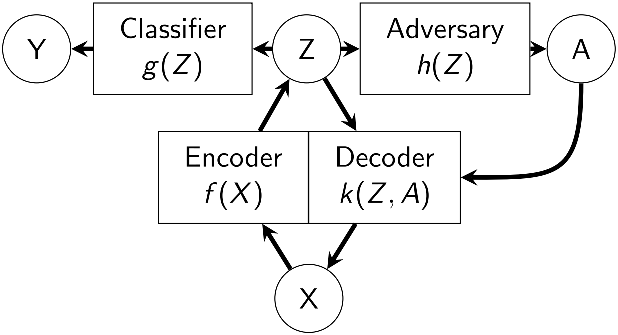
Learning Adversarially Fair Representations



- The classifier is the indifferent vendor, forcing the encoder to make the representations useful
- The adversary is the malicious vendor, forcing the encoder to hide the sensitive attributes in the representations

[Edwards and Storkey, 2015]

Adversarial Learning in LAFTR

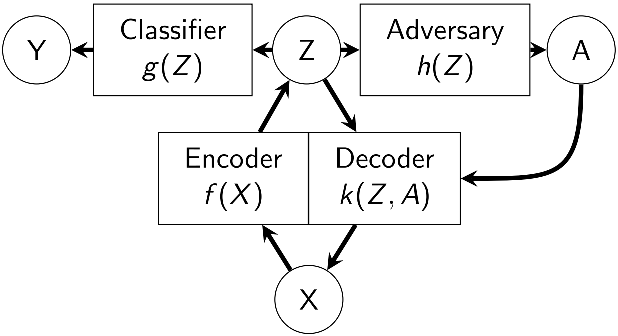


- Our game: encoder-decoder-classifier vs. adversary
- Goal: learn a fair encoder

$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[\mathcal{L}(f, g, h, k)\right].$$

$$\mathcal{L}(f, g, h, k) = \alpha\mathcal{L}_{Class} + \beta\mathcal{L}_{Dec} - \gamma\mathcal{L}_{Adv}$$

Adversarial Objectives



Choice of adversarial objective depends on fairness desideratum

- Demographic parity: $\mathcal{L}_{Adv}^{DP}(h) = \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a) \in \mathcal{D}_i} |h(f(x)) - a|$
- Equalized odds: $\mathcal{L}_{Adv}^{EO}(h) = \sum_{i,j \in \{0,1\}^2} \frac{1}{|\mathcal{D}_i^j|} \sum_{(x,a,y) \in \mathcal{D}_i^j} |h(f(x), y) - a|$
- Equal Opportunity: $\mathcal{L}_{Adv}^{EOpp}(h) = \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i^1|} \sum_{(x,a) \in \mathcal{D}_i^1} |h(f(x)) - a|$

From Adversarial Objectives to Fairness Definitions

In general: pick the right adversarial loss, encourage the right conditional independencies

- Demographic parity encourages $Z \perp A$ to fool adversary
- Equalized odds encourages $Z \perp A \mid Y$ to fool adversary
- Equal opportunity encourages $Z \perp A \mid Y = 1$ to fool adversary

Note that independencies of $Z = f(x)$ also hold for predictions $\hat{Y} = g(Z)$

We show: In the adversarial limit, these objectives guarantee these fairness metrics!

- The key is to connect predictability of A by the adversary $h(Z)$ to unfairness in the classifier $g(Z)$

Theoretical Properties

- Define $\Delta_{DP}(g) \triangleq$ DP-unfairness of classifier g
- Define $\mathcal{L}_{Adv}^{DP}(h) \triangleq$ adversarial loss (inv. weighted error)

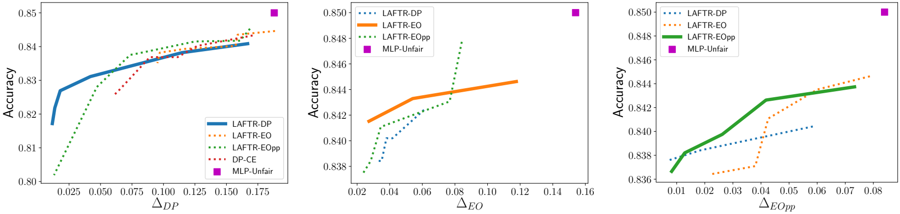
- We show: $\forall$ classifier $g(Z)$, we can construct an adversary $h(Z)$ s.t. $-\mathcal{L}_{Adv}^{DP}(h) = \Delta_{DP}(g)$

- Let $h^\star$ be the optimal adversary. Then

$$-\mathcal{L}_{Adv}^{DP}(h^\star) \geq -\mathcal{L}_{Adv}^{DP}(h) = \Delta_{DP} \tag{1}$$

- Takeaway: if $-\mathcal{L}_{Adv}^{DP}(h^\star)$ is forced to be small, Δ_{DP} will be too
- Holds for EO as well, but with h as a function of Y also

Results - Fair Classification (Adult)



- Train with two-step method to simulate owner $\rightarrow$ vendor framework
- Tradeoffs between accuracy and various fairness metrics yielded by different LAFTR loss functions
- Seems to work best for fairest solutions

Setup - Fair Transfer Learning

- Downstream vendors will have unknown prediction tasks
- Does **fairness** transfer?
- We test this as follows:
  1. Train encoder f on data X, with label Y
  2. Freeze encoder f
  3. On new data X', train classifier on top of $f(X')$, with new task label Y'
  4. Observe fairness and accuracy of this new classifier on new task Y'
- Compare LAFTR encoder f to other encoders
- We use Heritage Health dataset
  - Y is Charlson comorbidity index > 0
  - Y' is whether or not a certain type of insurance claim was made
  - Check for fairness w.r.t. age

Results - Fair Transfer Learning

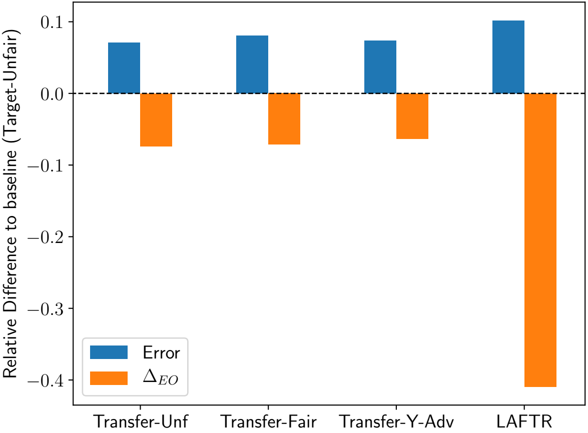


Figure 2: Fair transfer learning on Health dataset. Down is better in both metrics.

Conclusion

- Propose LAFTR: general model for fair representation learning
- Connect common fairness metrics to adversarial objectives
- Demonstrate that training with LAFTR improves transfer fairness
- Open questions:
  - Compare adversarial/non-adversarial methods?
  - Transfer fairness: datasets, limitations, better methods?
- Come check out our poster #44 tonight!

References

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM.

Edwards, H. and A. Storkey (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Hardt, M., E. Price, N. Srebro, et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork (2013). Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333.