Differential Privacy has Bounded Impact on Fairness in Classification

January 19, 2024

Seoul National University

Table of Contents

Contribution

- Distance between private model via output smoothing and optimal model, and the difference between their fairness levels are bounded by $O(\sqrt{p}/n)$.

Table of Contents

Notation

- $\mathcal{X}$: Feature space in $\mathbb{R}^p$
- $\mathcal{Y}$: Finite set of labels
- $\mathcal{S} \subset \mathcal{X}$: Set of sensitive attributes
- $\mathcal{D}$: Distribution over $\mathcal{X} \times \mathcal{Y}$
- $D = \{(x_1, y_1), \cdots, (x_n, y_n)$: i.i.d data from $\mathcal{D}$
- $\mathcal{H}$: function space of $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
- $H(x)$: $\text{argmax}_{y \in \mathcal{Y}} h(x, y)$
- $\rho(h, x, y) = h(x, y) - \max_{y' \neq y} h(x, y')$: Margin of a model h for an example-label pair (x, y)

Table of Contents

Fairness

- Focus on Group Fairness.
- As in Maheshwari & Perrot, when data can be partitioned into K disjoint groups by $D_1, \cdots, D_k$ (ex : $D_{(y=1,s=1)}, D_{(y=0,s=1)}, D_{(y=1,s=0)}, D_{(y=0,s=0)}$), fairness definitions can be written as

$$F_k(h, D) = C_k^0 + \sum_{k'=1}^{K} C_k^{k'} \mathbb{P}(H(X) = Y \mid D_{k'})$$

where the $C_k^{k'}$'s are group specific values independent of h.

Fairness

- Example : Equalized Odds (Hardt et al., 2016)

  - Let $\forall (y,s) \in \mathcal{Y} \times \mathcal{S}$, $\mathcal{Y} = \{0,1\}$

  - $F_{(y,s)}(h,D) = \mathbb{P}(H(X) = Y | Y = y, S = s) - \mathbb{P}(H(X) = Y | Y = y)$.

$$= C^0_{(y,s)} + \sum_{(y',s') \in \mathcal{Y} \times \mathcal{S}} C^{(y',s')}_{(y,s)} \mathbb{P}\left(H(x) = Y \mid Y = y', S = s'\right)$$

  with when $y = 1$

$$\begin{aligned} C^0_{(y,s)} &= 0 \\ C^{(y,s)}_{(y,s)} &= 1 - \mathbb{P}(S = s \mid Y = y) \\ \forall s' \neq s, C^{\left(y,s'\right)}_{(y,s)} &= -\mathbb{P}\left(S = s' \mid Y = y\right) \\ \forall y' \neq y, \forall s' \in \mathcal{S}, C^{\left(y',s'\right)}_{(y,s)} &= 0 \end{aligned}$$

when $y = 0$, $C^{\left(y',s'\right)}_{(y,s)} = 0$ for all $s \in$ S

Fairness

- Use the mean of the absolute fairness level of each group:

$$\text{Fair}(h, D) = \frac{1}{K} \sum_{k=1}^{K} |F_k(h, D)|$$

which is 0 when h is fair and positive when it is unfair.

Table of Contents

Definition(Dwork,2006)

- Let $\mathcal{A}^{\text{priv}} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ be a randomized algorithm.
- Define $\mathcal{A}^{\text{priv}}$ is (ϵ, δ)-differentially private if, for all neighboring datasets $D, D' \in (\mathcal{X} \times \mathcal{Y})^n$ and all subsets of hypotheses $\mathcal{H}' \subseteq \mathcal{H}$,

$$\mathbb{P}\left(\mathcal{A}^{\text{priv}}(D) \in \mathcal{H}'\right) \leq \exp(\epsilon)\mathbb{P}\left(\mathcal{A}^{\text{priv}}(D') \in \mathcal{H}'\right) + \delta$$

Table of Contents

Output perturbation

- Define h_n^* as

$$h_n^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h; x_i, s_i, y_i\right)$$

- Output perturbation make the non-private solution h_n^* be a private estimate by the Gaussian mechanism :

$$h^{\text{priv}} = \pi_{\mathcal{H}}\left(h^* + \mathcal{N}\left(\sigma^2 \mathbb{I}_p\right)\right)$$

where $\pi_{\mathcal{H}}$ is the projection on $\mathcal{H}$.

- It is known that given $\epsilon > 0$ and $\delta < 1$, h^{priv} is (ϵ, δ)-differentially private as long as

$$\sigma^2 \geq 2\Delta^2 \log(1.25/\delta)/\epsilon^2$$

where $\Delta = 2\Lambda/\mu n$

Assumption

- ρ is Lipschitz-continuous

$$\left|\rho(h, x, y) - \rho\left(h', x, y\right)\right| \leq L_{x,y} \left\|h - h'\right\|_{\mathcal{H}},$$

where $L_{x,y} < +\infty$ depends on the example (x, y) and $\|\cdot\|_{\mathcal{H}}$ is Eucildean and $\mathrm{H}isconvex$.

- Loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is Λ-Lipschitz and μ-strongly convex with respect to h.

Theorem

Theorem

Let h^{priv} be the vector released by output perturbation with noise $\sigma^2 = 8\Lambda^2 \log(1.25/\delta)/\mu^2 n^2 \epsilon^2$, and $0 < \zeta < 1$, then with probability at least $1 - \zeta$,

$$\left\| h^{priv} - h^* \right\|_2^2 \leq \frac{32 p \Lambda^2 \log(1.25/\delta) \log(2/\zeta)}{\mu^2 n^2 \epsilon^2}$$

Theorem

Theorem
With probability at least $1-\zeta$*,*

$$\left|F_k\left(h^{priv}, D\right) - F_k\left(h^*, D\right)\right| \leq \frac{\chi_k\left(h^{ref}, D\right) L\Lambda\sqrt{32p\log(1.25/\delta)\log(2/\zeta)}}{\mu n \epsilon}.$$

where $h^{ref} \in \{h^{priv}, h^*\}$ *and* $\chi_k(h, D) = \sum_{k'=1}^{K}\left|C_k^{k'}\right|\mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|}\,\middle|\, D_{k'}\right)$.

Experiment

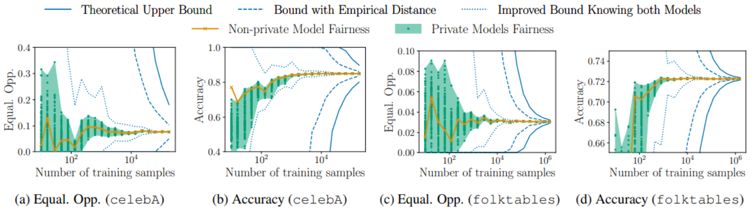

Figure 1: Experiment Result

- Private models mean $(1, 1/n^2)$-DP model learned by output perturbation.