Trustworthy ML Reading Group

31 Jan 2024

Conduct

- **From Ethos of MLC…**
  - https://mlcollective.org/wiki/code-of-conduct/
- **Highlights**
  - Expectation of Confidentiality
  - Reporting -> send me a direct message

Differentially Private Fair Learning

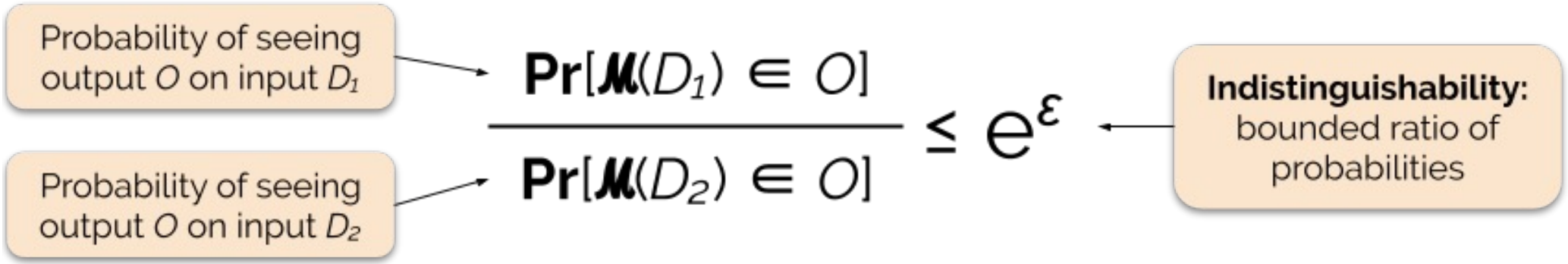
Matthew Jagielski, Michael Kearns, Jieming Mao, Aaron Oprea, Alina Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. ***Differentially private fair learning****. arXiv preprint arXiv:1812.02696, 2019.*

Differential Privacy (DP)

- Applies the following idea:
  - *No one should know if your information is a part of a dataset.*
  - *This should be true if you decide to remove or add your information to a dataset*
  - *The definition below is a relaxation of DP*

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta.$$

(ε,δ) differential privacy



Differential Privacy

- *Small ε => hard to distinguish between two datasets*
- *Large ε => very easy to distinguish two datasets*

$$P_0 [C = r \mid Y = y] = P_1 [C = r \mid Y = y] \; \forall \; r, y$$

C => class
Y => output
r => a protected attribute

Fairness

- **Equalized Odds**
  - *Model Performance (Accuracy in this case) should be similar across different protected attributes*
  - *Paper relaxes the basic definition to* ***γ-equalized odds***

Building Blocks for Trust

Bias and Fairness

Explainability and Interpretability

Privacy and Security

Robustness

Causality

https://www.trustworthyml.org/

Privacy and Fairness

- Enforcing Fairness
  - Can make a system leak private information
- Enforcing Privacy
  - Can make a system “unfair”
- Sometimes enforcing both causes utility to suffer

Paper

- **Big Idea**
  - *Create a model that combines differential privacy with Equalized Odds*
  - ***Solution 1:*** *Follow Equalized Odds approach. Use protected attributes to measure fairness during test-time. Use DP in Equalized Odds approach*
  - ***Solution 2:*** *minimize violating fairness while improving utility of a private model (zero-sum game)*

Paper

“Our results therefore suggest that the requirement that we not use the protected attribute at test time (i.e. that we be avoid “disparate treatment”) ***might be extremely burdensome*** *if we also want the protections of differential privacy and have only" small dataset sizes.”*

Interesting Takeaways

- *There is a tradeoff when optimizing fairness, privacy, and utility*
  - *There is a constant tradeoff between Privacy and Utility*
- *Only private info is guaranteed "differentially private" (not all info)*

Fairness and Privacy solution

Algorithm	Assumptions on $\mathcal{H}$	Fairness Guarantee	Needs access to A at test time?	Does it guarantee privacy of X as well?	Error	Fairness Violation
DP-postprocessing	None	Equalized Odds	Yes	No	$\tilde{O}\left(\frac{\lvert\mathcal{A}\rvert}{m\epsilon}\right)$ [1]	$\tilde{O}\left(\frac{1}{\min \hat{q}_{ay} m\epsilon}\right)$
DP-oracle-learner	$d_{\mathcal{H}} < \infty$ $d_{\mathcal{H}} := VC(\mathcal{H})$	Equalized Odds	No	No	$\tilde{O}\left(\frac{B}{\min \hat{q}_{ay}}\sqrt{\frac{\lvert\mathcal{A}\rvert d_{\mathcal{H}}}{m\epsilon}}\right)$	$B^{-1} + \tilde{O}\left(\frac{1}{\min \hat{q}_{ay}}\sqrt{\frac{\lvert\mathcal{A}\rvert d_{\mathcal{H}}}{m\epsilon}}\right)$
	$\lvert\mathcal{H}\rvert < \infty$	Equalized Odds	No	Yes	$\tilde{O}\left(\frac{B}{\min \hat{q}_{ay}}\sqrt{\frac{\lvert\mathcal{A}\rvert \ln(\lvert\mathcal{H}\rvert)}{m\epsilon}}\right)$	$B^{-1} + \tilde{O}\left(\frac{1}{\min \hat{q}_{ay}}\sqrt{\frac{\lvert\mathcal{A}\rvert \ln(\lvert\mathcal{H}\rvert)}{m\epsilon}}\right)$
	$\lvert\mathcal{H}\rvert < \infty$, $\mathcal{H}$ has maximally discriminatory classifiers	Equalized False Positive Rate	Yes	Yes	$\tilde{O}\left(\frac{\lvert\mathcal{A}\rvert}{\min \hat{q}_{ay}}\sqrt{\frac{\lvert\mathcal{A}\rvert \ln(\lvert\mathcal{H}\rvert)}{m\epsilon}}\right)$	$\tilde{O}\left(\frac{\lvert\mathcal{A}\rvert}{\min \hat{q}_{ay}}\sqrt{\frac{\lvert\mathcal{A}\rvert \ln(\lvert\mathcal{H}\rvert)}{m\epsilon}}\right)$

Complementary reading

- Bagdasaryan, E., & Shmatikov, V. (2019). **Differential privacy has disparate impact on model accuracy**. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1905.12101
- Mangold, P. (2022, October 28). **Differential Privacy has Bounded Impact on Fairness in Classification.** arXiv.org. https://arxiv.org/abs/2210.16242
- Arcolezi, H. H., Makhlouf, K., & Palamidessi, C. (2023). **(Local) Differential Privacy has NO Disparate Impact on Fairness**. In *Lecture Notes in Computer Science* (pp. 3–21). https://doi.org/10.1007/978-3-031-37586-6_1

What I’m thinking about

- **Me…**
  - DP appears to be a very strong guarantee of one’s privacy. If we use a domain specific case (e.g., student information) could we use another privacy solution? (e.g., MPC, FL, Anonymity, etc.)
  - Contradictions of Fairness and Privacy exist in law. We normally treat this case by case. Is there any use to this idea when applied to AI (use domain knowledge to change models)?
  - What happens when other accuracy metrics are used?
  - *Bounded (worse case) fairness, error, and privacy definitions*
- **What are you thinking about?**