Machine Learning Fairness

Sarah Kurihara,Varchasvi Vedula,Wenhui Fang,
Krista Zhang,Sharon Meng
Team 8
Statistics GR5243
Columbia University

I. Implemented Algorithm

- A4: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment (DM and DM-sen)
- A6: Handling Conditional Discrimination (LM and LPS)
- Pre-processing: A6
- In-Processing: A4

II. Dataset Overview & Processing

- A database containing the criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County from 2013 and 2014
- 7214 Observations * 53 Features

- Select features with less than 15% missing, not string or meaningless
- Binarize categorical variables, take logarithm on time variables
- 7214 Observations * 53 Features -> 5730 Observations * 16 Features
- Train:Validation:Test = 5:1:1 for model training

III. A6: Handling Conditional Discrimination

Setup:

- Some differences in decisions made across sensitive groups can be explained by other (correlated, non-sensitive) attributes; therefore we only want to remove the bad discrimnation and keep the explainable discrimination
- Local techniques for handling conditional discrimination when one of the attributes is considered to be explanatory
- Instead of removing S and E, we augment some training data near the decision boundary to control for bad bias
- S (sensitive attribute): Race; E (explanatory variable): Charge Degree

Local Massaging (LM)

Goal: **Modify data labels** until for all $charge_i$ in the range of charge,

$$P'(+|AA, charge_i) = P'(+|C, charge_i) = (P(+|AA, charge_i) + P(+|C, charge_i))/2$$

P: Probability before modifying data, P': Probability after modifying data

Process:

- For each $charge_i$, train a classifier. Switch the target labels for a calculated number of AA and C observations near the decision boundary

Local Massaging (LM)

```
DELTA(African American) =  46 African Americans changed from 1 to 0
DELTA(Caucasian) =  39 Caucasians changed from 0 to 1
DELTA(African American) =  90 African Americans changed from 1 to 0
DELTA(Caucasian) =  54 Caucasians changed from 0 to 1
```

After the data augmentation, a logistic regression model is trained on this new data.

Improved parity and calibration, compared to the baseline logistic regression model.

Local Preferential Sampling (LPS)

Goal: Modify data composition by **deleting and duplicating training observations** until for all $charge_i$ in the range of charge,

$$P'(+|AA, charge_i) = P'(+|C, charge_i) = (P(+|AA, charge_i) + P(+|C, charge_i))/2$$

P: Probability before modifying data, P': Probability after modifying data

Process:

- For each $charge_i$, train a classifier. Delete and duplicate a certain number of of AA and C observations near the decision boundary to remove discrimination in training data.

Local Preferential Sampling (LPS)

Size of training data remains the same (but composition changed)

After the data augmentation, a logistic regression model is trained on this new data.

Improved calibration compared to the baseline logistic regression model.

Highest overall accuracy after the baseline models.

IV. A4: Disparate Mistreatment

- Avoiding disparate treatment: $P(\hat{y} \mid x, z) = P(\hat{y} \mid x)$
  - Given the information of sensitive feature, the prob will not change
- Goal: The misclassification rates for different groups of people having different values of the sensitive feature z are the same
  - Minimizing differences of FPR and FNR for each group
- Notation:

overall misclassification rate (OMR):

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1),$$

false positive rate (FPR):

$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1),$$

false negative rate (FNR):

$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1),$$

Disparate Mistreatment With Sensitive

- To restrict the overall misclassification rate, we put constraints on loss function optimization problem with threshold ε

$$
\begin{aligned}
\text{minimize} \quad & L(\boldsymbol{\theta}) \\
\text{subject to} \quad & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \\
& P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon,
\end{aligned} \tag{8}
$$

Rewrite Problem into DCCP

- DCCP: Disciplined Convex-Concave Program
- We use logistic regression for modeling and training
- $$\begin{array}{ll} \text{minimize} & -\sum_{(\mathbf{x},y)\in\mathcal{D}} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \le c \\ & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \ge -c. \end{array}$$
- $g_\theta(y,x)$ means signed distance to the boundary

Training Process

- Getting θ from Loss function DCCP problem with training data
- Predict response by using θ in logistic regression
- Evaluate results

V. Evaluation Method

- Accuracy: When controlling fairness, does overall accuracy fall greatly?
- Parity: Are probability for positive prediction differs in two groups?
- Calibration: Do accuracies differ in two groups?
- False Positive Rate: Is it more likely to test positive for one group?
- False Negative Rate: Is it more likely to neglect positive individuals for one group?

