Fairness Constraints: Mechanisms for Fair Classification
VS
Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

GR5243 Project 4

Group 3

Starting data processing

	age	race	sex	priors_count	c_charge_degree	jail_time	decile_score	score_text	two_year_recid
1	34	African–American	Male	0	F	10.0	3	Low	1
2	24	African–American	Male	4	F	1.0	4	Low	1
6	41	Caucasian	Male	14	F	6.0	6	Medium	1
8	39	Caucasian	Female	0	M	2.0	1	Low	0
9	21	Caucasian	Male	1	F	0.0	3	Low	1

- First of all, we define a list of feature names called 'feauture_name', including age, race, sex, priors count, charge degree, jail time, decile score, score text. We use those features to predict the likelihood of individuals reoffending. Also ,we retain only samples where the race is 'African-American' or ‘Caucasian'.

Fairness Constraints: Mechanisms for Fair Classification

- Explainability and Transparency
- Compliance with Legal and Ethical Standards
- Building Trust
- Promoting Diversity and Inclusivity
- Business Value and Social Impact

Fairness Constraints: Mechanisms for Fair Classification

2 Fairness Constraints: Mechanisms for Fair Classification

This method [2] considers the signed distance from the users' feature vectors to the decision boundary $\{d_\theta(x_i)\}_{i=1}^{N}$, and compute

$$Cov(z, d_\theta(x)) \approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \tag{5}$$

where z is the protected feature. This is a convex function with respect to the decision boundary parameters θ.

```
train accuracy: 0.9617433414043584
test accuracy: 0.9607843137254902
train calibration: 0.006170192851669975 6
test calibration: 0.006230752822441593
```

- incorporating fairness constraints in the classifier design
- constructing covariance function
- significance of the convex nature of the function with respect to optimization and model training

Fairness Beyond Disparate Treatment & Disparate Impact:Learning Classification without Disparate Mistreatment

- Comprehensive Fairness
- Maintained Model Performance
- Flexibility and Adaptability
- Reduced Bias

- Introducing Fairness Constraints
- Reweight the sample
- Logistical Regression Fairness-aware Loss Functions
- Fairness-aware Regularization

Fairness Beyond Disparate Treatment & Disparate Impact:Learning Classification without Disparate Mistreatment

Data processing

	index	intercept	age	race	sex	priors_count	c_charge_degree	jail_time	score_text	decile_score_1	decile_score_2	decile_score_3	decile_score_4	decile_score_5	decile_score_6	decile_score_7	decile_score_8	decile_score_9	decile_score_10
0	1	1.0	−0.048742	0	1	−0.733669	0	−0.167787	0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	2	1.0	−0.894973	0	1	0.055933	0	−0.340683	0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	6	1.0	0.543619	1	1	2.029939	0	−0.244630	1	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	8	1.0	0.374373	1	0	−0.733669	1	−0.321473	0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	9	1.0	−1.148842	1	1	−0.536269	0	−0.359894	0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

- Intercept Column Addition
- Feature Standardization:
- Label Encoding of Categorical Variables
- Mapping of Score Text
- Index Resetting

Fairness Beyond Disparate Treatment & Disparate Impact:Learning Classification without Disparate Mistreatment

Denote the user feature vectors as x, class labels as $y \in \{-1, 1\}$, sensitive features $z \in \{0, 1\}$, and the training dataset as $\mathcal{D}$. This method [3] considers the covariance between the users' sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary,

$$Cov(z, g_\theta(y, x)) \approx \frac{1}{N} \sum_{(x,y,z)\in\mathcal{D}} (z - \bar{z}) g_\theta(y, x) \tag{9}$$

where g_θ can be defined as

$$g_\theta(y, x) = min(0, y d_\theta(x))$$
$$g_\theta(y, x) = min(0, \frac{1-y}{2} y d_\theta(x))$$
$$g_\theta(y, x) = min(0, \frac{1+y}{2} y d_\theta(x))$$

