On learning fairness and accuracy on multiple subgroups

Kim Choeun
April 15, 2025

Seoul National University

Introduction

- This work focus on the criteria of **group sufficiency**.
  - ensures that the conditional expectation of ground-truth label $\mathbb{E}[Y \mid f(X), A]$ is identical across different subgroups given the predictor's output
  - e.g., the ML algorithm is used to assess the clinic risk $\rightarrow$ $\mathbb{E}[Y \mid f(X), A = \text{black}] \gg \mathbb{E}[Y \mid f(X), A = \text{white}]$
- Aims to propose a novel principled framework for ensuring group sufficiency, as well as preserving an informative prediction with a small generalization error.
- In particular, focused on one challenge scenario : *the data includes multiple or even a large number of subgroups, some with only limited samples.*

Preliminaries

- $X \in \mathcal{X}$: input, $Y \in \{0,1\}$: label, $A \in \mathcal{A}$:sensitive attribute (scalar discrete random variable)
- $(X, Y, A) \sim \mathcal{D}(X, Y, A)$
- $f : \mathcal{X} \rightarrow [0,1]$: predictor

Group Sufficiency

A predictor f satisfies group sufficiency with respect to the sensitive attribute A if $\mathbb{E}[Y \mid f(X)] = \mathbb{E}[Y \mid f(X), A]$.

Group Sufficiency Gap

The group sufficiency gap of a predictor f is defined as

$$\mathbf{Suf}_f = \mathbb{E}_{A,X}[|\mathbb{E}[Y \mid f(X)] - \mathbb{E}[Y \mid f(X), A]|]$$

UB of Group Sufficiency Gap

a-group Bayes Predictor

The a-group Bayes predictor $f_{A=a}^{Bayes}$ is defined as

$$f_{A=a}^{Bayes}(X) = \mathbb{E}[Y \mid X, A = a]$$

Theorem

Group sufficiency gap $\mathbf{Suf}_f$ *is upper bounded by*

$$\mathbf{Suf}_f \leq 4\mathbb{E}_{A,X}[|f - f_A^{Bayes}|]$$

Specifically, if A takes finite value and follows uniform distribution with $\mathcal{D}(A = a) = 1/|\mathcal{A}|$. Then the group sufficiency gap is futher simplified as

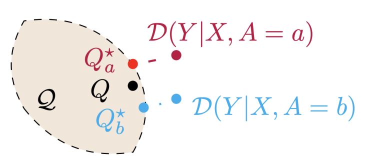
$$\mathbf{Suf}_f \leq \frac{4}{|\mathcal{A}|} \sum_a \mathbb{E}_X[|f - f_{A=a}^{Bayes}| \mid A = a]$$

This implies that using a probabilistic framework to approximate predictor $f(x) \approx \mathbb{E}(Y \mid X)$ results in both group sufficiency gap and prediction error being small. (under the assumption that f_A^{Bayes}'s are quite similar)

Principled Approach

- Considered a randomized algorithm that learns a predictive distribution Q over scoring predictors from the data.
  - the predictor is drawn from the posterior distribution. $\tilde{f} \sim Q$
  - in the inference, the predictor's output is formulated as $f(X) = \mathbb{E}_{\tilde{f} \sim Q} \tilde{f}(X)$
- In practice, we should restrict the predictive distribution Q within a distribution family $Q \in \mathcal{Q}$ such as Gaussian distribution.
- We also denote $Q_a^* \in \mathcal{Q}$ as the optimal prediction-distribution w.r.t. $A = a$ under BCE loss within $\mathcal{Q}$, that is $Q_a^* = arg\,\min_{Q_a \in \mathcal{Q}} \mathbb{E}_{\tilde{f} \sim Q_a} \mathcal{L}_a^{BCE}(\tilde{f}_a)$