VI. Result

Algorithm 4:

- Consider more on the FPR and FNR
- Not too much accuracy loss

- Consider less on parity
- Takes longer time
- DM-sen algorithm takes sensitive feature in learning — can result in disparate treatment though this effect is unobservable
- Unstable when features are limited

Result

Algorithm 6:

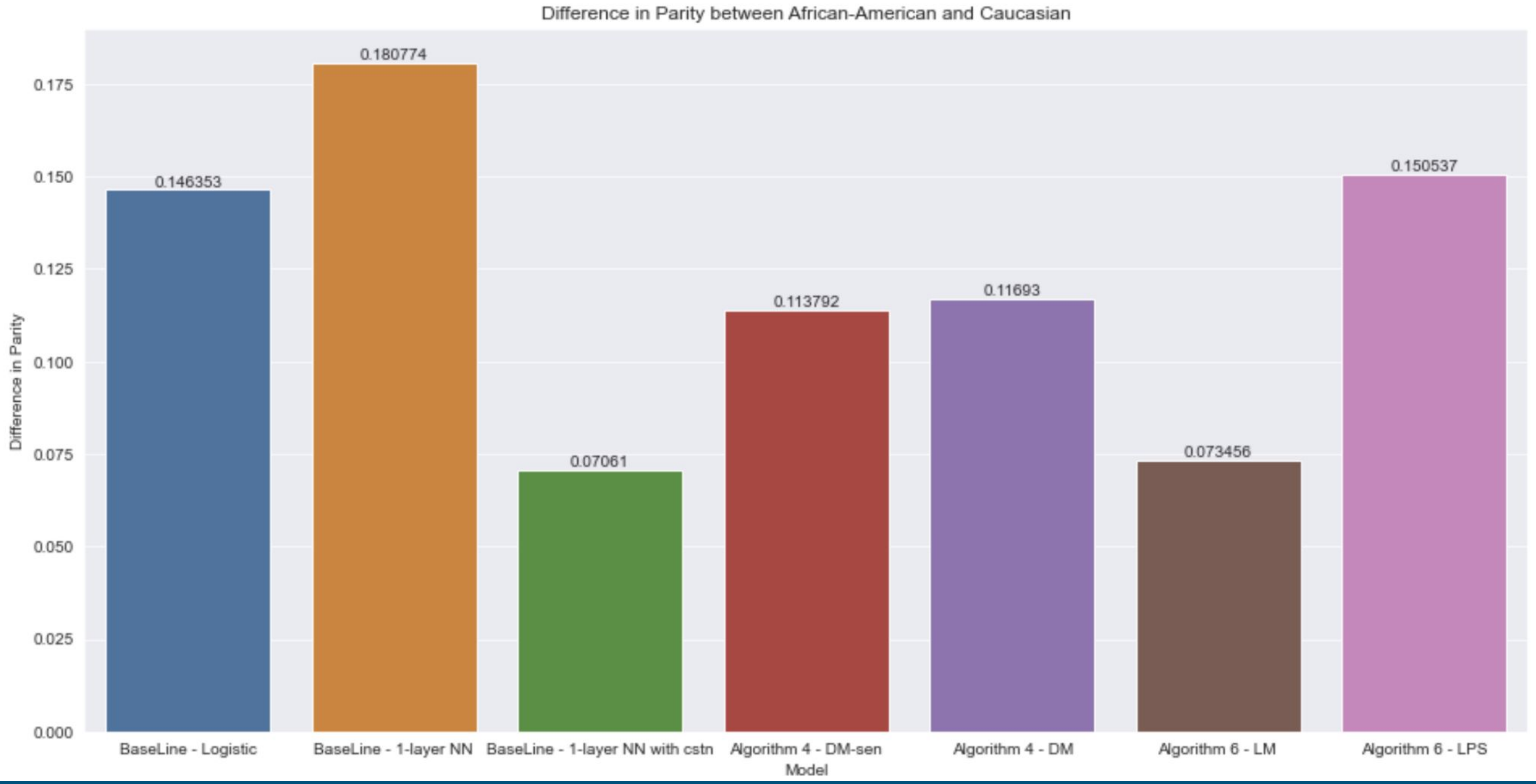
LM:

