Fair Infinitesimal Jackknife:
Mitigating the influence of biased training data points without refitting

Neurips 2022

SeHyun Park
January 30, 2023

Seoul National University

Outline

Introduction

- Bias mitigation algorithm can be categorized into pre-processing, in-processing, and post-processing approaches.
- Pre-processing and in-processing often require retraining a model from scratch and can be intractable in many real-world situation.

$\Rightarrow$ Post-processing approaches are the only viable option in such cases.

- Contribution:

1) Developing a post-processing fairness algorithm that improves the fairness characteristics of a pre-trained model without requiring it to be refit.
2) Developing IHVP-WoodFisher, a WoodFisher based on Inverse-Hessian Vector Product(IHVP) scheme for computing the fairness influence score.

Preliminaries

Notations

- Data set : $\mathcal{D} = \{\mathbf{z}_n = (\mathbf{x}_n, s_n, y_n)\}_{n=1}^{N}$
  - $\mathbf{x}_n \in \mathbb{R}^p$: feature
  - $s_n \in [0, 1, \cdots, k]$: sensitive attribute
  - $\mathbf{y}_n \in \mathcal{Y}$: response
- Parameter : $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^D$
- Model : $h_{\boldsymbol{\theta}}(\mathbf{x}_n) \in \mathcal{Y}$
- loss function $\ell : \Theta \times \mathcal{Y} \to \mathbb{R}$

Empirical and Weighted Risk Minimization

- Weighted risk minimization problem
  - Let $\mathbf{w} = [w_1, w_2, \cdots, w_N]^T \in \mathbb{R}^N$ be weights vector.
  - Then $\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \underset{\theta \in \Theta}{argmin} \frac{1}{N} \sum_{n=1}^{N} w_n \ell(h_{\boldsymbol{\theta}}(x_n), y_n)$
  - Setting all the weights to one, $\mathbf{1} \stackrel{def}{=} [w_1 = 1, w_2 = 1, \cdots, w_N = 1]^T$, then $\widehat{\boldsymbol{\theta}}(\mathbf{1}) = \widehat{\boldsymbol{\theta}} = \underset{\theta \in \Theta}{argmin} \frac{1}{N} \sum_{n=1}^{N} \ell(h_{\boldsymbol{\theta}}(x_n), y_n)$.
    It recovers the standard empirical risk minimization problem.

- Since $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ is a function of the weights $\mathbf{w}$, we can form a first-order Taylor approximation to it about $\mathbf{1}$:

$$\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \widehat{\boldsymbol{\theta}} + \nabla_{\mathbf{w}}\widehat{\boldsymbol{\theta}}(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{1}} (\mathbf{w} - \mathbf{1}) + \mathcal{O}\left((\mathbf{w} - \mathbf{1})^2\right)$$

- This first order Taylor approximation is often referred to as the Infinitesimal Jackknife Approximation.

- When $\widehat{\boldsymbol{\theta}}$ is a stationary point of $\frac{1}{N}\sum_{n=1}^{N}\ell(h_{\boldsymbol{\theta}}(x_n), y_n) \overset{let}{=} L(\boldsymbol{\theta})$,

$$\left.\frac{d\widehat{\boldsymbol{\theta}}(\mathrm{w})}{dw_n}\right|_{\mathrm{w}=\mathbf{1}} = -H^{-1}g_n$$

where $H \overset{\text{def}}{=} \nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, and $g_n \overset{\text{def}}{=} \nabla_{\boldsymbol{\theta}}\ell\left(y_n, h_{\boldsymbol{\theta}}\left(\mathrm{x}_n\right)\right)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$

- To measure the influence of training instance on a differentiable functional, $M(\widehat{\boldsymbol{\theta}}(\mathrm{w}), \mathrm{w})$, of $\widehat{\boldsymbol{\theta}}(\mathrm{w})$, apply chain rule to arrive at,

$$\mathcal{I}_{M,n} \overset{\text{def}}{=} \left.\frac{dM(\widehat{\boldsymbol{\theta}}(\mathrm{w}), \mathrm{w})}{dw_n}\right|_{\mathrm{w}=\mathbf{1}} = -\left.\nabla_{\widehat{\boldsymbol{\theta}}} M(\widehat{\boldsymbol{\theta}}(\mathrm{w}), \mathrm{w})\right|_{\mathrm{w}=\mathbf{1}}^{T} H^{-1}g_n$$

