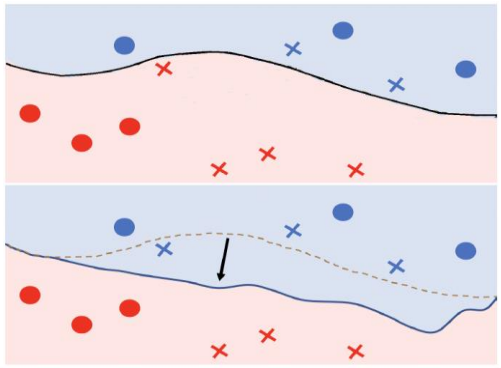
FaiR-N: Fair and Robust Neural Networks for Structured Data

Shubham Sharma, Alan Gee, David Paydarfar, Joydeep Ghosh
University of Texas at Austin

Introduction

- FaiR-N uses a novel distance to the boundary formulation in order to:
  - reduce the disparity in the average ability of recourse (i.e. the change needed to get a positive outcome) between individuals in each protected group
  - increase the average distance of data points to the boundary to promote adversarial robustness.



FaiR-N framework

- Given a binary classifier with a decision boundary, let $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ be an input vector, let $\boldsymbol{s(x)}$ be a stratification function that partitions the input dataset into k groups based on one or more protected attributes.
- The fairness loss is:

$$\mathcal{L}_{fairness} = \left| \mathop{\mathbb{E}}_{\mathbf{x}|\, s(\boldsymbol{x})=a} [d(\mathbf{x}, \mathcal{B})] - \mathop{\mathbb{E}}_{\mathbf{x}|\, s(\boldsymbol{x})=b} [d(\mathbf{x}, \mathcal{B})] \right|$$

- The robustness index:

$$\mathcal{I}_{robust} = \mathop{\mathbb{E}}_{\mathbf{x}} [d(\mathbf{x}, \mathcal{B})]$$

- The overall objective is:

$$\mathcal{L}_{overall} = \mathcal{L}_{cross} + \lambda_F \cdot \mathcal{L}_{fairness} + \lambda_R \cdot 1/\mathcal{I}_{robust}$$

Distance to the Boundary

- Let $f_0(\boldsymbol{x})$ and $f_1(\boldsymbol{x})$ be outputs of the softmax layer of a neural network. The decision boundary is defined as:

$$\mathcal{B} = \{\mathbf{x} | f_0(\mathbf{x}) = f_1(\mathbf{x})\}$$

- The distance to the boundary has been approximated as:

$$d(\mathbf{x}, \mathcal{B}) = \frac{|f_0(\mathbf{x}) - f_1(\mathbf{x})|}{||\nabla_{\mathbf{x}} f_0(\mathbf{x}) - \nabla_{\mathbf{x}} f_1(\mathbf{x})||_2}$$

- However this requires slow Hessian computation. We propose a new formulation:

Let $g_0(\boldsymbol{x})$ and $g_1(\boldsymbol{x})$ represent the inputs (i.e. logits) to the softmax function. We show that the distance can be approximated as:

$$\hat{d}(\mathbf{x}, \mathcal{B}) = |g_0(\mathbf{x}) - g_1(\mathbf{x})|$$

- Theorem: The relation between d $(\boldsymbol{x}, \boldsymbol{\mathcal{B}})$ and $\hat{d}$ $(\boldsymbol{x}, \boldsymbol{\mathcal{B}})$ is monotonic and is expressed as:

$$d(\mathbf{x}, \mathcal{B}) = \frac{e^{2\hat{d}(\mathbf{x}, \mathcal{B})} - 1}{2\, e^{\hat{d}(\mathbf{x}, \mathcal{B})}\, \nabla_{\mathbf{x}} \hat{d}(\mathbf{x}, \mathcal{B})}.$$

- Corollary: Near the decision boundary, d $(\boldsymbol{x}, \boldsymbol{\mathcal{B}})$ $\simeq$ $\hat{d}$ $(\boldsymbol{x}, \boldsymbol{\mathcal{B}})$