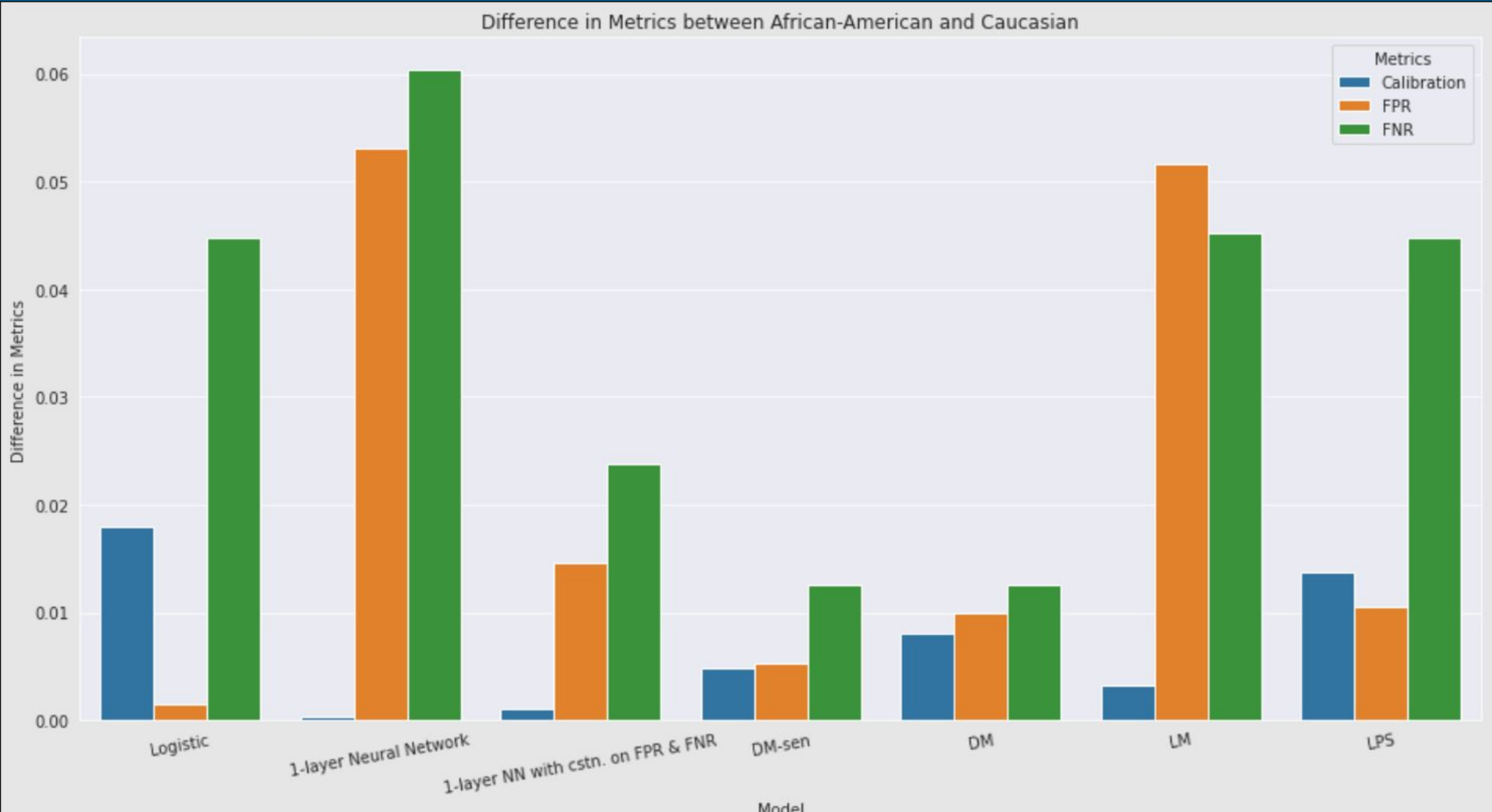
- Considers more on parity and calibration
- Accuracy loss is higher than LPS
- Not controlling the difference in FNR and FPR