Principled Approach



Corollary

The group sufficiency gap **Suf**$_f$ *in randomized algorithm w.r.t. learned predictive-distribution* Q *is upper bounded by*

$$\begin{aligned}\textbf{Suf}_f &\le \frac{4}{|\mathcal{A}|}\sum_a \mathbb{E}_X\left[|\mathbb{E}_{\tilde{f}\sim Q}\tilde{f}(x) - \mathbb{E}_{\tilde{f}\sim Q_a^*}\tilde{f}(x)| + |\mathbb{E}_{\tilde{f}\sim Q_a^*}\tilde{f}(x) - \mathbb{E}_{\tilde{f}\sim \mathcal{D}(y|x,a)}\tilde{f}(x)|\right] \\ &\le \frac{4}{|\mathcal{A}|}\sum_a [TV(Q_a^*\|Q) + TV(Q_a^*\|\mathcal{D}(y \mid x, a))] \\ &\le \frac{2\sqrt{2}}{|\mathcal{A}|}\sum_a \left[\sqrt{KL(Q_a^*\|Q)} + \sqrt{KL(Q_a^*\|\mathcal{D}(Y \mid X, A = a))}\right]\end{aligned}$$

Principled Approach

- Challenge in learning limited samples
  - $\hat{Q}_a^* = arg\min_{Q_a \in \mathcal{Q}} \mathbb{E}_{\tilde{f}_a \sim Q_a} \hat{\mathcal{L}}_a^{BCE}(\tilde{f}_a)$
  - each subgroup contains limited number of samples $\rightarrow$ overfitting

Theorem

Supposing that datasets $\{S_a\}_{a=1}^{|\mathcal{A}|}$ with $S_a = \{(x_i^a, y_i^a)\}_{i=1}^m$ are i.i.d. sampled from $\mathcal{D}(x, y \mid A = a)$, the BCE loss is upper bounded by L, $Q_a \in \mathcal{Q}$ is any learned distribution from dataset S_a and $Q \in \mathcal{Q}$ is any distribution. Then with high probability $\geq 1 - \delta$ with $\forall \delta \in (0, 1)$, we have:

$$\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a} \mathcal{L}_a^{BCE}(\tilde{f}_a) \leq \frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a} \hat{\mathcal{L}}_a^{BCE}(\tilde{f}_a) + \frac{L}{\sqrt{|\mathcal{A}| m}} \sum_a \sqrt{KL(Q_a \| Q)} + L \sqrt{\frac{\log(1/\delta)}{|\mathcal{A}| m}}$$

Bilevel Objective

$$\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a}\hat{\mathcal{L}}_a^{BCE}(\tilde{f}_a) + \frac{L}{\sqrt{|\mathcal{A}|m}}\sum_a \sqrt{KL(Q_a \| Q)} + L\sqrt{\frac{\log(1/\delta)}{|\mathcal{A}|m}}$$

From the theorem, we can construct bi-level objective as follows:

$$\min_{Q \in \mathcal{Q}} \frac{1}{|\mathcal{A}|}\sum_a KL(\bar{Q}_a^* \| Q) \quad (1)$$

$$\bar{Q}_a^* = arg \min_{Q_a \in Q}\{\mathbb{E}_{\tilde{f}_a \sim Q_a}\hat{\mathcal{L}}_a^{BCE}(\tilde{f}_a) + \lambda KL(Q_a \| Q)\}, \forall a \in \mathcal{A} \quad (2)$$

Upper level objective 1 and Lower level objective 2

Practical Implementations

- Parametric models
  - Choose the isotropic gaussian (with diagonal covariance matrix) as the distribution family $\mathcal{Q}$
  - Thus, we need to learn the parameter (θ, σ) for Q
  - For the subgroup $A = a$, we learn parameters (θ_a, σ_a) for $\bar{Q}_a^*$
  - For the single predictor $\tilde{f}$, we use parametric NN models and assume $\tilde{f}$ is parametrized by a d-dimensional vector $w \in \mathbb{R}^d$, denoted as $\tilde{f}_w$

- Gradient Estimation
  - $KL(Q_a \| Q)$: Since both Q_a and Q are factorized Gaussian, $KL(Q_a \| Q) = \frac{1}{2} \sum_{i=1}^{d} \{\log \frac{\sigma_a^2[i]}{\sigma^2[i]} + \frac{\sigma_a^2[i] + (\theta_a[i] - \theta[i])^2}{\sigma^2[i]} - 1\}$
  - $\mathbb{E}_{\tilde{f}_{w_a} \sim Q_a} \hat{\mathcal{L}}_a^{BCE}(\tilde{f}_{w_a})$: re-parametrize $w_a = \theta_a + \sigma_a \epsilon$, $\epsilon \sim \mathcal{N}(0, I) \rightarrow$
    $\nabla_{(\theta_a, \sigma_a)} \mathbb{E}_{w_a \sim N(\theta_a, \sigma_a)} \hat{\mathcal{L}}_a^{BCE}(\tilde{f}_{w_a}) = \nabla_{(\theta_a, \sigma_a)} \mathbb{E}_{\epsilon \sim N(0, I)} \hat{\mathcal{L}}_a^{BCE}(\tilde{f}_{w_a(\theta_a, \sigma_a)})$
    $\rightarrow$ approximate using Monte Carlo sampling w.r.t ϵ