Fair Classification through Post-Hoc Interventions

- Develop : Post-processing fairness algorithm
- Given :
  (1) a pre-trained model $\widehat{\boldsymbol{\theta}}_{pre}$
  (2) access to the training data and a validation set
  (3) a twice differentiable loss function and a once differentiable surrogate to the fairness metric
  (4) an invertible Hessian at a local optimum of the loss

Fair Classification

- Two common fairness metrics
  1. Demographic parity(DP)
  - $\Delta\mathrm{DP}(\boldsymbol{\theta}) = |P(h_{\boldsymbol{\theta}}(X) = 1 \mid S = 1) - P(h_{\boldsymbol{\theta}}(X) = 1 \mid S = 0)|$
  2. Equality of odds(EO)
  - $\Delta\mathrm{EO}(\boldsymbol{\theta}) = \sum_{y} |P(h_{\boldsymbol{\theta}}(X) = 1 \mid S = 1, Y = y) - P(h_{\boldsymbol{\theta}}(X) = 1 \mid S = 0, Y = y)|$
- Smooth surrogate to $\Delta\mathrm{DP}$ and $\Delta\mathrm{EO}$
  - $M_{\mathcal{D}}^{\Delta\mathrm{DP}}(\boldsymbol{\theta}) = \left|\mathbb{E}_{p_{\mathcal{D}}(X=\mathsf{x}|S=1)}[h_{\boldsymbol{\theta}}(\mathsf{x})] - \mathbb{E}_{p_{\mathcal{D}}(X=\mathsf{x}|S=0)}[h_{\boldsymbol{\theta}}(\mathsf{x})]\right|$
  - $M_{\mathcal{D}}^{\Delta\mathrm{EO}}(\boldsymbol{\theta}) = \sum_{y}\left|\mathbb{E}_{p_{\mathcal{D}}(X=\mathsf{x}|S=0,Y=y)}[h_{\boldsymbol{\theta}}(\mathsf{x})] - \mathbb{E}_{p_{\mathcal{D}}(X=\mathsf{x}|S=1,Y=y)}[h_{\boldsymbol{\theta}}(\mathsf{x})]\right|$

- Assume : use a held-out validation set $\mathcal{D}_{\text{val}} = \{\mathbf{x}_n, s_n, y_n\}_{n=1}^{N_{\text{val}}}$
- Influence function for group fairness :
  - on $\Delta\mathrm{DP}$

    $$\mathcal{I}_{\Delta\mathrm{DP},n} = -\nabla_{\widehat{\boldsymbol{\theta}}} M^{\Delta\mathrm{DP}}_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}})^T H^{-1} g_n$$

  - on $\Delta\mathrm{EO}$

    $$\mathcal{I}_{\Delta\mathrm{EO},n} = -\nabla_{\widehat{\boldsymbol{\theta}}} M^{\Delta\mathrm{EO}}_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}})^T H^{-1} g_n$$

Post-Hoc Mitigation

- Post-hoc improvement of pre-trained $\widehat{\boldsymbol{\theta}}$

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\text{fair}} \stackrel{\text{def}}{=} \widehat{\boldsymbol{\theta}}\left(\mathrm{w}_{\text{fair}}\right) &= \widehat{\boldsymbol{\theta}} + \sum_{n=1}^{N} \left.\frac{d\boldsymbol{\theta}(\mathrm{w})}{dw_n}\right|_{\mathrm{w}=\mathbf{1}} \left(w_n^{\mathrm{fair}} - 1\right), \\
&= \widehat{\boldsymbol{\theta}} - \sum_{n=1}^{N} H^{-1} g_n \left(w_n^{\mathrm{fair}} - 1\right)
\end{aligned}
$$

- By searching for a weight vector $\mathrm{w}_{fair} = [w_1^{\text{fair}}, w_2^{\text{fair}}, \cdots, w_N^{\text{fair}}]^T \in \mathbb{R}^N$ such that optimizing $M^b_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}}(\mathrm{w}), \mathrm{w})$ with respect to w, $b = \Delta\mathrm{DP}$ or $\Delta\mathrm{EO}$
- However, optimizing $M(\widehat{\boldsymbol{\theta}}(\mathrm{w}), \mathrm{w})$ without any constraint on w will likely result in fair but inaccurate classifier, and the optimized weights will typically not be interpretable.
- Circumventing these issues by constraining the elements of w to be binary.

