Project : Machine Learning Fairness (Paper A1 & A4)

Group 3

A1 Learning Fair Representations

Paper A1

Summary of Paper A1

- Goal: We want to solve fairness problems (both group and individual fairness).
- How:
  - We develop a learning approach to solve fairness problems.
  - We learn a restricted form of distance function so that we can eliminate assumption that the distance function is given apriori.
  - We formulate fairness as optimization problem by finding intermediate representation of data, so that we can best encode the data and obfuscate information about membership in the protected group.
  - In the model, we map each individual (a data point in the given input space) to a probability distribution in a new representation space. Then we do the classification based on the new representations.

Model

- Minimize Objective Function:

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

$$L_z = \sum_{k=1}^{K} \left| M_k^+ - M_k^- \right|$$

$$M_k^+ = \mathbb{E}_{\mathbf{x} \in X^+} P(Z = k|\mathbf{x}) = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k}$$

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2$$

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k$$

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)$$

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k$$

$$M_{n,k} = P(Z = k|\mathbf{x}_n) \quad \forall n, k$$

$$P(Z = k|\mathbf{x}) = \exp(-d(\mathbf{x}, \mathbf{v}_k)) / \sum_{j=1}^{K} \exp(-d(\mathbf{x}, \mathbf{v}_j))$$

Results

	Accuracy			Calibration	Run time
	African-American	Caucasian	Overall		
Train	0. 48	0. 6	0. 53	0. 12	663s
Validation	0. 47	0. 59	0. 52	0. 12	
Test	0. 44	0. 64	0. 53	0. 2	0. 44s

A4

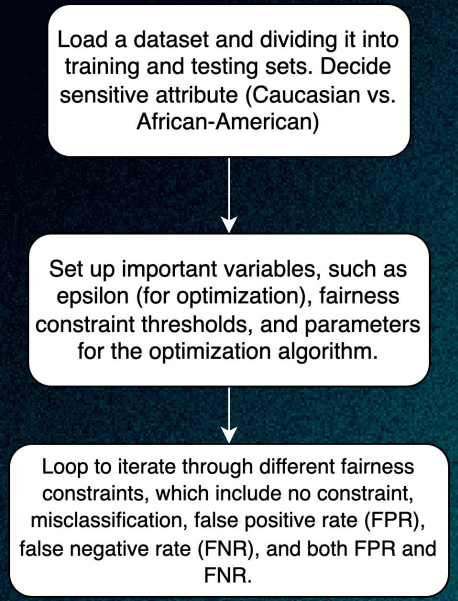
Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment (DM and DM-sen)

Paper A4

Summary of Paper

- Previous fairness definitions, such as disparate treatment and disparate impact, are **insufficient** because they **focus on avoiding bias in the outcome of the model rather than in the treatment of individuals.**
- The authors propose a new definition of fairness, called "disparate mistreatment," which takes into account the harmful effects of false positives and false negatives on different subgroups of the population. (**Minimize loss function while constrained on Covariance between Sensitive variable and the signed distance between the feature vectors of misclassified users and the classifier decision boundary**)
- The DM constraint is defined as follows: for each subgroup of the population (e.g., based on race or gender), **the difference between the false positive rate and the false negative rate should be less than or equal to a predefined threshold**. This threshold is chosen based on the desired level of fairness and the base rate of the subgroups

Algorithm Diagram



For each constraint type:
a. Set up the constraint parameters accordingly.
b. Train a classifier considering the given constraints using the train_model function.
c. Evaluate the classifier's performance and fairness by calculating accuracy, FPR, and FNR for each group (e.g., Caucasian and African-American individuals).

Constraints

$$\begin{aligned}\mathrm{Cov}(z, g_\theta(y, \mathbf{x})) &= \mathbb{E}[(z - \bar{z})(g_\theta(y, \mathbf{x}) - \bar{g}_\theta(y, \mathbf{x}))] \\ &\approx \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_\theta(y, \mathbf{x}),\end{aligned}$$

Covariance Approximation

$$g_\theta(y, \mathbf{x}) = min(0, y d_\theta(\mathbf{x})),$$

$$g_\theta(y, \mathbf{x}) = min\left(0, \frac{1-y}{2} y d_\theta(\mathbf{x})\right), \text{ or}$$