Practical Implementations

- Algorithm

Algorithm 1 Fair and Informative Learning for Multiple Subgroups (FAMS)

Input: Parameters w.r.t. distribution Q:$(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, datasets $\{S_a\}$, $a \in \mathcal{A}$.
for Sampling a subset of $\{S_a\}$, where $a \in \mathcal{A}' \subseteq \mathcal{A}$ **do**
    ### Solving the lower-level ###
    Fix Q, optimizing the loss w.r.t. $Q_a = \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ through SGD for each $a \in \mathcal{A}'$
$$\mathbb{E}_{\tilde{f}_{\mathbf{w}_a} \sim Q_a} \hat{\mathcal{L}}_a^{\mathrm{BCE}}(\tilde{f}_{\mathbf{w}_a}) + \lambda \mathrm{KL}(Q_a \| Q)$$
    Obtaining the solution $\overline{Q}_a^\star$, $a \in \mathcal{A}'$.
    ### Solving the upper-level ###
    Fix $\overline{Q}_a^\star$ with $a \in \mathcal{A}'$, optimizing the loss w.r.t. Q through SGD: $\frac{1}{|\mathcal{A}'|} \sum_a \mathrm{KL}(\overline{Q}_a^\star \| Q)$
    Obtaining updated parameter $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ in Q
end for
Return: Parameter of distribution Q: $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$

- Inference
    - In the inference, use MC method to sample the weights of the NN from distribution $w \sim \mathcal{N}(\theta, \sigma^2)$, then averaging the output w.r.t. different sampled weights to approximate $f(x) = \mathbb{E}_{\tilde{f}_{w \sim Q}} \tilde{f}_w(x)$.

- Dataset : Amazon Review

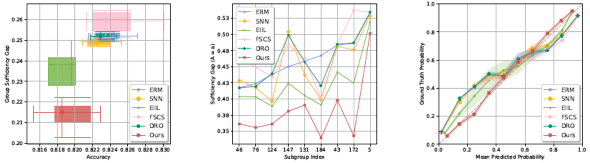
  Aim to predict the sentiment (classification) from the review.

  Each user has limited number of reviews, ranging from 75 to 400.
- The *user* is treated as a subgroup.

  Draw and fix 200 users from the original dataset, i.e., $|\mathcal{A}| = 200$.
- Adopt DistilBERT to learn the embedding with dimension $\mathbb{R}^{768}$.
- Then adopt $\tilde{f}_w$ and $\tilde{f}_{w_a}$ as the four-layer FCN, where $w \sim Q$ and $w_a \sim Q_a$.

Experiment

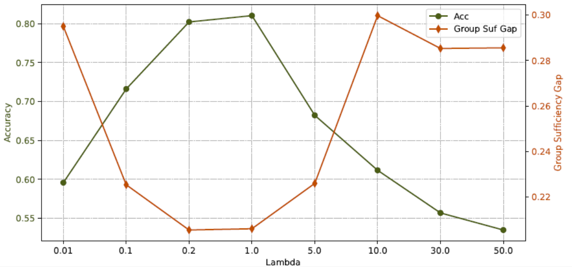
- Baselines
  - ERM : training a deep model w.o. considering the sensitive attribute
  - SNN : stochastic NN through the vanilla training from the whole dataset. find a predictive distribution Q to minimize $\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{f\sim Q}\hat{\mathcal{L}}_a^{BCE}(f)$.
  - EIIL : IRM based approach to promote the group sufficiency
  - FSCS : adopted the conditional MI constraint $I(A, Y \mid f(X))$ to promote the sufficiency
  - DRO : re-weighting approach to assign the importance of the task
- Since $f(X)$ is continuous, the group sufficiency gap is calculated by splitting the output of predictor into multiple intervals in $[0, 1]$ and computing the conditional expectation within each interval.