Post-Hoc Mitigation

- Let $\overline{M}^b_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{w})$ be a linearized approximation to $M^b_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{w})$ about $\mathbf{1}$

Proposition

Let $\mathbf{w}_{\text{fair}} \in \{0,1\}^N$ be a N dimensional binary vector such that its n^{th} coordinate is $w^{fair}_n = 1 - \mathbb{I}\left[\mathcal{I}_{b,n} > 0\right]$, then,

$$\mathbf{w}_{fair} = \underset{\mathbf{w}\in\{0,1\}^N}{\operatorname{argmin}} \bar{M}^b_{\mathcal{D}_{val}}(\widehat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{w}) - M^b_{\mathcal{D}_{val}}(\widehat{\boldsymbol{\theta}}(\mathbf{1}), \mathbf{1}),$$

$$\textit{and } \bar{M}^b_{\mathcal{D}_{val}}\left(\widehat{\boldsymbol{\theta}}(\mathbf{w}_{fair}), \mathbf{w}_{fair}\right) - M^b_{\mathcal{D}_{val}}(\widehat{\boldsymbol{\theta}}(\mathbf{1}), \mathbf{1}) \leq 0$$

- It follows that $M^b_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}}(\mathbf{w}_{\text{fair}}), \mathbf{w}_{\text{fair}}) \approx\leq M^b_{\mathcal{D}_{\text{val}}}(\widehat{\boldsymbol{\theta}}(\mathbf{1}), \mathbf{1})$
- Define $\mathcal{D}_{-} = \{\mathbf{z}_n |\ \mathbf{z}_n \in \mathcal{D} \text{ and } \mathcal{I}_{M,n} > 0\}$

$\Rightarrow \widehat{\boldsymbol{\theta}}_{\text{fair}} = \widehat{\boldsymbol{\theta}} + \sum_{m\in\mathcal{D}_{-}} H^{-1} g_m$

Compute Hessian-Vector Product(HVP)

- In this paper, they develop an alternative iterative procedure based on the recently proposed WoodFisher approximation.
- **WoodFisher approximation**

$$\widehat{H}_{n+1}^{-1}=\widehat{H}_{n}^{-1}-\frac{\widehat{H}_{n}^{-1}\nabla_{\theta}\ell\left(y_{n+1},h_{\boldsymbol{\theta}}\left(\mathrm{x}_{n+1}\right)\right)\nabla_{\theta}\ell\left(y_{n+1},h_{\boldsymbol{\theta}}\left(\mathrm{x}_{n+1}\right)\right)^{\top}\widehat{H}_{n}^{-1}}{N+\nabla_{\theta}\ell\left(y_{n+1},h_{\boldsymbol{\theta}}\left(\mathrm{x}_{n+1}\right)\right)^{\top}\widehat{H}_{n}^{-1}\nabla_{\theta}\ell\left(y_{n+1},h_{\boldsymbol{\theta}}\left(\mathrm{x}_{n+1}\right)\right)}$$

with $\widehat{H}_{0}^{-1}=\lambda^{-1}I_{D}$, and λ a small positive scalar.

Compute Hessian-Vector Product(HVP)

- IHVP-WoodFisher approximation

$\Rightarrow$ purpose : $H^{-1}\mathbf{v}$

Proposition

Let $\boldsymbol{o}_1 = \nabla_\theta \ell(\mathrm{z}_1), \boldsymbol{k}_1 = \mathbf{v}$, and N denote the number of training instances. The Hessian-vector product $H^{-1}\mathbf{v}$ is approximated by iterating through the IHVP-WoodFisher recurrence in under equation and computing $\boldsymbol{k}_N$.

$$\boldsymbol{o}_{n+1} = \boldsymbol{o}_n - \frac{\boldsymbol{o}_n \nabla_\theta \ell(\mathrm{z}_{n+1})^\top \boldsymbol{o}_n}{N + \nabla_\theta \ell(\mathrm{z}_{n+1})^\top \boldsymbol{o}_n}, \quad \boldsymbol{k}_{n+1} = \boldsymbol{k}_n - \frac{\boldsymbol{o}_n \nabla_\theta \ell(\mathrm{z}_{n+1})^\top \boldsymbol{k}_n}{N + \nabla_\theta \ell(\mathrm{z}_{n+1})^\top \boldsymbol{o}_n}$$

where, we use $\ell(\mathrm{z}_{n+1})$ as shorthand for $\ell(y_{n+1}, h_{\boldsymbol{\theta}}(\mathrm{x}_{n+1}))$.