However, since the problem

$$\begin{aligned} \min \quad & L(\theta) \\ \text{s.t.} \quad & \frac{1}{N} \sum_{(x,y,z)\in\mathcal{D}} (z - \bar{z}) g_\theta(y, x) \le c \\ & \frac{1}{N} \sum_{(x,y,z)\in\mathcal{D}} (z - \bar{z}) g_\theta(y, x) \ge -c \end{aligned} \tag{10}$$

is nonconvex, the constraints are converted into a Disciplined Convex Concave Program which can be solved efficiently.

$$\begin{aligned} \min \quad & L(\theta) \\ \text{s.t.} \quad & \frac{-N_1}{N} \sum_{(x,y)\in\mathcal{D}_0} g_\theta(y, x) + \frac{N_0}{N} \sum_{(x,y)\in\mathcal{D}_1} g_\theta(y, x) \le c \\ & \frac{-N_1}{N} \sum_{(x,y)\in\mathcal{D}_0} g_\theta(y, x) + \frac{N_0}{N} \sum_{(x,y)\in\mathcal{D}_1} g_\theta(y, x) \ge -c \end{aligned} \tag{11}$$

where $\mathcal{D}_0$ and $\mathcal{D}_1$ are the subsets of the training dataset $\mathcal{D}$ taking values $z = 0$ and $z = 1$, respectively. $N_0 = |D_0|$ and $N_1 = |D_1|$.

- Function9:emphasizing its role in measuring the disparity in the distance from the decision boundary for misclassified individuals.
- Function10:converting the constraints into a Disciplined Convex Concave Program (DCCP) to solve it efficiently.
- Function11: setting a fairness boundary, ensuring the model's errors do not systematically favor or disadvantage any group.

Result

Unconstrained classifier
Accuracy: 0.69236

	Sensitive Attribute	FPR	FNR	TNR	TPR	Accuracy
0	0	0.33	0.30	0.67	0.70	0.686492
1	1	0.21	0.45	0.79	0.55	0.701467

FPR constraint classifier
Accuracy: 0.6765

	Sensitive Attribute	FPR	FNR	TNR	TPR	Accuracy
0	0	0.26	0.36	0.74	0.64	0.688716
1	1	0.33	0.36	0.67	0.64	0.657463

