Fair Machine Learning Algorithms Comparison

Group 4

LFR Model:

Intuition: Find a sanitized intermediate representation of X that

- Keep as much as possible information about attributes
- Remove information regarding individual membership of protected group.

Goals:

Find a good prototype variable Z such that

- Mapping X to Z
- **Statistical Parity:** Probability of mapping X in protected group to Z equals probability of mapping X in unprotected group to Z

$$P(Z = k | x^+ \in \mathbb{X}^+) = P(Z = k | x^- \in \mathbb{X}^-) \ for \ all \ k$$

- Preserve information in X as much as possible
- Mapping Z to Y

How to achieve Statistical Parity in LFR:

- Parameters: Vk (location of prototype)
  Wk(mapping prototypes to Y)

Soft max

$$P(Z=k|\mathbf{x}) = \exp(-d(\mathbf{x},\mathbf{v}_k)) / \sum_{j=1}^{K} \exp(-d(\mathbf{x},\mathbf{v}_j))$$

- Define probability of mapping Xn to Z:

$$M_{n,k} = P(Z=k|\mathbf{x}_n) \quad \forall n,k$$

Statistical Parity

$$M_k^+ = M_k^-, \forall k$$

$$M_k^+ = \mathbb{E}_{\mathbf{x}\in X^+} P(Z=k|\mathbf{x}) = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k}$$

Objective function: $L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$

- Constraint to achieve Statistical Parity

$$L_z = \sum_{k=1}^{K} |M_k^+ - M_k^-|$$

- Constraint to preserve information left in new representation of X

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \quad \hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k$$

- Constraint to assure predictions' accuracy

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)$$

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k$$

PR Model:

Fairness-aware Classifier with Prejudice Remover Regularizer

- **Three causes of unfairness:**
  Prejudice
  Underestimation
  Negative Legacy
- **Three types of prejudice:**
  Direct prejudice
  Indirect prejudice
  Latent prejudice
- **Prejudice Removal Techniques**
  Reduce indirect prejudice
  Implemented as a regularizer

PR Model:

Tune parameters by maximizing the log-likelihood

$$\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) = \sum_{(y_i,\mathbf{x}_i,s_i)\in\mathcal{D}} \ln \mathcal{M}[y_i|\mathbf{x}_i, s_i;\boldsymbol{\Theta}]$$

Two Regularizers(L2 regularizer and prejudice remover regularizer) added to the objective function

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Use logistic regression model as prediction model

$$\mathcal{M}[y|\mathbf{x}, s;\boldsymbol{\Theta}] = y\sigma(\mathbf{x}^\top \mathbf{w}_s) + (1-y)(1-\sigma(\mathbf{x}^\top \mathbf{w}_s))$$

PR Model:

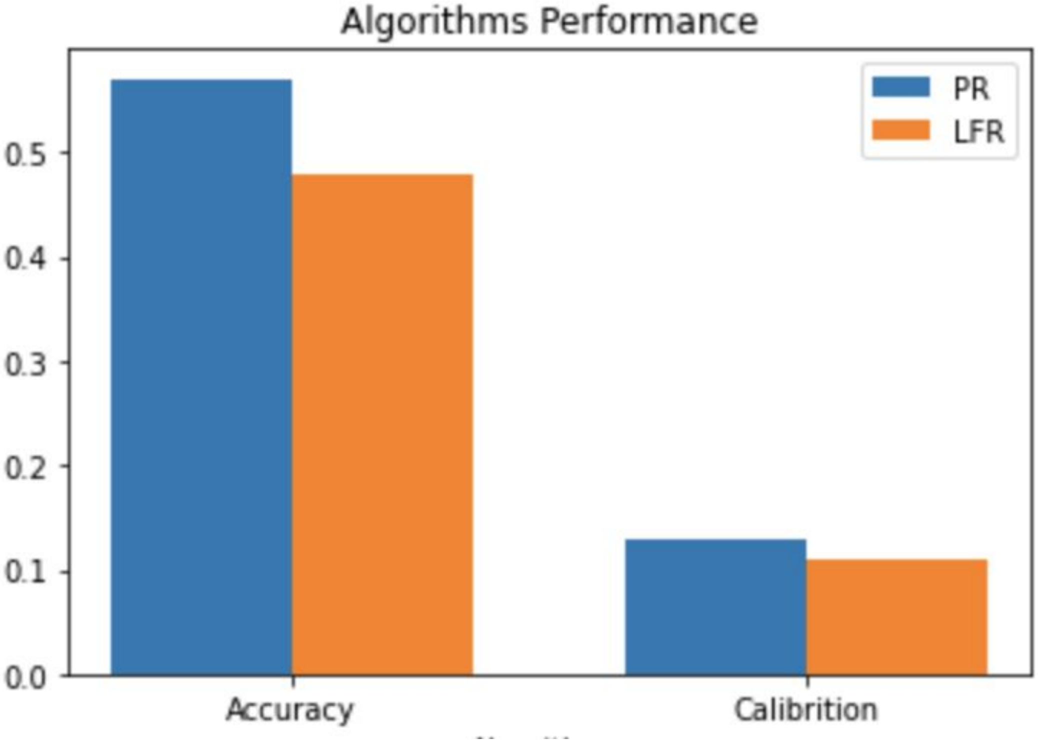
Prejudice Removal Regularizer $\mathrm{R}_{\mathsf{PR}}(\mathcal{D}, \boldsymbol{\Theta})$

$$\sum_{(\mathbf{x}_i, s_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \mathcal{M}[y | \mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y | s_i]}{\hat{\Pr}[y]}$$

The final objective function

$$\sum_{(y_i, \mathbf{x}_i, s_i)} \ln \mathcal{M}[y_i | \mathbf{x}_i, s_i; \boldsymbol{\Theta}] + \eta \, \mathrm{R}_{\mathsf{PR}}(\mathcal{D}, \boldsymbol{\Theta}) + \frac{\lambda}{2} \sum_{s \in \mathcal{S}} \|\mathbf{w}_s\|_2^2$$

Model result and Algorithms Comparison



- PR algorithm have higher accuracy than LFR algorithm
- LFR have lower calibration than PR Algorithm.

Thank you!