Experiment

1. Adult dataset
   - Task : to predict if a person has an income above a threshold
   - Sensitive attribute : **Gender** $\in$ [Female, Male]
   - Response : **Income** $\in$ ['<=50k', '>50k']

2. ACSPublicCoverage dataset
   - Task : to predict whether an individual is covered by public health insuarance
   - Sensitive attribute : **Race** $\in$ [white, black]
   - Response : **PUBCOV(**Public health coverage) $\in [0, 1]$

Algorithm

Algorithm 1 `Fair-IJ`

Input: Pre-trained model parameters $\hat{\boldsymbol{\theta}}$, training set $\mathcal{D}$, loss function ℓ, a validation set $\mathcal{D}_{\text{val}}$ and a smooth surrogate to the fairness metric $b \in \{\Delta\text{DP}, \Delta\text{EO}\}$, $M^b_{\mathcal{D}_{\text{val}}}$.

Calculate: $\nabla_{\boldsymbol{\theta}} M(\hat{\boldsymbol{\theta}}, \mathbf{1})$ using Equation 7 or Equation 8.

Calculate: $\boldsymbol{r} = H^{-1}\nabla_{\boldsymbol{\theta}} M(\hat{\boldsymbol{\theta}}, \mathbf{1})$ by setting $\boldsymbol{k}_1 = \nabla_{\boldsymbol{\theta}} M(\hat{\boldsymbol{\theta}}, \mathbf{1})$ and iterating through Equation 14 for B iterations.

Calculate: the fairness influence $\mathcal{I}_{b,n}$ of each training instance $\mathbf{z}_n$ on $\mathcal{D}_{\text{val}}$ by computing dot product between g_n and $\boldsymbol{r}$.

Construct: the set $\mathcal{D}_-$ and denote its cardinality, $|\mathcal{D}_-| = K$.

Initialize: $\hat{\boldsymbol{\theta}}^0_{\text{fair}} := \hat{\boldsymbol{\theta}}$

for $k \in [1, \ldots, K]$ **do**

  Construct: $\mathcal{D}^k_- = \{\mathbf{z}_n \in \mathcal{D}_- \mid \mathcal{I}_{b,n} > \mathcal{I}_{b,(K-k)}\}$, where $\mathcal{I}_{b,(K-k)}$ denotes the $(K-k)^{\text{th}}$ order statistic of the influence scores $[\mathcal{I}_{b,1}, \ldots, \mathcal{I}_{b,K}]$.

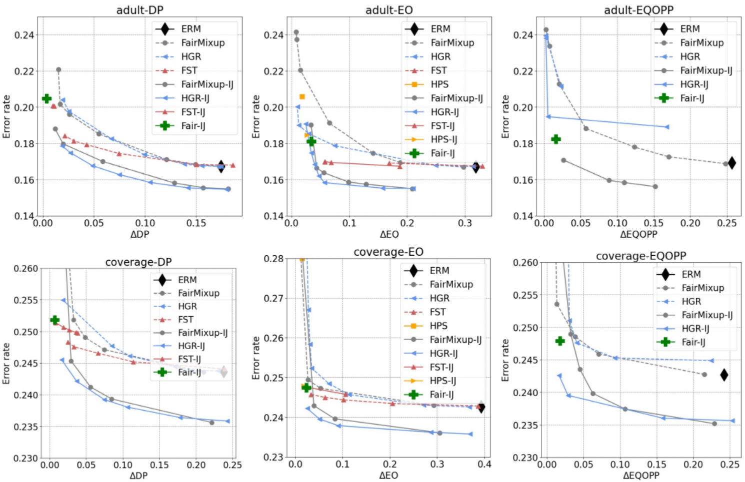
  Calculate: $\hat{\boldsymbol{\theta}}^k_{\text{fair}}$ by replacing $\mathcal{D}_-$ with using $\mathcal{D}^k_-$ in Equation 12.

  If $b_{\mathcal{D}_{\text{val}}}(\hat{\boldsymbol{\theta}}^k_{\text{fair}}) < b_{\mathcal{D}_{\text{val}}}(\hat{\boldsymbol{\theta}}^{k-1}_{\text{fair}})$ **set** $\hat{\boldsymbol{\theta}}_{\text{fair}} := \hat{\boldsymbol{\theta}}^k_{\text{fair}}$ **else set** $\hat{\boldsymbol{\theta}}_{\text{fair}} := \hat{\boldsymbol{\theta}}^{k-1}_{\text{fair}}$ and break out of the for loop.

end for

Return: fair model parameters $\hat{\boldsymbol{\theta}}_{\text{fair}}$.

Result



- ▶ Accuracy and fairness Pareto frontier for the Adult and the Coverage datasets averaged over 10 runs.
- ▶ Points closer to the bottom-left achieve the best fairness/accuracy trade-off.

End