Fair Bayes-Optimal Classifiers under Predictive Parity

Xianli Zeng, Edgar Dobriban, Guang Cheng

January 30, 2023

Reviewer : Park Seok Hoon , Seoul National University IDEA LAB

Table of Contents

Table of Contents

Contribution

1. They identify a sufficient condition under which all fair Bayes-optimal classifiers are GWTR
2. They proposed the FairBayes-DPP algorithm for binary fair classification.

Table of Contents

Preliminaries

- $X \in \mathcal{X}$: Feature data
- $A \in \mathcal{A}$: protected variable
- $Y \in \{0, 1\}$: ground truth label
- $\eta_a(x) = P(Y = 1 | X = x, A = a)$

Definition of Randomized Classifier

A randomized classifier is a measurable function $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$, indicating the probability of predicting $\hat{Y} = 1$ when observing $X = x$ and $A = a$. We denote $\hat{Y}_f = \hat{Y}_f(x, a)$ the prediction induced by the classifier f

Definition of GWTR classifier

A classifier f is a GWTR if there are constants $t_a, a \in \mathcal{A}$, and functions $r_a : \mathcal{X} \to [0,1], a \in \mathcal{A}$ such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$

$$f(x,a) = I(\eta_a(x) > t_a) + r_a(x) I(\eta_a(x) = t_a)$$

Definition of Predictive parity and DPP

A classifier f satisfies predictive parity if for all $a \in \mathcal{A}$,

$$P(Y = 1|\hat{Y}_f = 1, A = a) = P(Y = 1|\hat{Y}_f = 1)$$

Also,

$DPP(f) = \sum_{a \in \mathcal{A}} |P(Y = 1|\hat{Y}_f = 1, A = a) - P(Y = 1|\hat{Y}_f = 1)|$

Definition of cost sensitive zero-one risk

When the predictive parity is commonly considered, the false positives are particularly harmful. So, we consider cost-sensitive classfication.

For a cost parameter $c \in [0, 1]$, the cost-sensitive zero-one risk of the classifier f is as belows

$$R_c(f) = cP(\hat{Y}_f = 1, Y = 0) + (1 - c)P(\hat{Y}_f = 0, Y = 1)$$

Main contribution in paper

In this paper, they proposed fair Bayes-optimal classifier that satisfied Predictive Parity.

$$f^*_{PPV} = \text{argmin}_{f:DPP(f)=0} R_c(f)$$

Table of Contents

Condition

Condition (*)

$$min_{a\in\mathcal{A}}P(Y=1|\eta_a(X)\geq c, A=a)\geq max_{a\in\mathcal{A}}P(Y=1|A=a)$$

: the performances of different groups vary only moderately

Theorem 1 - Main

Consider the cost-sensitive 0-1 risk with cost parameter c. If the condition (*) holds, then all fair Bayes-optimal classifiers under predictive parity are GWTR. Thus, for all f^*_{PPV}, there are $(t^*_a)_{a=1}^{|\mathcal{A}|}$ and functions $r^*_a(x) : \mathcal{X} \to [0,1]$ such that, for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$f^*_{PPV}(x, a) = I(\eta_a(x) > t^*_a) + r^*_a(x) I(\eta_a(x) = t^*_a)$$

Table of Contents

- The DPP constraint is non-convex with respect to the model $\hat{\eta}_a(x)$ parameters.
- In such cases, incorporating fairness constraints as a penalty in the training objective may be hard due to potential local minima.
- So, they consider post-processing algorithms.

Step 1

- They apply standard ML algorithms to learn the feature and group-conditional label probabilities $\eta_a(X)$ based on the whole datasets.

- L : loss function, $\mathcal{G} = \{g_\theta, \theta \in \Theta\}$

- $\hat{\eta}_a(X) = g_{\hat{\theta}}(x, a)$
  where $\hat{\theta} \in \text{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, g_\theta(x_i, a_i))$

Step 2

- If the emirical version of Condition (*) holds,

  Find threshold $t_a, a \in \mathcal{A}$ which minimize cost-sensitive risk and satisfy sample predictive parity by using gridsearch.

Table of Contents

Data

They sampled 70%, 50% and 30% as training, validation and test dataset with replacement.

1. Adult
   : protected variable : gender
2. Compas
   : protected variable : race
3. CelebA
   : protected variable : gender

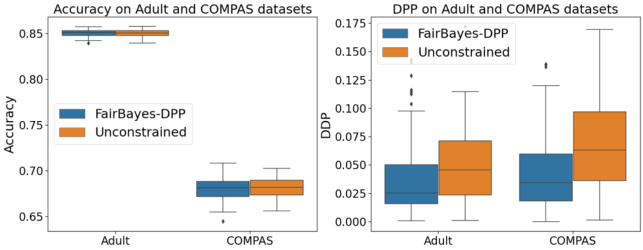
Setting

- They experiments many times by sampling train,test and validation datsets.
- cost parameter $c = 0.5$

Adult and Compas

- The conditional probability η is learned via a three-layer MLP with 32 hidden neurons per layer.

- Batch size of Adult : 512, Batch size of Compas : 2048



CelebA

- The conditional probability η is learned via a Resnet50.

Attributes	Per-attribute Accuracy		Per-attribute DPP	
	FairBayes-DPP	Unconstrained	FairBayes-DPP	Unconstrained
Arched Eyebrows	0.838(0.003)	0.838(0.003)	0.027(0.015)	0.099(0.041)
Attractive	0.825(0.002)	0.826(0.003)	0.075(0.011)	0.169(0.016)
Bags Under Eyes	0.853(0.002)	0.852(0.002)	0.024(0.015)	0.056(0.034)
Bangs	0.959(0.001)	0.959(0.001)	0.007(0.007)	0.069(0.029)
Big Lips	0.706(0.002)	0.717(0.003)	0.023(0.015)	0.115(0.027)
Big Nose	0.845(0.002)	0.847(0.003)	0.083(0.020)	0.145(0.023)