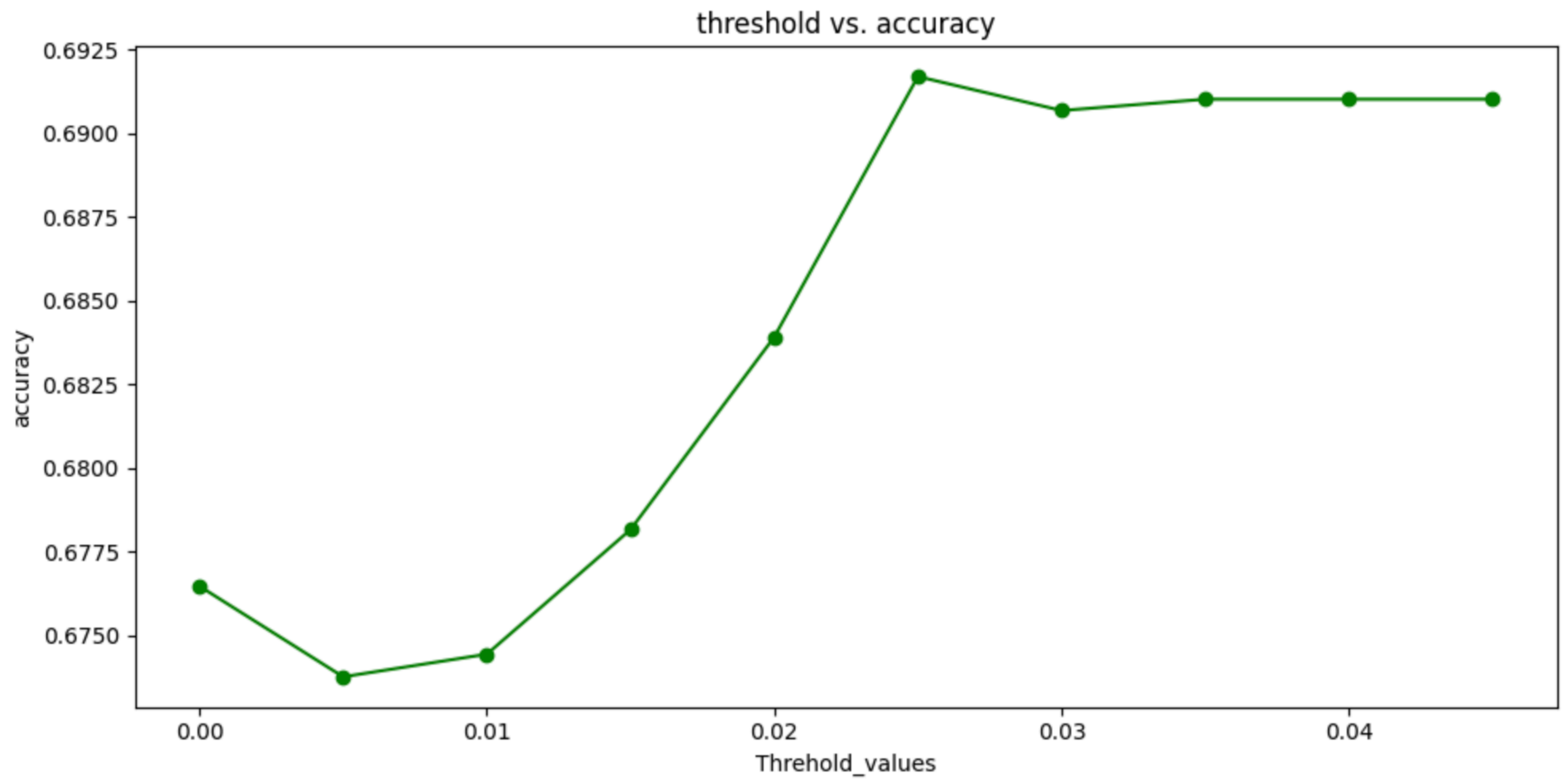
FNR constraint classifier
Accuracy: 0.6836

	Sensitive Attribute	FPR	FNR	TNR	TPR	Accuracy
0	0	0.26	0.36	0.74	0.64	0.682601
1	1	0.26	0.41	0.74	0.59	0.685073

