Paper: Fair Mixup: Fairness via Interpolation [1]

Jiaxin Liu

Group Reading

June 28, 2021

Overview

Fairness

- Fairness in ML.

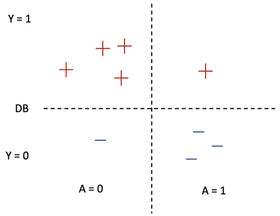


Figure: Fairness example.

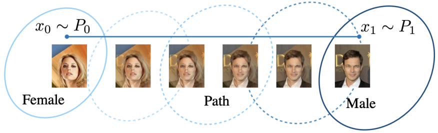


Figure: Interpolation example.

- Optimization problem:

$$\min_f \mathbb{E}_{(x,y)\sim P}[l(f(x),y)] + \lambda \cdot \text{fairness measurements}$$

Fairness measurements

- **For a binary classification setup**
  - Inputs $X \in \mathscr{X} \subset \mathbb{R}^d$.
  - Labels $Y \in \mathscr{Y} = \{0, 1\}$
  - Sensitive attribute $A \in \{0, 1\}$.
  - $\hat{Y} \in [0, 1]$ from model $f : \mathbb{R}^d \to [0, 1]$

Fairness measurements

- **For a binary classification setup**
  - Inputs $X \in \mathscr{X} \subset \mathbb{R}^d$.
  - Labels $Y \in \mathscr{Y} = \{0, 1\}$
  - Sensitive attribute $A \in \{0, 1\}$.
  - $\hat{Y} \in [0, 1]$ from model $f : \mathbb{R}^d \to [0, 1]$
- **Demographic parity (DP)**
  - $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$
  - requires $\hat{Y}$ to be independent of A.
- **Equalized odds (EO)**
  - $P(\hat{Y}|A = 1, Y = y) = P(\hat{Y}|A = 0, Y = y)$ for $y \in \{0, 1\}$
  - requires $\hat{Y}$ and A to be conditionally independent w.r.t. Y.

Metrics

- $\mathbf{\Delta DP(f)} = |\mathbb{E}_{x\sim P_0} f(x) - \mathbb{E}_{x\sim P_1} f(x)|$
- $\mathbf{\Delta EO(f)} = \sum_{y\in\{0,1\}} |\mathbb{E}_{x\sim P_0^y} f(x) - \mathbb{E}_{x\sim P_1^y} f(x)|$

where $P_a = P(\cdot|A=a)$ and $P_a^y = P(\cdot|A=a, Y=y), a, y \in \{0,1\}$.

Gap Regularization

$$\min_f \mathbb{E}_{(x,y)\sim P}[l(f(x),y)] + \lambda \Delta DP(f)$$

where l is the classification loss.

Data augmentation

- Mixup [3]: generate samples via convex combinations of pairs of examples.
  - $z_i, z_j \in \mathbb{R}^d$
  - virtual samples $tz_i + (1-t)z_j$ for $t \in [0, 1]$.

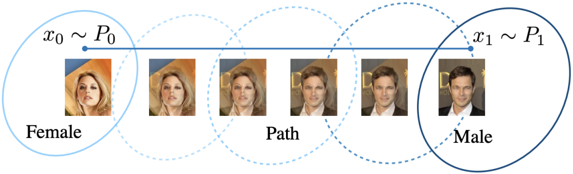


Figure: Example for interpolation.

Dynamic Formulation of fairness

- Interpolator $T(x_0, x_1, t)$ between x_0 and x_1 based on step t.
  - $T(x_0, x_1, 0) = x_0$
  - $T(x_0, x_1, 1) = x_1$
- $\mathbf{\Delta DP(f)} = |\mathbb{E}_{x \sim P_0} f(x) - \mathbb{E}_{x \sim P_1} f(x)|$

Dynamic Formulation of fairness

- Interpolator $T(x_0, x_1, t)$ between x_0 and x_1 based on step t.
  - $T(x_0, x_1, 0) = x_0$
  - $T(x_0, x_1, 1) = x_1$
- $\mathbf{\Delta DP(f)} = |\mathbb{E}_{x\sim P_0} f(x) - \mathbb{E}_{x\sim P_1} f(x)|$

Lemma 1.

