A Fair Generative Model Using LeCam Divergence

March 21, 2024

Seoul National University

Table of Contents

Contribution

- Generate fair synthetic data via LeCam divergence and unlabelled reference dataset.

Table of Contents

Notation

- $\mathcal{X}$: Data in $\mathbb{R}^D$
- $\mathcal{S}$: Set of sensitive attributes

Table of Contents

Fairness

- Fair synthetic data satisfies

$$P_G(S = 0) = P_G(S = 1)$$

  where P_{syn} is distribution of synthetic data.
- However train data does not satisfy above condition.

$$P_{bias}(S = 0) \neq P_{bias}(S = 1)$$

Table of Contents

Setup

- Suppose that the information of S is not available.
- To train fair synthetic data, adopt reference dataset $\mathcal{D}^{\text{ref}}$ which may satisfy

$$P_{ref}(S = 0) \approx P_{ref}(S = 1)$$

- Let $\mathcal{D}^{\text{bias}}$ be train data.
- When the number of train data and reference data are m_{bias} and m_{ref} respectively, we assume that

$$m_{bias} \gg m_{ref}$$

Original Method

- To train fair synthetic data, we optimize generator G by minimizing

$$\min_G (1-\lambda) \cdot \mathrm{D}_f\left(\mathbb{P}_{\text{bias}} \| \mathbb{P}_G\right) + \lambda \cdot \mathrm{D}_{\text{fair}}\left(\mathbb{P}_{\text{ref}} \| \mathbb{P}_G\right)$$

where D_f is f-divergence and D_{fair} is fair discrepancy.

Table of Contents

LC-divergence

- Propose Lecam divergence based fairness discrepancy

$$\min_G (1-\lambda) \cdot \mathrm{D}_f \left(\mathbb{P}_{\text{bias}} \| \mathbb{P}_G\right) + \lambda \cdot \mu \mathrm{D}_\Delta \left(\mathbb{P}_{\text{ref}} \| \mathbb{P}_G\right)$$

  where μ denotes a non-negative weight, and $\mathrm{D}_\Delta \left(\mathbb{P}_{\text{ref}} \| \mathbb{P}_G\right)$ indicates the LC-divergence between $\mathbb{P}_{\text{ref}}$ and $\mathbb{P}_G$:

$$\mathrm{D}_\Delta \left(\mathbb{P}_{\text{ref}} \| \mathbb{P}_G\right) := \sum_{x \in \mathcal{X}} \frac{\left(\mathbb{P}_{\text{ref}}(x) - \mathbb{P}_G(x)\right)^2}{\mathbb{P}_{\text{ref}}(x) + \mathbb{P}_G(x)}$$

Training

- f-GAN

$$\min_G \max_D \mathbb{E}_{\mathbb{P}_{\textbf{bias}}}[D(X)] - \mathbb{E}_{\mathbb{P}_G}[f^*(D(X))]$$

where D is discriminator and f^* is conjugate of f.

Training

- fair f-GAN

$$\max_{D} \mathbb{E}_{\mathbb{P}_{\textbf{bias}}} [D(X)] - \mathbb{E}_{\mathbb{P}_G} [f^*(D(X))]$$

$$\max_{D_{\textbf{ref}}} \mathbb{E}_{\mathbb{P}_{\textbf{ref}}} [D_{\text{ref}}(X)] - \mathbb{E}_{\mathbb{P}_G} [D_{\text{ref}}(X)] - \frac{1}{2(\mu + \alpha)} R_\Delta$$

$$\min_{G} -(1-\lambda)\mathbb{E}_{\mathbb{P}_G} [f^*(D(X))] - \lambda \mathbb{E}_{\mathbb{P}_G} [D_{\text{ref}}(X)]$$

where α denotes an exponential moving average of D_{ref} v.r.t. reference samples and R_Δ indicates a regularization term for D_{ref} defined as:

$$R_\Delta := \mathbb{E}_{\mathbb{P}_{\textbf{ref}}} \left[\|D_{\text{ref}}(X) + \alpha\|^2\right] + \mathbb{E}_{\mathbb{P}_G} \left[\|D_{\text{ref}}(X) - \alpha\|^2\right]$$

Experiment

- Baseline 1 : Unfair method with train and reference data
- Baseline 2 : Unfair method with reference data
- Fairness measure : $\sqrt{\sum_{s=1}^{S} (P_{ref}(S = s) - P_G(S = s))^2}$

Experiment

Reference set size		25%	10%	5%	2.5%	1%
Baseline I	Intra FID	12.00 ± 0.069	**12.73** ± **0.053**	**13.54** ± **0.074**	**13.79** ± **0.072**	**15.89** ± **0.094**
	Fairness	0.495 ± 0.001	0.554 ± 0.002	0.559 ± 0.001	0.566 ± 0.002	0.576 ± 0.002
Baseline II	Intra FID	23.81 ± 0.118	32.31 ± 0.109	40.07 ± 0.062	67.70 ± 0.112	92.34 ± 0.131
	Fairness	0.093 ± 0.002	0.115 ± 0.002	0.120 ± 0.003	0.150 ± 0.003	0.455 ± 0.002
Choi et al. (2020)	Intra FID	20.68 ± 0.076	25.74 ± 0.079	30.15 ± 0.037	30.40 ± 0.041	31.49 ± 0.074
	Fairness	0.065 ± 0.002	0.104 ± 0.002	0.126 ± 0.001	0.237 ± 0.003	0.344 ± 0.002
Proposed	Intra FID	**11.48** ± **0.814**	14.50 ± 0.996	14.64 ± 0.626	17.16 ± 1.607	23.11 ± 0.797
	Fairness	**0.037** ± **0.007**	**0.039** ± **0.013**	**0.118** ± **0.007**	**0.129** ± **0.010**	**0.146** ± **0.022**