Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 5: ~~Course Projects~~
Finish Multicalibration

Sep 25, 2026

Bias and Group Conditional Bias

Definition: A predictor f has bias β for a distribution D if

$$\mathbb{E}_{D}[f(X) - Y] = \beta(f)$$

Definition: A predictor f has group conditional bias β for group g and distribution D if

$$\mathbb{E}_{D}[f(X) - Y \mid g(X) = 1] = \beta(f, g)$$

Group Class Conditional Bias / Multiaccuracy

- $\mathcal{G}$ be a class of groups, $\mathcal{G} \subseteq 2^{\mathcal{X}}$

Definition: A predictor f has group class conditional bias α for class $\mathcal{G}$ and distribution $\mathcal{D}$ if

$$\forall g \in \mathcal{G} \quad \mathbb{E}[f(x) - y \mid g(x) = 1]^2 \cdot \mathbb{P}[g(x) = 1] \leq \alpha^2$$

$$\beta(f,g)^2 \cdot \mu(g) \leq \alpha^2$$

Reducing Group Class Bias

Definition: A predictor f has group class conditional bias α for class $\mathcal{G}$ and distribution $\mathcal{D}$ if

$$\forall g \in \mathcal{G} \quad \mathbb{E}[f(x) - y \mid g(x) = 1]^2 \cdot \mathbb{P}[g(x) = 1] \leq \alpha^2$$

$$\beta(f,g)^2 \cdot \mu(g) \leq \alpha^2$$

Lemma: If f is a predictor and g is a group and

$$f_{\beta(g)\to 0}(x) := f(x) - \beta(f,g) \cdot g(x)$$

then

$$\mathrm{MSE}(f_{\beta(g)\to 0}) = \mathrm{MSE}(f) - \beta(f,g)^2 \mu(g)$$

Proof:

$$\mathrm{MSE}(f) - \mathrm{MSE}(f_{\beta(g)\to 0}) = \mathbb{E}[(f(x)-y)^2 - (f_\beta(x)-y)^2]$$

$$= \underline{\mu \cdot \mathbb{E}[(f(x)-y)^2 - (f_\beta(x)-y)^2 \mid g(x)=1]} + \underbrace{\cancel{(1-\mu)\, \mathbb{E}[(f(x)-y)^2 - (f_\beta(x)-y)^2 \mid g(x)=0]}}_{=0}$$

$$= \mu \cdot \mathbb{E}[(f(x)-y)^2 - (f(x)-y-\beta)^2 \mid g(x)=1]$$

$$= \mu \cdot (2\beta\, \underbrace{\mathbb{E}[f(x)-y \mid g(x)=1]}_{\beta} - \beta^2) = \mu \cdot \beta^2 \quad \square$$

Reducing Group Class Bias

Definition: A predictor f has group class conditional bias α for class $\mathcal{G}$ and distribution $\mathcal{D}$ if

$$\forall g \in \mathcal{G} \quad \mathbb{E}[f(x) - Y \mid g(x)=1]^2 \cdot \mathbb{P}[g(x)=1] \leq \alpha^2$$

$$\beta(f,g)^2 \cdot \mu(g) \leq \alpha^2$$

Input: $f, \mathcal{G}, \alpha$

1. Set $f_0 = f$ and $t = 0$
2. While there is $g_t \in \mathcal{G}$ s.t.

$$\beta(f_t, g_t)^2 \mu(g_t) \geq \alpha^2$$

choose any such group

$$f_{t+1}(x) := f_t(x) - \beta(f_t, g_t) g_t(x)$$

and increment t

Output: the predictor f_t

Lemma: If f is a predictor and g is a group and

$$f_{\beta(g)\to 0}(x) := f(x) - \beta(f,g) \cdot g(x)$$

then

$$MSE(f_{\beta(g)\to 0}) = MSE(f) - \beta(f,g)^2 \mu(g)$$

Analysis: The algorithm stops after $T \leq 1/\alpha^2$ updates

Proof: ① $MSE(f_0) \leq 1$

② $MSE(f_t) \geq 0$

③ $MSE(f_t) - MSE(f_{t+1}) \geq \alpha^2$

(By lemma)

Algorithmic Aspects

<u>Input:</u> $f, \mathcal{G}, \alpha$

1. Set $f_0 = f$ and $t = 0$
2. While there is $g_t \in \mathcal{G}$ s.t.

   $$\beta(f_t, g_t)^2 \mu(g_t) \geq \alpha^2$$

   choose any such group

   $$f_{t+1}(x) := f_t(x) - \beta(f_t, g_t) g_t(x)$$

   and increment t

<u>Output:</u> the predictor f_t

How do I implement this?

- Naive search over $\mathcal{G}$ takes $|\mathcal{G}|$ time
- Can write this step as an optimization

$$g_t = \arg\max_{g \in \mathcal{G}} \beta(f, g)^2 \mu(g)$$

$$= \arg\max_{g \in \mathcal{G}} \mathbb{E}[f(X) - Y \mid g(X) = 1]^2 \, \mathbb{P}[g(X) = 1]$$

$$= \arg\max_{g \in \mathcal{G}} \underbrace{\mathbb{E}[(f(X) - Y) g(X)]^2}_{\text{look at this}} / \mathbb{P}[g(X) = 1]$$

Algorithmic Aspects

or $\mathbb{E}[(Y-f(x))g(x)]$

Want to solve $\underset{g \in G}{\operatorname{argmax}} \ \mathbb{E}[(f(x)-Y)g(x)]$

training examples

$$(x,y) \longmapsto (x, z) \qquad z = \tfrac{1}{2} + \tfrac{1}{2}(f(x)-y)$$

$$(\vec{x}_i, 1) \quad f(x) = .8 \longmapsto (\vec{x}, .4)$$

Try to learn g that predicts z as well as you can

$$\boxed{\underset{g \in G}{\operatorname{argmin}} \ \mathbb{E}[(g(x)-z)^2]} = \underset{g \in G}{\operatorname{argmax}} \ \mathbb{E}[(f(x)-Y)g(x)]$$

Assume $g(x) \in \{0,1\}$