(a) Accuracy-$\mathbf{Suf}_f$ (b) $\mathbf{Suf}_f$ on each subgroup (c) Probability Calibration

Figure 3: Amazon Review dataset. (a) Boxplot of accuracy and group sufficiency gap $\mathbf{Suf}_f$ with 5 repeats: median, 75th percentile and minimum-maximum value. (b) Group sufficiency gap on subgroup $A = a$, which is the difference between $\mathbb{E}[Y|f(X)]$ and $\mathbb{E}[Y|f(X), A = a]$. We visualize the top-9 users' group sufficiency gap in ERM, whereas the result for all users is delegated to the Appendix. (c) Probability calibration curve over 5 repeats with mean and standard deviation. i.e $(f(X), \mathbb{E}[Y|f(X)])$. The proposed approach demonstrated a consistently improved probability calibration.

Experiment



Accuracy-$\mathbf{Suf}_f$ curve under different λ in 2

Appendix : proof of Theorem

Step 1 We first demonstrate the following Lemma, which is based on [50, 60].

Lemma E.1. *Let f be a random variable taking value in A and let $X_1, \ldots, X_l$ be l independent variables with each X_k distributed to μ_k over the set A_k. For function $g_k : A \times A_k \to [a_k, b_k]$, $k = 1, \ldots, l$. Let $\zeta_k(f) = \mathbb{E}_{X_k \sim \mu_k} g_k(f, X_k)$ for any fixed value of f. Then for any fixed distribution π on A and any λ, $\delta > 0$, the following inequality holds with high probability $1 - \delta$ over the sampling $X_1, \ldots, X_l$ for all distribution ρ over A.*

$$\mathbb{E}_{f\sim\rho} \sum_{k=1}^{l} \zeta_k(f) - \mathbb{E}_{f\sim\rho} \sum_{k=1}^{l} g_k(f, X_k) \leq \frac{1}{\lambda} \left(KL(\rho \| \pi) + \frac{\lambda^2}{8} \sum_{k=1}^{l} (b_k - a_k)^2 + \log \frac{1}{\delta} \right)$$

Step 2 Then we could use the aforementioned Lemma to demonstrate the main theorem.

Proof. We adopt the lemma for the union of the whole training samples $S = \cup_{a\in\mathcal{A}} S_a$.

We set

$$\rho = \underbrace{(Q_1 \otimes Q_2 \otimes \cdots \otimes Q_{|\mathcal{A}|})}_{|\mathcal{A}| \text{ times}} \qquad \pi = \underbrace{(Q \otimes Q \otimes \cdots \otimes Q)}_{|\mathcal{A}| \text{ times}}$$

We also set $X_k = (x_i^a, y_i^a)$, $l = |\mathcal{A}|m$, $f = (\tilde{f}_1, \ldots, \tilde{f}_a, \ldots, \tilde{f}_{|\mathcal{A}|})$, $g_k(f, X_k) = \frac{1}{|\mathcal{A}|m} \ell^{\text{BCE}}(\tilde{f}_a(x_i^a), y_i^a)$. Since we adopt the binary cross entropy loss, $a_k = 0$ and $b_k = L/(|\mathcal{A}|m)$,

then with high probability $1 - \delta$, we have:

$$\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a} \mathcal{L}_a^{\text{BCE}}(\tilde{f}_a) \leq \frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) + \frac{1}{\lambda}(\text{KL}(Q_1 \otimes \cdots \otimes Q_{|\mathcal{A}|} \| Q \otimes \cdots \otimes Q) + \log(\frac{1}{\delta})) + \frac{\lambda L}{8|\mathcal{A}|m}$$

Through the decomposition property of KL divergence, we finally have:

$$\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f} \sim Q_a} \mathcal{L}_a^{\text{BCE}}(\tilde{f}) \leq \frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f} \sim Q_a} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) + L \sqrt{\frac{1}{2|\mathcal{A}|m} (\sum_a \text{KL}(Q_a \| Q) + \log(\frac{1}{\delta}))}$$

$$\leq \frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f} \sim Q_a} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) + \frac{L}{\sqrt{|\mathcal{A}|m}} \sum_a \sqrt{\text{KL}(Q_a \| Q)} + L \sqrt{\frac{\log(1/\delta)}{|\mathcal{A}|m}}$$

□