Let $T : \mathscr{X}^2 \times [0,1] \to \mathscr{X}$ be a function continuously differentiable w.r.t. t such that $T(x_0, x_1, 0) = x_0$ and $T(x_0, x_1, 1) = x_1$. For any differentiable function f, we have

$$\Delta\mathrm{DP}(f) = \left| \int_0^1 \frac{d}{dt} \int f(T(x_0, x_1, t)) dP_0(x_0) dP_1(x_1) dt \right| =: \left| \int_0^1 \frac{d}{dt} \mu_f(t) dt \right|$$

where we define $\mu_f(t) = \mathbb{E}_{x_0\sim P_0, x_1\sim P_1} f(T(x_0, x_1, t))$, the expected output of f w.r.t. $T(x_0, x_1, t)$.

-

Smoothness Regularization

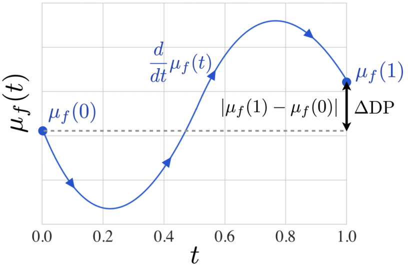


Figure: $\mu_f(t) - t$ figure.

- $\mu_f(t) = \mathbb{E}_{x_0 \sim P_0, x_1 \sim P_1} f(T(x_0, x_1, t))$

Smoothness Regularizer

$$R_T(f) = \int_0^1 \left| \frac{d}{dt}\mu_f(t) \right| dt \geq \Delta\mathrm{DP}(f)$$

Fair Mixup

Smoothness Regularizer

$$R_T(f) = \int_0^1 \left| \frac{d}{dt} \mu_f(t) \right| dt \geq \Delta\mathrm{DP}(f)$$

Input Mixup

Smoothness regularizer:

$$R_{\mathrm{mixup}}^{\mathrm{DP}}(f) = \int_0^1 \left| \int < \nabla_x f(tx_0 + (1-t)x_1), x_0 - x_1 > dP_0(x_0) dP_1(x_1) \right| dt.$$

Manifold Mixup

- Manifold Mixup [2]
  - Compositional hypothesis $f \circ g$
  - feature encoder $g : \mathscr{X} \to \mathscr{Z}$
  - predictor $f : \mathscr{Z} \to \mathscr{Y}$
-

$$
\begin{aligned}
\Delta \mathrm{DP}(f \circ g) &= \left| \int_0^1 \frac{d}{dt} \int f(T(x_0, x_1, t)) dP_0(x_0) dP_1(x_1) dt \right| \\
&= \left| \int_0^1 \frac{d}{dt} \int f(tg(x_0) + (1-t)g(x_1)) dP_0(x_0) dP_1(x_1) dt \right|
\end{aligned}
$$

Smoothness Regularizer

Input Mixup

Smoothness regularizer:

$$R^{\mathrm{DP}}_{\mathrm{mixup}}(f) = \int_0^1 \left| \int < \nabla_x f(tx_0 + (1-t)x_1), x_0 - x_1 > dP_0(x_0)dP_1(x_1) \right| dt.$$

Manifold Mixup

Smoothness regularizer:

$$R^{\mathrm{DP}}_{\mathrm{mixup}}(f)$$
$$= \int_0^1 \left| \int < \nabla_z f(tg(x_0) + (1-t)g(x_1)), g(x_0) - g(x_1) > dP_0(x_0)dP_1(x_1) \right| dt.$$

Equalized Odds (EO)

