ICONIP 2024

31st International Conference on Neural Information Processing

December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Towards Private and Fair Machine Learning: Group-Specific Differentially Private Stochastic Gradient Descent with Threshold Optimization

Zhi Yang, CSE, SUSTech, Shenzhen, China

Changwu Huang, CSE, SUSTech, Shenzhen, China

Xin Yao, School of Data Science, Lingnan University, Hong Kong, China

Sponsors: g.tec, 100% Pure New Zealand, AUT Knowledge Engineering & Discovery Research Innovation, Springer

Supported by: AUT University, University of Auckland, Nottingham Trent University, Asia Pacific Neural Network Society

Outline

SUSTech
Southern University of Science and Technology

1. Introduction

2. Related Work

3. Methodology

4. Experimental Study

5. Conclusion

1.Introduction

SUSTech
Southern University of Science and Technology

Emphasize the importance of data privacy protection.

《GDPR》

Technique

- Differential Privacy
- Homomorphic Encryption
- Federated Learning

……

➢ **Differential Privacy (DP) has emerged as the predominant choice for ensuring data privacy[1].**

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

1.Introduction

《ECOA》

Explicitly prohibits discrimination based on protected traits.

Existing ML algorithms exhibit varying degrees of discrimination in their decisions.

Fairness measurements

- Demographic Parity
- Accuracy Parity

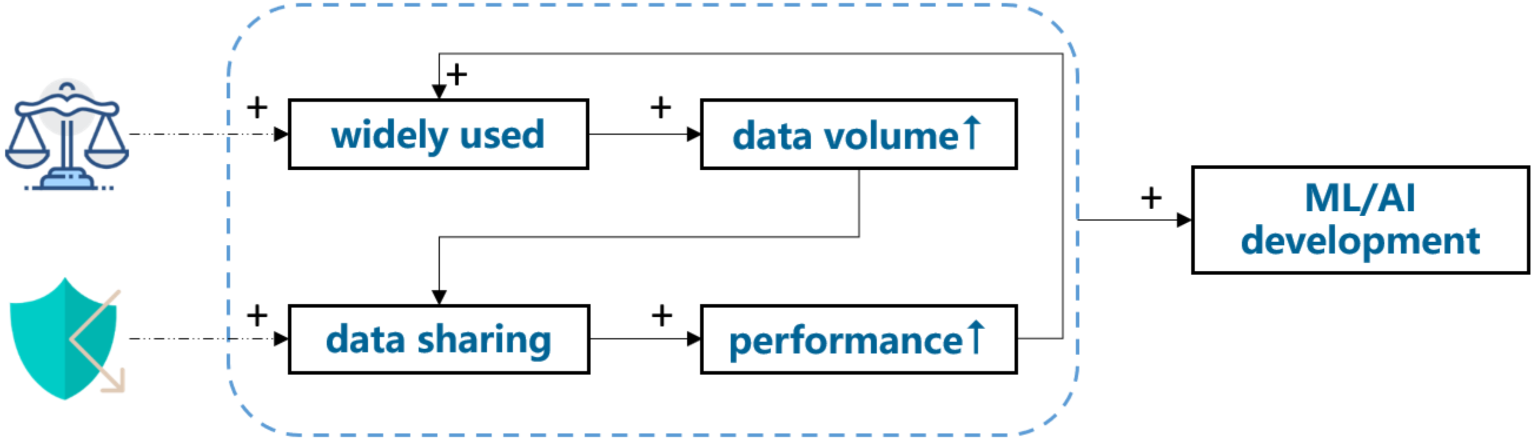
Fairness-aware ML methods

- Pre-process
- In-process
- Post-process

1.Introduction

SUSTech
Southern University of Science and Technology

- From both ethical and legal perspectives, **fairness** and **privacy** are two crucial aspects for the development of ML/AI.



They are interrelated research issues rather than isolated challenges.

1.Introduction

SUSTech
Southern University of Science and Technology

- Combining privacy and fairness poses challenges in two main categories: **addressing amplified unfairness due to DP** and **achieving outcome fairness in the ML model.**
- However, current methods often address the two objectives in isolation, overlooking their combined impact.
- To bridge this gap, we introduce **a group-specific DP stochastic gradient descent (DP-SGD) training mechanism with classification threshold optimization**, which concurrently addressing accuracy and outcome fairness issues in differentially private models.