Relation to other fairness metrics

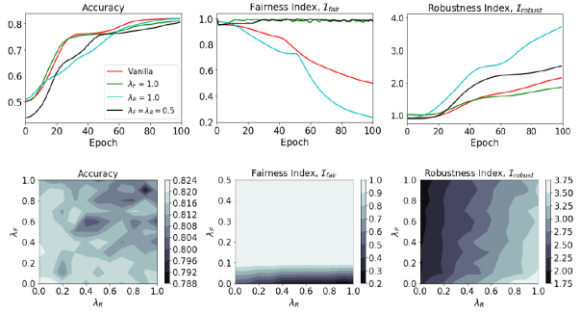
- There can be no formal guarantees on the relation of fairness via reducing recourse gap and fairness through other measures (the former relies on the distance to the boundary while the latter depend only on error rates).
- However, we show that our fairness loss also encourages the reduction in true and false positive rates between groups i.e., using this loss also helps improve on equalized odds

Experiments

- Results of training a vanilla neural network and FaiR-N for the datasets (UCI Adult, German Credit, MEPS):

Dataset	Acc.	$\mathcal{I}_{fair}$	$\mathcal{I}_{robust}$	|Δ TPR|	|Δ FPR|
Vanilla ($\lambda_F = \lambda_R = 0$)					
Adult	**0.821±0.002**	0.502±0.058	2.16±0.052	0.399±0.033	0.105±0.007
German	**0.767±0.007**	0.941±0.047	1.43±0.061	0.120±0.058	0.215±0.078
MEPS*	**0.848±0.002**	0.726±0.044	2.24±0.038	0.104±0.029	0.023±0.007
FaiR-N ($\lambda_F = \lambda_R = 0.5$)					
Adult	0.808±0.007	**0.968±0.024**	**2.52±0.178**	**0.050±0.051**	**0.017±0.009**
German†	0.750±0.007	**0.946±0.056**	**1.90±0.061**	**0.053±0.037**	**0.082±0.057**
MEPS*	0.838±0.010	**0.954±0.028**	**3.25±0.702**	**0.035±0.028**	**0.010±0.011**

- Results of training a vanilla neural network and FaiR-N for the UCI Adult dataset with different hyperparameter combinations:



- Results compared to other baselines:
  - Comparing our distance formulation to the original distance formulation
  - Comparing if the distance calculated is comparable to distances calculated in the input space via a comparison with CERTIFAI
  - Comparing to four state of the art methods that reduce error rate gaps

Method	Accuracy	|ΔTPR|	|ΔFPR|	Time (s)	$\Delta\text{Burden}_{CER}$
VNN-2	0.82±0.09	0.34±0.041	0.13±0.01	**318.2** ± 7.2	0.68± 0.04
FaiR-N-2	0.81± 0.09	**0.041± 0.067**	**0.021 ± 0.01**	**324.4± 8.52**	**0.06± 0.03**
FaiR-N-4	**0.83 ± 0.12**	0.07 ± 0.09	0.05 ± 0.01	344.9 ± 11.92	0.1 ± 0.04
P.R.†	0.78±0.06	0.12 ± 0.03	0.10 ± 0.02	371.4 ±8.44	0.16 ± 0.05
O.P.P.†	0.77± 0.11	0.09 ± 0.03	0.04 ± 0.08	424.2 ± 16.10	0.23 ± 0.04
DataAug†	0.78 ± 0.10	0.04 ± 0.02	0.03±0.01	520.1±13.6	0.22± 0.07
RedApp†	0.81 ± 0.7	0.05 ± 0.01	0.04±0.02	419.4±11.4	0.19± 0.04
LargeM-2	0.808 ± 0.09	0.044 ± 0.008	0.020 ±0.006	471.7 ± 19.48	**0.03 ± 0.01**
LargeM-4	0.829±0.12	0.049 ± 0.004	0.08 ± 0.01	631.5 ± 19.91	0.10 ± 0.05