- $\mathbf{\Delta EO(f)} = \sum_{y\in\{0,1\}} |\mathbb{E}_{x\sim P_0^y} f(x) - \mathbb{E}_{x\sim P_1^y} f(x)|$
- $\mathbf{\Delta EO}(f) = \sum_{y\in\{0,1\}} \left| \int_0^1 \frac{d}{dt} \int f(T(x_0, x_1, t)) dP_0^y(x_0) dP_1^y(x_1) dt \right|$

Input Mixup

Smoothness regularizer:

$$R_{\text{mixup}}^{\text{EO}}(f)$$
$$= \sum_{y\in\{0,1\}} \int_0^1 \left| \int < \nabla_x f(tx_0 + (1-t)x_1), x_0 - x_1 > dP_0^y(x_0) dP_1^y(x_1) \right| dt.$$

Optimization Problem

Fair Mixup:

$$\min_f \mathbb{E}_{(x,y)\sim P}[l(f(x), y)] + \lambda R_{\text{mixup}}(f)$$

Task 1: predict income$\sim \$50k$; attribute: gender.

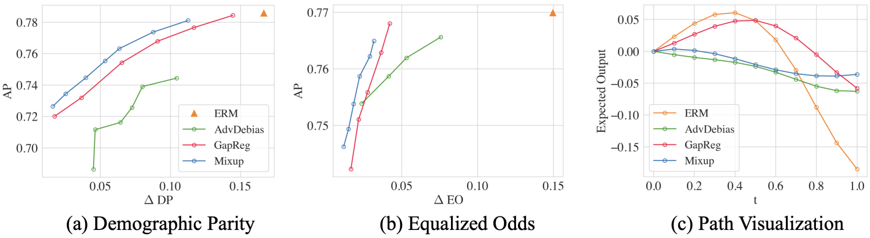


Figure: Adult Dataset. Task 1

Experiments

Task 2: celebrity faces 2.1 attractive, 2.2 smile, 2.3 wavy hair. Attribute: gender.

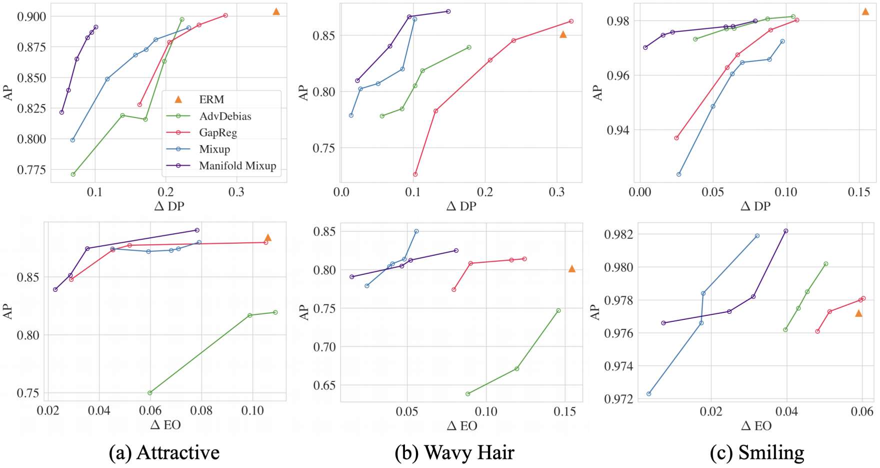
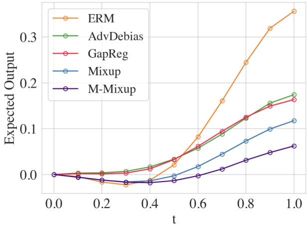
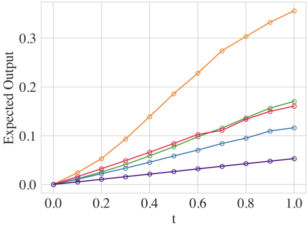


Figure: CelebA Dataset. Task 2

Task 2: Visualization of calibrated paths



(a) Input Path



(b) Latent Path

Figure: CelebA Dataset. Task 2.1 attractive classification.