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

2.Related Work

SUSTech
Southern University of Science and Technology

■Fairness measurements:

- Demographic Parity (DemParity) [2]:

$$|P(\hat{y} = 1|s = s_a) - P(\hat{y} = 1|s = s_b)| \leq \theta$$

- Accuracy Parity (AccParity) [3]:

$$|P(\hat{y} = y|s = s_a) - P(\hat{y} = y|s = s_b)| \leq \theta$$

2.Related Work

SUSTech
Southern University of Science and Technology

- **Differentially Private Stochastic Gradient Descent (DP-SGD)[4]:**

Algorithm 1 DP-SGD

Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, the parameterized model $f_w(\cdot)$, loss function $\ell(\hat{y}, y)$ for prediction $\hat{y}$ and label y, iterations T, batch size b, learning rate η, noise scale σ, gradient norm bound C.

```
1: Initialize w^(0) randomly.
2: for t = 0, 1, ..., T − 1 do
3:     Sample a batch B^(t) from D with sampling probability b/N for each data point.
4:     for i ∈ B^(t) do
5:         g_i ← ∇ℓ(f_{w^(t)}(x_i), y_i)
6:         ḡ_i ← g_i · min(1, C/||g_i||_2)
7:     end for
8:     g̃ ← (1/b)(Σ_{i∈B^(t)} ḡ_i + N(0, σ²C²I))
9:     w^(t+1) ← w^(t) − ηg̃
10: end for
```

Output: Model $f_{w^{(T)}}(\cdot)$ and accumulated (ϵ, δ).

$$\bar{g}_i \leftarrow g_i \cdot min(1, \frac{C}{\|g_i\|_2})$$

$$\tilde{g} \leftarrow \frac{1}{b}\left(\sum_{i \in B^{(t)}} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$$

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

2.Related Work

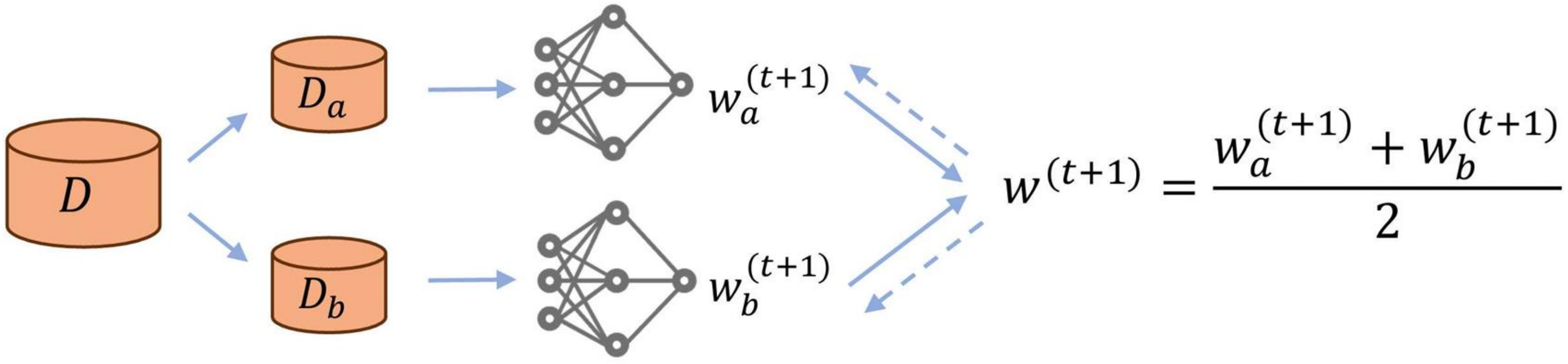
- **Mitigating the unfairness amplified by DP.** Recent studies have found that incorporating DP into models can increase AccParity measurement between sensitive groups[3, 5-6]. Various efforts have been made to address this issue[6, 7].

- **Achieving outcome fairness in differentially private models.** Several works study how to achieve outcome fairness using fairness-aware learning when enforcing DP in the private model[8, 9].