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

Overall Misclassification

Focus on False Positive

Focus on False Negative

Results (Accuracy)

Summary for sensitive attribute value White (Caucasian):

	Constraint_type	Accuracy	FPR	FNR
0	No Constraint	0.657033	0.368421	0.323024
1	Misclassification	0.660886	0.372807	0.312715
2	Only FPR	0.630058	0.271930	0.446735
3	Only FNR	0.637765	0.333333	0.384880
4	Both FPR and FNR	0.641618	0.315789	0.391753

Summary for sensitive attribute value Black (African-American):

	Constraint_type	Accuracy	FPR	FNR
0	No Constraint	0.668712	0.117318	0.591837
1	Misclassification	0.662577	0.162011	0.551020
2	Only FPR	0.674847	0.178771	0.503401
3	Only FNR	0.662577	0.223464	0.476190
4	Both FPR and FNR	0.662577	0.223464	0.476190

Result (Calibration)

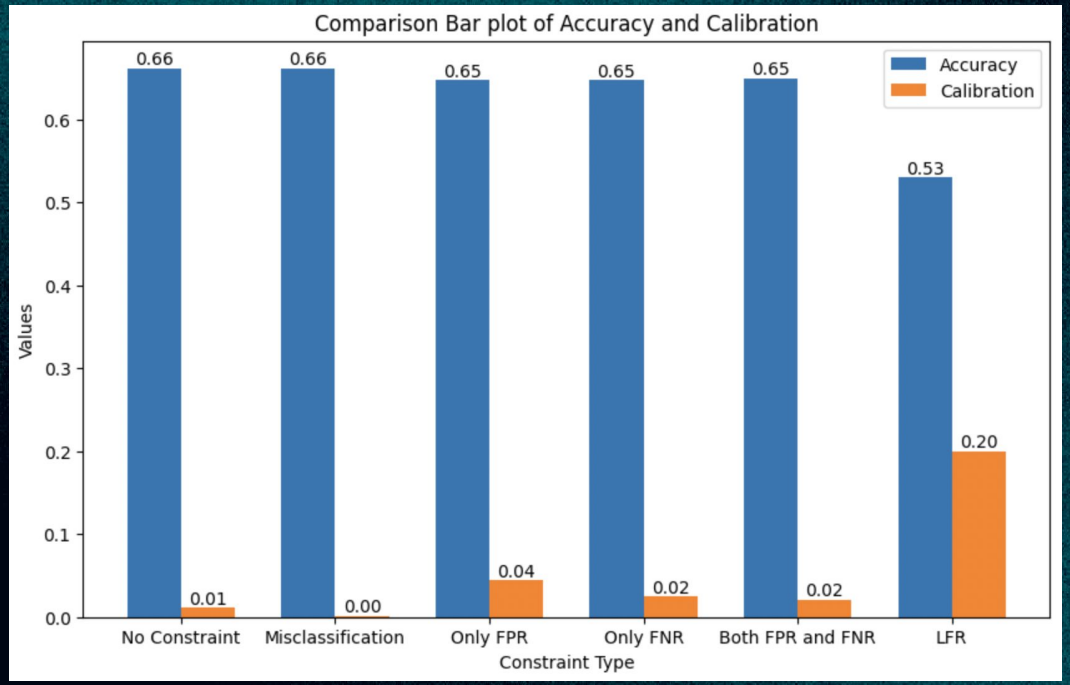
Summary of differences between sensitive attribute values:

	Constraint_type	Overall Accuracy	Diff_Accuracy	Diff_FPR	Diff_FNR
0	No Constraint	0.661538	0.011679	-0.251103	0.268813
1	Misclassification	0.661538	0.001690	-0.210796	0.238306
2	Only FPR	0.647337	0.044789	-0.093159	0.056666
3	Only FNR	0.647337	0.024812	-0.109870	0.091311
4	Both FPR and FNR	0.649704	0.020958	-0.092326	0.084438

Summary of Hinge Loss + Misclassification

```
Training accuracy: 0.96167247386759581
Test accuracy: 0.9631436314363143
Calibration train: 0.18490127758420438
Calibration test: 0.1869918699186992
```

Comparison Results & Conclusion



Paper A4 (Disparate Treatment) algorithms (Having fpr constraint, fnr constraint, fpr and fnr constraint) provides a better accuracy as well as better calibration than LFR algorithm