Optimized Pre-Processing for Discrimination Prevention

Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, Kush R. Varshney

Presented by:
Shorya Consul
Mónica Ribero

Introduction

- Motivated to deal with *disparate impact*
    - Focus on preprocessing by modifying training data.

- Notion of fairness addressed: ***group fairness*** (similar outcomes for all groups) and ***individual fairness,*** (similar individuals treated similarly irrespective of group)

- Additionally, address discrimination control/utility of data trade-off.

Formulation

- Training data: $\{(D_i, X_i, Y_i)\}_{i=1}^{n} \sim p_{D,X,Y}$

- Objective: Find $p_{\hat{X},\hat{Y}|D,X,Y}$ to obtain $\{(D_i, \hat{X}_i, \hat{Y}_i)\}_{i=1}^{n}$ that satisfies **discrimination** and **distortion** control, but maintains model and data **utility.**

I. Discrimination Control

$O(|\mathcal{D}|)$ *constraints*

- Limit dependence of $\{\hat{X}_i, \hat{Y}_i)\}_{I=1}^{n}$ on protected variables D
- Approach:
  a. Make $p_{\hat{Y}|D}$ close to a target distribution p_{Y_T} : $J(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)) \leq \epsilon_{y,d} \quad \forall d, y$

  b. Make the conditional probability for any two values of D :

$$J(p_{\hat{Y}|D}(y|d_1), p_{\hat{Y}|D}(y|d_1)) \leq \epsilon_{y,d_1,d_2} \quad \forall d_1, d_2, y$$

E.g. J(p,q) = |p/q -1|

$O(|\mathcal{D}|^2)$ *constraints*

II. Distortion Control

- Avoid large changes (e.g. low credit score mapped to very high credit score)

$$\mathbb{E}[\delta((x, y), (\hat{X}, \hat{Y}))|d, x, y] \leq c_{d,x,y} \quad \forall d, x, y$$

Distortion metric:

E.g. Binary-valued:
Desirables and non-desirable mappings

Conditional expectation to guarantee low distortion even for individuals with low probability.

III. Utility Preservation

- Model learned under new dataset is not too different from one learned from original dataset.

$$\Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$$

Dissimilarity metric

E.g. total variation distance:

$$\Delta(p,q) = \frac{1}{2}\sum_{x}|p_X(x) - q_X(x)|$$

Optimization formulation

Utility

$$\min_{p_{\hat{X},\hat{Y}|X,Y,Z}} \quad \Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$$

Discrimination

$$\text{s.t.} \quad J(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)) \leq \epsilon_{y,d}$$

Distortion

$$\mathbb{E}[\delta((x,y),(\hat{X},\hat{Y}))|d,x,y] \leq c_{d,x,y} \quad \forall d,x,y$$

$$p_{\hat{X},\hat{Y}|X,Y,Z} \quad \text{is a valid distribution}$$

Generalizability of Discrimination Control

Q. Do the same guarantees apply to unseen/test data?

Assumption: Model approximates conditional distribution well.

If model allowed to depend on D,

$$p_{\tilde{Y}|D}(\tilde{y}|d) = \sum_{\hat{x}} p_{\tilde{Y}|\hat{X},D}(\tilde{y}|\hat{x},d) p_{\hat{X}|D}(\hat{x}|d) \approx \sum_{\hat{x}} p_{\hat{Y}|\hat{X},D}(\tilde{y}|\hat{x},d) p_{\hat{X}|D}(\hat{x}|d) = p_{\hat{Y}|D}(\tilde{y}|d).$$

Output of model

Mapping of Y

Generalizability of Discrimination Control

If model cannot depend on D, $$p_{\tilde{Y}|D}(\tilde{y}|d) \approx \sum_{\hat{x}} p_{\hat{Y}|\hat{X}}(\tilde{y}|\hat{x}) p_{\hat{X}|D}(\hat{x}|d),$$

- Not generally equal to $p_{\hat{Y}|D}(y|d)$
  - More difficult to control
- Guarantees can be preserved with Markov assumption but problem loses convexity

$$p_{\hat{Y}|\hat{X},D} = p_{\hat{Y}|\hat{X}}$$

- Also shown to be robust to mismatch in distributions of training and test data

Learning Fair Representations (Zemel, et al, 2013)

- Idea: Find randomized mapping to prototypes in input space
- Mapping hopes to “lose” information pertaining to membership to a protected group
- Also aims to achieve group and individual fairness