3.Methodology

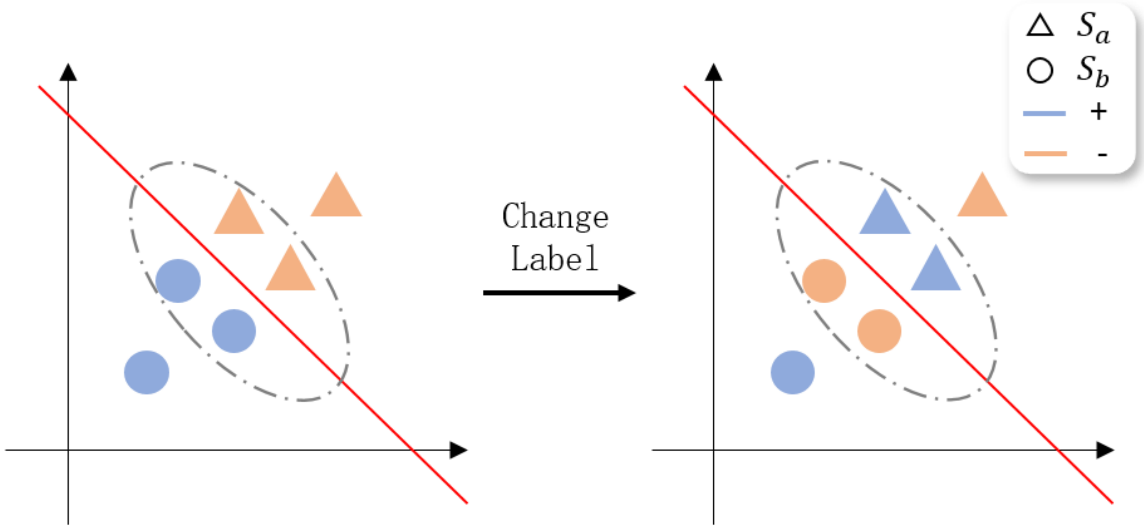
SUSTech
Southern University of Science and Technology

- We introduce GS-DP-SGD, a group-specific training strategy for DP-SGD that aims to alleviate the accuracy discrepancy exacerbated by DP-SGD.



3.Methodology

- To mitigate the outcome unfairness of the private model trained by GS-DP-SGD, we incorporate a post-processing method called reject option based classification (ROC)[10] .



Algorithm 4 Threshold Optimization based Classification (TOC)

Input: Validation dataset $D_{valid} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$, fairness constraint θ, model $f_{w^{(T)}}$.

for $\gamma \in$ `linspace(0.5, 1, 100)` **do**
    for $\mathbf{x}_i \in D_{valid}$ **do**
        $\hat{y}_i = \mathrm{ROC}(f_{w^{(T)}}(\cdot), \gamma, \mathbf{x}_i)$
    end for
    $m_\gamma = |P(\hat{y}=1|s=s_a) - P(\hat{y}=1|s=s_b)|$
    if $m_\gamma \le \theta$ **then Return** γ
end for
Return the γ that has minimal m_γ

Output: The threshold γ.

3.Methodology

- Overall, we implement GS-DP-SGD with Threshold Optimization (referred to as GS-DP-SGD-TO) to address both the exacerbated accuracy disparity (i.e., AccParity)and the outcome fairness (i.e., DemParity) issues.

Algorithm 5 GS-DP-SGD-TO

Input: Training dataset D_{train}, validation dataset D_{valid}, the parameterized model $f_w(\cdot)$, loss function $\ell(\cdot,\cdot)$, iterations T, batch size b, learning rate η, noise scale σ, gradient norm bound C, fairness constraint θ.

1: $f_{w^{(T)}}, (\epsilon, \delta) = \text{GD-DP-SGD}(D_{train}, f(\cdot), \ell(\cdot,\cdot), T, b, \eta, \sigma, C)$

2: $\gamma = \text{TOC}(D_{valid}, \theta, f_{w^{(T)}})$

Output: The ROC model $\text{ROC}(f_{w^{(T)}}, \gamma, \cdot)$ and the accumulated (ϵ, δ).

4.Experimental Study

■Experimental Setup：

- ***Datasets:*** Six commonly used binary classification datasets relevant to fairness research: **Adult, Dutch, Bank, Credit, Compas, and Law**.
- ***Comparison Methods:*** Methods aiming to reduce accuracy disparity: **DP-SGD-F[6], DP-SGD-A[7], and GS-DP-SGD.** Methods aiming to achieve fairness in model decisions: **FairDP[8] and DP-SGD-P[9].**
- ***Model:*** **MLP** with two hidden layers of 256 units each, a maximum of 20 iterations.
- ***Evaluation Metrics:*** Two common group fairness metrics: **Accuracy Parity (AccParity) and Demographic Parity (DemParity)**.