FPR & FNR constraint classifier
Accuracy: 0.6792

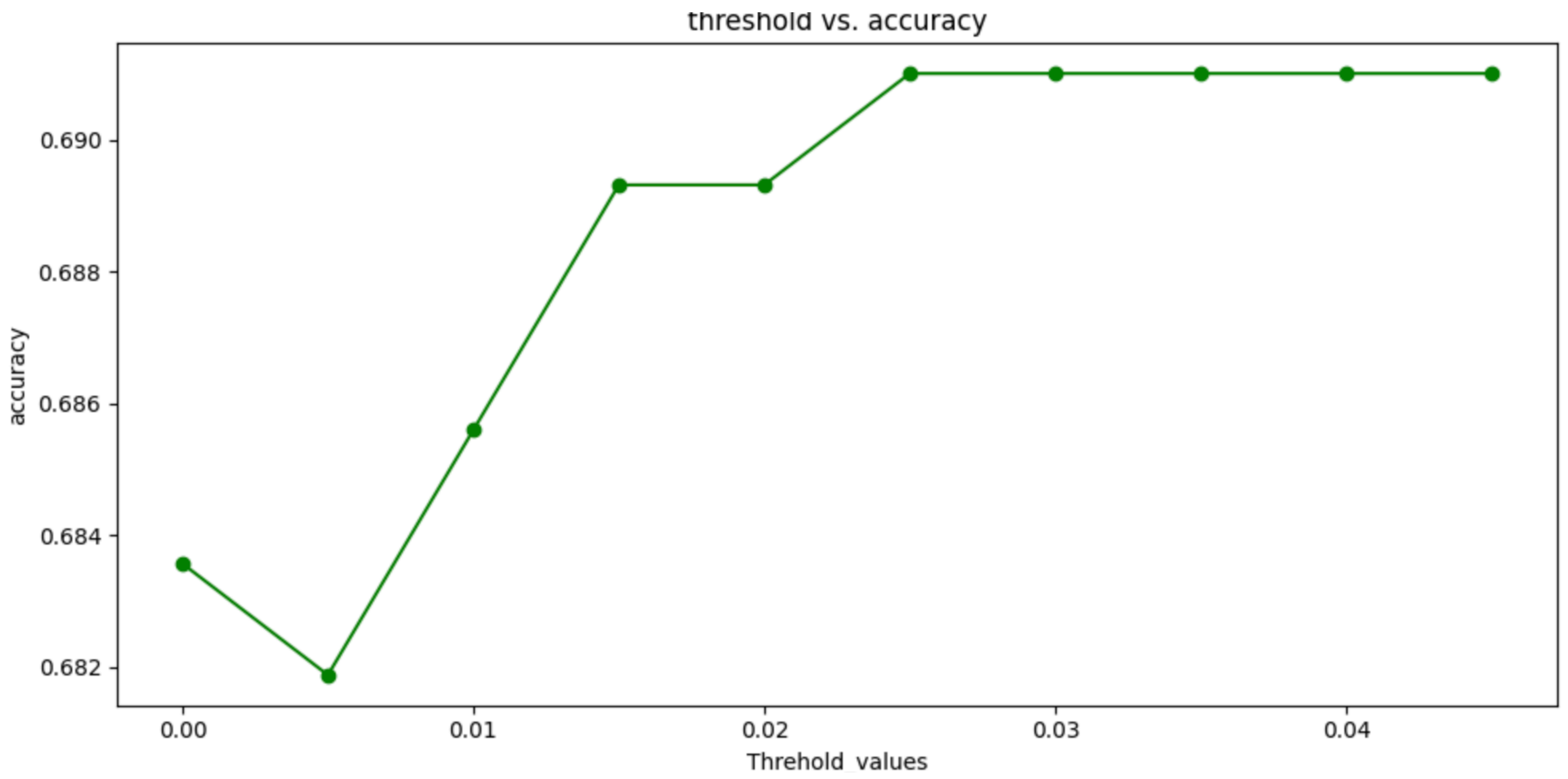
	Sensitive Attribute	FPR	FNR	TNR	TPR	Accuracy
0	0	0.24	0.37	0.76	0.63	0.690384
1	1	0.32	0.38	0.68	0.62	0.661777

- Under this condition of threshold is 0. FNR will give us the most close accuracy.