Metrics

- Second formulation of discrimination control with *J(p,q)* = |*p/q* -*1*|
- Discrimination:

$$\max_{d,d' \in \mathcal{D}} J(p_{\tilde{Y}|D}(1|d), p_{\tilde{Y}|D}(1|d'))$$

- Two levels of discrimination control, ϵ = {0.05,0.1}

$$J\left(p_{\hat{Y}|D}(y|d_1), p_{\hat{Y}|D}(y|d_2)\right) \leq \epsilon_{y,d_1,d_2} \; \forall \; d_1, d_2 \in \mathcal{D}, y \in \{0,1\}.$$

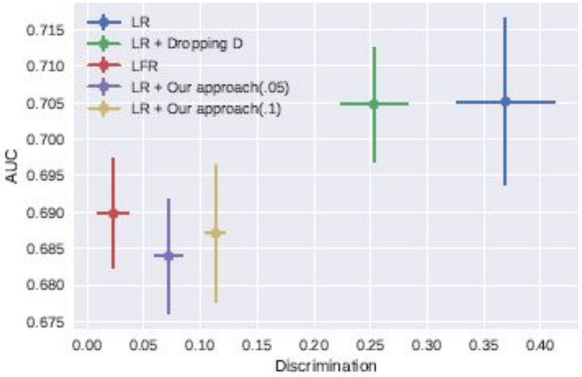
- AUC

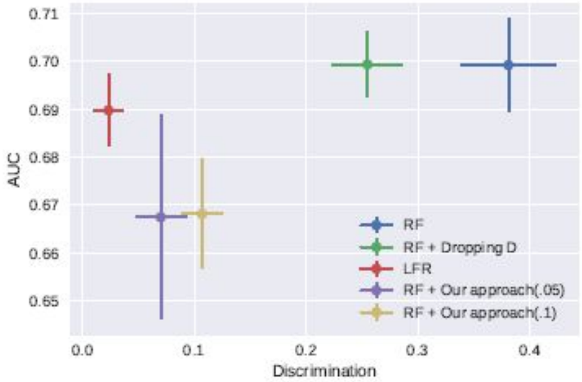
Results - COMPAS

- *X*
  - Severity of charge
  - No. of prior crimes
  - Age category

- ***Y*:** If person reoffended
- ***D*:** Race

Example: Jump of more than one age category penalized by 10^4 in distortion





Results - COMPAS

Marginal outcome distributions before and after transformation

Table 2: Dependence of the outcome variable on the discrimination variable before and after the proposed transformation. F and M indicate Female and Male, and A-A, and C indicate African-American and Caucasian.

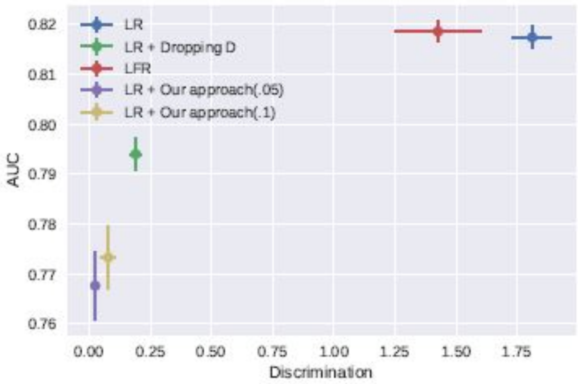
D	Before transformation		After transformation	
(gender, race)	$p_{Y\|D}(0\|d)$	$p_{Y\|D}(1\|d)$	$p_{\hat{Y}\|D}(0\|d)$	$p_{\hat{Y}\|D}(1\|d)$
F, A-A	0.607	0.393	0.607	0.393
F, C	0.633	0.367	0.633	0.367
M, A-A	0.407	0.593	0.596	0.404
M, C	0.570	0.430	0.596	0.404

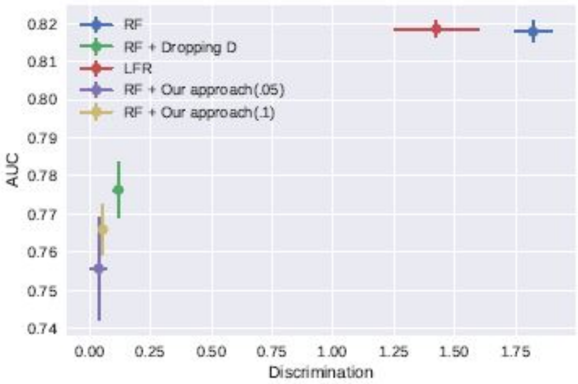
Results - UCI Adult

- *X*
  - Age (in decades)
  - Education (in years)

- *Y:* Income (binary)
- *D:* Gender

Example: δ = 1 if income decreases, age unchanged and education increases by up to a year





Conclusions

- Present pre-processing framework to mitigate disparate impact when training models
  - Strives for both group and individual fairness
- Objective formulated as a convex optimization problem
  - Solved to get mapping
- Results are not conclusive - is the proposed approach better?