4.Experimental Study

SUSTech
Southern University of Science and Technology

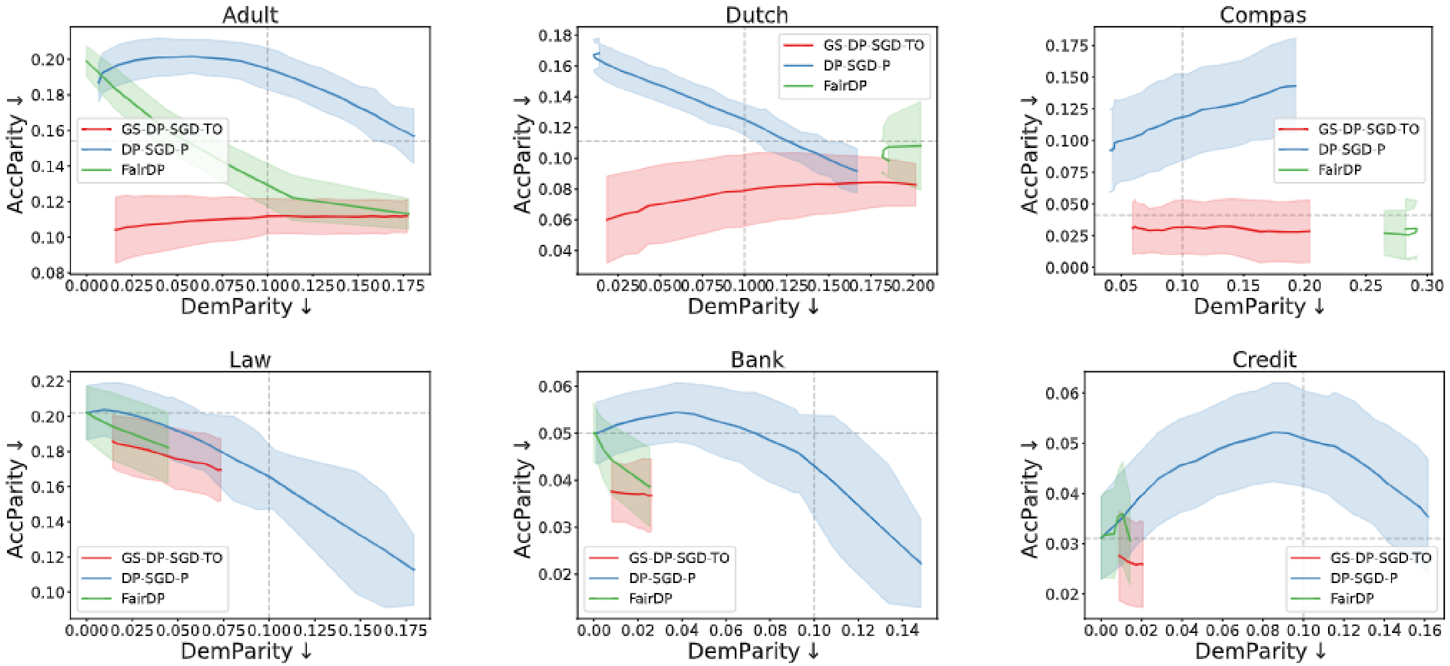
■Compare with methods of alleviating the accuracy unfairness intensified by DP.