LPS:

- With least accuracy loss
- Parity and Calibration may not hold
- Not controlling the difference in FNR and FPR

Overall Result Comparison on Algorithm 4 and Algorithm 6							
Metrics			Accuracy	Parity	Calibration	False Positive Rate (FPR)	False Negative Rate (FNR)
BaseLine	Logistic	Overall	0.93	-	-	0.067146	0.073107
		African-American	-	0.537657	0.937238	0.067873	0.058366
		Caucasian	-	0.391304	0.919255	0.066327	0.103175
		Difference	-	*0.146353*	*0.017983*	*0.001546*	*-0.044809*
	1-layer Neural Network	Overall	0.92875	-	-	0.079137	0.062663
		African-American	-	0.562762	0.92887	0.104072	0.042802
		Caucasian	-	0.381988	0.928571	0.05102	0.103175
		Difference	-	*0.180774*	*0.000299*	*0.053052*	*-0.060373*
	1-layer NN with *cstn.* on FPR & FNR	Overall	0.80375	-	-	0.201439	0.190601
		African-American	-	0.520921	0.803347	0.19457	0.198444
		Caucasian	-	0.450311	0.804348	0.209184	0.174603
		Difference	-	*0.07061*	*-0.001001*	*-0.014614*	*0.023841*
Algorithm 4	DM-sen	Overall	0.924084	-	-	0.088496	0.063683
		African-American	-	0.427966	0.927966	0.084507	0.053191
		Caucasian	-	0.541758	0.923077	0.089835	0.065708
		Difference	-	*-0.113792*	*0.004889*	***-0.005328***	***-0.012517***
	DM	Overall	0.925829	-	-	0.084956	0.063683
		African-American	-	0.423729	0.932203	0.077465	0.053191
		Caucasian	-	0.540659	0.924176	0.08747	0.065708
		Difference	-	*-0.11693*	*0.008027*	*-0.010005*	***-0.012517***
Algorithm 6	LM	Overall	0.915	-	-	0.069544	0.101828
		African-American	-	0.495816	0.916318	0.045249	0.116732
		Caucasian	-	0.42236	0.913043	0.096939	0.071429
		Difference	-	***0.073456***	***0.003275***	*-0.05169*	*0.045303*
	LPS	Overall	**0.9275**	-	-	0.071942	0.073107
		African-American	-	0.541841	0.933054	0.076923	0.058366
		Caucasian	-	0.391304	0.919255	0.066327	0.103175
		Difference	-	*0.150537*	*0.013799*	*0.010596*	*-0.044809*


Difference in Parity between African-American and Caucasian
Difference in Parity
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.146353
0.180774
0.07061
0.113792
0.11693
0.073456
0.150537
BaseLine - Logistic
BaseLine - 1-layer NN
BaseLine - 1-layer NN with cstn
Algorithm 4 - DM-sen
Algorithm 4 - DM
Algorithm 6 - LM
Algorithm 6 - LPS
Model


Difference in Metrics between African-American and Caucasian
Metrics
Calibration
FPR
FNR
Difference in Metrics
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Logistic
1-layer Neural Network
1-layer NN with cstn. on FPR & FNR
DM-sen
DM
LM
LPS
Model

Thank you!