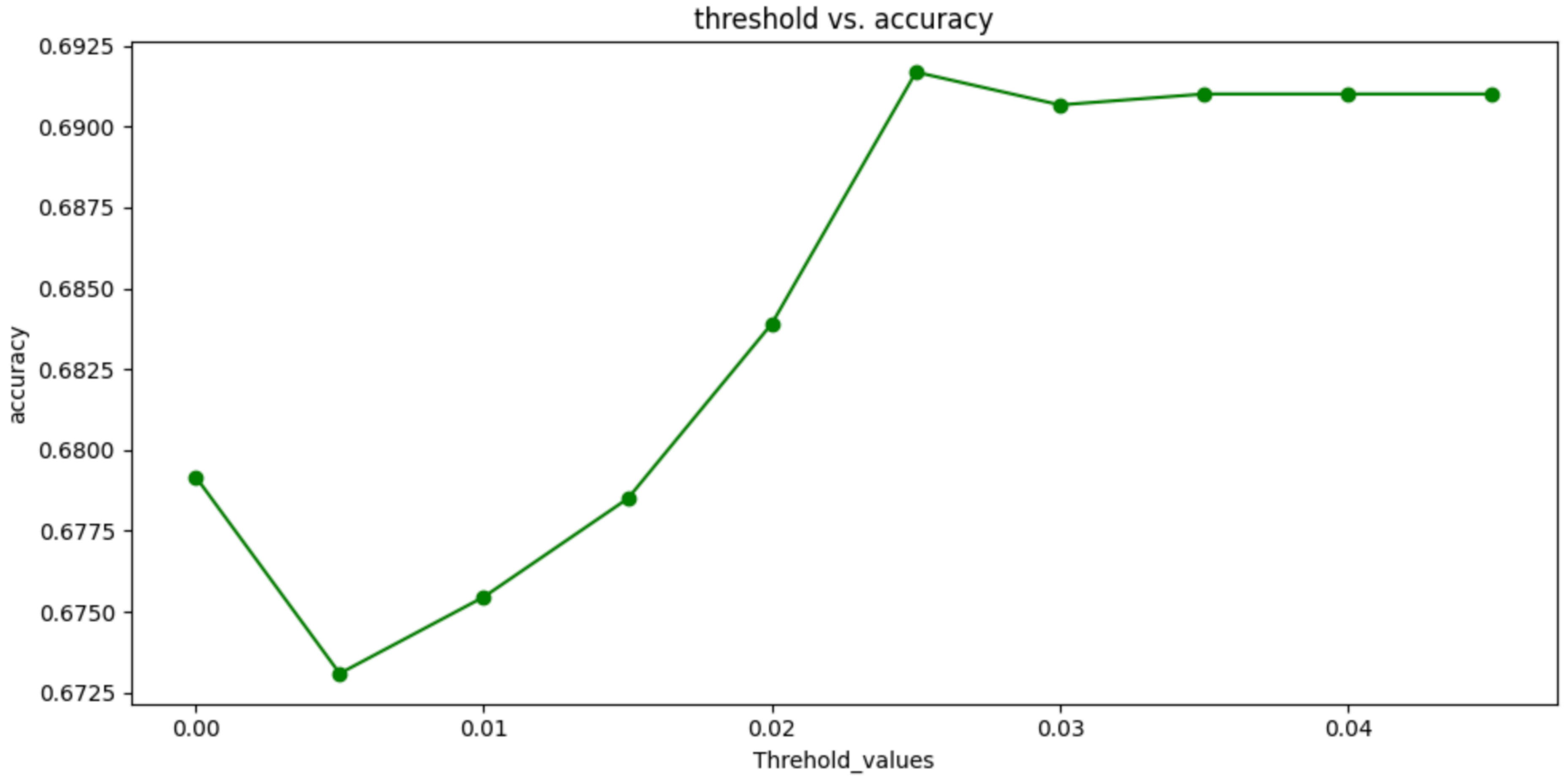
For FPR

When the threshold from 0.005 to 0.025, the accuracy always increase. And it reach its highest value on 0.025.



For FNR

When the threshold value reaches about 0.05, it will begin to increase fastly, and when the threshold reached 0.025, it will stop to increase and achieve the highest value.



For both FPR and FNR

The accuracy will decrease firstly, and when the threshold value reaches about 0.05, it will begin to increase fastly, and when the threshold reached 0.025, it will stop to increase and decrease slightly.

Reference

- Zafar, Muhammad Bilal, et al. “Fairness beyond Disparate Treatment & Disparate Impact.” Proceedings of the 26th International Conference on World Wide Web - WWW ’17, 2017, arxiv.org/pdf/1610.08452.pdf, https://doi.org/10.1145/3038912.3052660.

- https://github.com/TZstatsADS/ADS_Teaching/blob/master/Tutorials/wk10-Overview_Machine_Learning_Fairness_Methods.pdf