Dataset	Method	ϵ	Acc (↑)	AccParity (↓)	DemParity (↓)
Adult	SGD		0.848 ± 0.003	0.114 ± 0.008	0.191 ± 0.007
	DP-SGD	2.654	0.789 ± 0.005	0.154 ± 0.010	0.064 ± 0.005
	DP-SGD-F	2.667	0.829 ± 0.003	0.112 ± 0.008	0.210 ± 0.005
	DP-SGD-A	2.661	0.848 ± 0.003	0.114 ± 0.007	0.191 ± 0.010
	GS-DP-SGD	2.657	0.832 ± 0.005	0.116 ± 0.010	0.199 ± 0.057
	GS-DP-SGD-TO	2.658	0.837 ± 0.005	0.108 ± 0.015	0.047 ± 0.015
Dutch	SGD		0.834 ± 0.003	0.070 ± 0.006	0.335 ± 0.016
	DP-SGD	2.269	0.793 ± 0.005	0.111 ± 0.011	0.231 ± 0.033
	DP-SGD-F	2.280	0.812 ± 0.005	0.092 ± 0.008	0.232 ± 0.024
	DP-SGD-A	2.275	0.834 ± 0.004	0.069 ± 0.006	0.332 ± 0.017
	GS-DP-SGD	2.267	0.805 ± 0.008	0.081 ± 0.015	0.208 ± 0.064
	GS-DP-SGD-TO	2.267	0.786 ± 0.006	0.070 ± 0.026	0.052 ± 0.020
Compas	SGD		0.673 ± 0.014	0.027 ± 0.015	0.291 ± 0.021
	DP-SGD	4.118	0.623 ± 0.025	0.041 ± 0.022	0.170 ± 0.057
	DP-SGD-F	4.204	0.632 ± 0.021	0.035 ± 0.019	0.184 ± 0.044
	DP-SGD-A	4.161	0.678 ± 0.013	0.024 ± 0.016	0.281 ± 0.020
	GS-DP-SGD	4.113	0.676 ± 0.011	0.024 ± 0.020	0.287 ± 0.037
	GS-DP-SGD-TO	4.113	0.668 ± 0.010	0.029 ± 0.017	0.073 ± 0.060
Law	SGD		0.898 ± 0.004	0.160 ± 0.014	0.190 ± 0.018
	DP-SGD	3.671	0.888 ± 0.004	0.202 ± 0.015	0.000 ± 0.000
	DP-SGD-F	3.694	0.888 ± 0.004	0.202 ± 0.015	0.001 ± 0.001
	DP-SGD-A	3.683	0.898 ± 0.004	0.158 ± 0.015	0.182 ± 0.017
	GS-DP-SGD	3.679	0.895 ± 0.004	0.163 ± 0.021	0.139 ± 0.053
	GS-DP-SGD-TO	3.679	0.894 ± 0.004	0.176 ± 0.017	0.049 ± 0.024
Bank	SGD		0.900 ± 0.003	0.036 ± 0.007	0.035 ± 0.004
	DP-SGD	2.831	0.884 ± 0.004	0.050 ± 0.006	0.002 ± 0.001
	DP-SGD-F	2.845	0.887 ± 0.004	0.047 ± 0.007	0.007 ± 0.003
	DP-SGD-A	2.838	0.902 ± 0.003	0.036 ± 0.006	0.037 ± 0.005
	GS-DP-SGD	2.837	0.888 ± 0.005	0.043 ± 0.013	0.048 ± 0.031
	GS-DP-SGD-TO	2.837	0.901 ± 0.004	0.037 ± 0.008	0.026 ± 0.013
Credit	SGD		0.812 ± 0.004	0.024 ± 0.010	0.030 ± 0.006
	DP-SGD	3.365	0.778 ± 0.005	0.031 ± 0.008	0.000 ± 0.000
	DP-SGD-F	3.381	0.779 ± 0.006	0.029 ± 0.008	0.005 ± 0.006
	DP-SGD-A	3.373	0.817 ± 0.005	0.023 ± 0.011	0.033 ± 0.007
	GS-DP-SGD	3.366	0.809 ± 0.004	0.031 ± 0.013	0.035 ± 0.023
	GS-DP-SGD-TO	3.366	0.821 ± 0.006	0.026 ± 0.009	0.020 ± 0.009

4.Experimental Study

SUSTech
Southern University of Science and Technology

■Compare with methods of mitigating outcome unfairness for differentially private models.

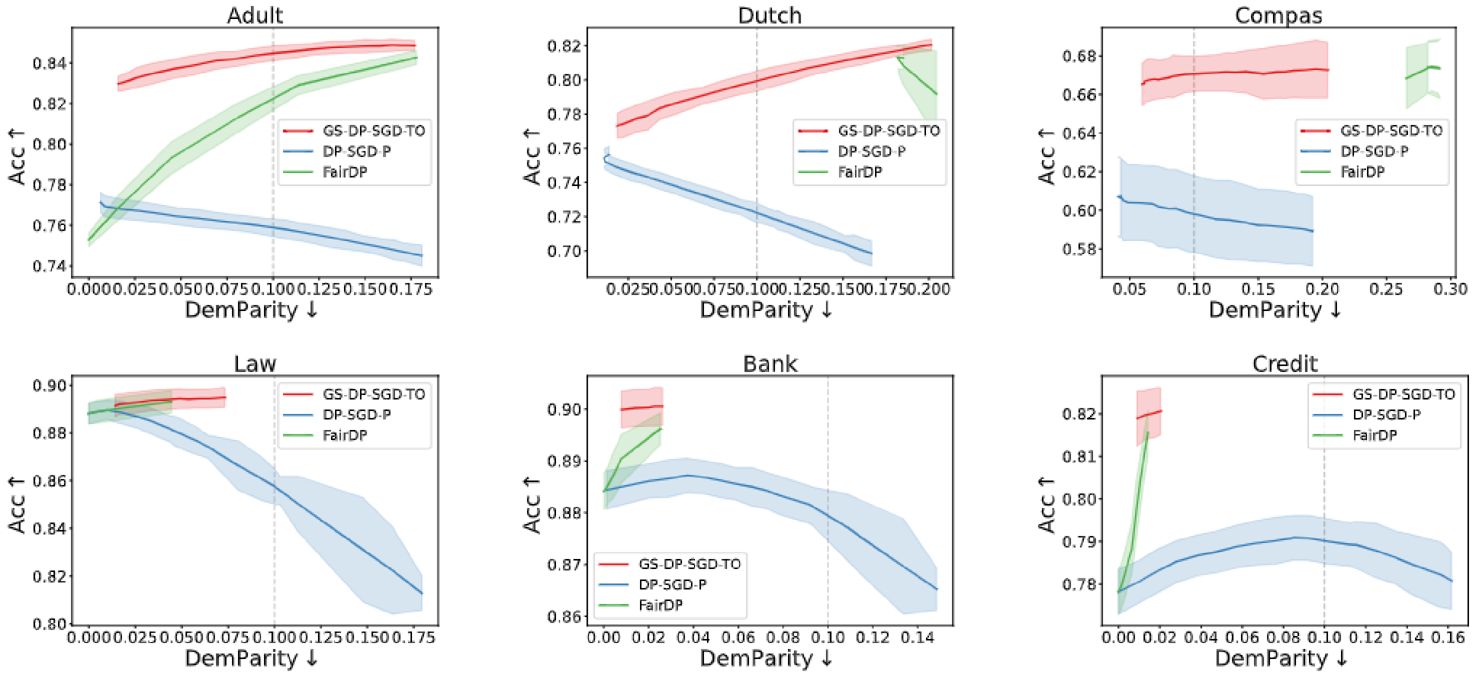


GS-DP-SGD-TO reliably and effectively reduces the AccParity value.

4.Experimental Study

SUSTech
Southern University of Science and Technology

■Compare with methods of mitigating outcome unfairness for differentially private models.



GS-DP-SGD-TO achieves the highest accuracy at the same DemParity value.

5.Conclusion

SUSTech
Southern University of Science and Technology

- Our approach uses group-specific DP-SGD during training to reduce accuracy disparity, followed by threshold optimization to improve outcome fairness.

- Extensive experiments confirm its effectiveness across datasets, balancing AccParity and DemParity with reasonable utility.

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

5.Conclusion

■Limits:

- The outcome fairness metric may exceed the predefined constraint due to threshold selection based on the validation set.

- It currently applies only to binary classification tasks and single sensitive attributes.

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Thank you!

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

References

[1] Sanyal, A., Hu, Y., Yang, F.: How unfair is private learning? In: Uncertainty in Artificial Intelligence. pp. 1738–1748. PMLR (2022)

[2] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)

[3] Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential privacy has disparate impact on model accuracy. Advances in neural information processing systems 32 (2019)

[4] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)

[5] Tran, C., Dinh, M., Fioretto, F.: Differentially private empirical risk minimization under the fairness lens. Advances in Neural Information Processing Systems 34, 27555–27565 (2021)

[6] Xu, D., Du, W., Wu, X.: Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1924–1932 (2021)

[7] Esipova, M.S., Ghomi, A.A., Luo, Y., Cresswell, J.C.: Disparate impact in differential privacy from gradient misalignment. arXiv preprint arXiv:2206.07737 (2022)

[8] Tran, K., Fioretto, F., Khalil, I., Thai, M.T., Phan, N.: Fairdp: Certified fairness with differential privacy. arXiv preprint arXiv:2305.16474 (2023)

[9] Pannekoek, M., Spigler, G.: Investigating trade-offs in utility, fairness and differential privacy in neural networks. arXiv preprint arXiv:2102.05975 (2021)

[10] Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th international conference on data mining. pp. 924–929. IEEE (2012)