Report on ML Fairness
Project 4 - Group 7

Comparative Study of Machine Learning Fairness Papers

Team Member:

Han Wang, Yihan Zhang, Shaohuan Wu, Mannah, Shefali

Disparate Mistreatment (DM):

a concept that highlights unfairness in decision-making where misclassification rates differ across groups defined by sensitive attributes like race or gender.

Impact in Decision-Making Systems:

historical decisions used in training these systems

Examples in Classification Systems:

a classification system consistently misclassifies one racial group more often than another.

Goal:

Equitable Misclassification Rates

Approach:

incorporating DM measures into the decision-making algorithms

Unfairness Notions:

disparate treatment: sensitive attributes impacts on the decision, when non-sensitive be the same
disparate impact: different groups have different positive decision rate
disparate mistreatment: different groups have different false negative rates

User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-sensitive					
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

Figure 1: Decisions of three fictitious classifiers (C_1, C_2 and C_3) on whether (1) or not (0) to stop a pedestrian on the suspicion of possessing an illegal weapon. Gender is a sensitive attribute, whereas the other two attributes (suspicious bulge in clothing and proximity to a crime scene) are non-sensitive. Ground truth on whether the person is actually in possession of an illegal weapon is also shown.

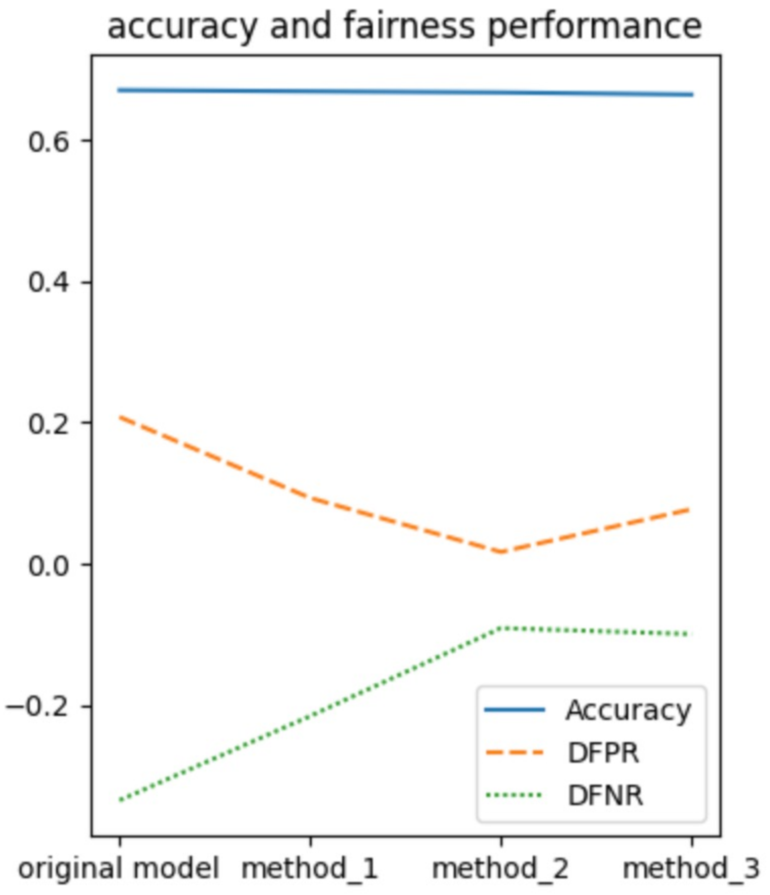
DM
Approach, Algorithm and Basic Elements

Algorithm:

Limit DM by introducing constraints to the false positive rates, false negative rates, or both

Evaluation:

calculating the accuracy of the algorithm, the difference between the false positive rates of sensitive attributes (DFPR), and the difference between the false negative rates of sensitive attributes DFNR)



	Accuracy	DFPR	DFNR
original model	0.669593	0.207324	-0.334695
method_1	0.667967	0.092858	-0.215562
method_2	0.666341	0.016511	-0.091154
method_3	0.663415	0.076943	-0.099580

We can conclude that Method 2 (constraining the false negative rate) is the best at reducing disparate mistreatment while also not compromising the accuracy a lot.

Introduction to Local Massaging (LM) and Local Preferential Sampling (LPS)

Local Massaging (LM) and Local Preferential Sampling (LPS):

used to adjust a dataset or a model's decisions to reduce bias
quantifies the objectively explainable part of discrimination
introduces conditional discrimination-aware classification

Conditional Discrimination:

Discrimination conditioned on sensitive attributes

Traditional Solutions:

Remove all discrimination
Limitations: reverse discrimination

Effects:

eliminate "bad" discrimination while allowing justified differences in decisions

Quantify Explainable discrimination:

independent attributes
explanatory attributes (e)
sensitive attributes (s)
favored group: m
deprived group: f

Bad Discrimination:

$$D_{bad} = D_{all} - D_{expl}$$

$$\begin{aligned} D_{expl} &= \sum_{i=1}^{k} P(e_i|m)P^{\star}(+|e_i) - \sum_{i=1}^{k} P(e_i|f)P^{\star}(+|e_i) \\ &= \sum_{i=1}^{k} \left(P(e_i|m) - P(e_i|f)\right) P^{\star}(+|e_i), \end{aligned}$$

$$P^{\star}(+|e_i) := \frac{P(+|e_i, m) + P(+|e_i, f)}{2},$$

$$P_c(+|e_i, m) = P_c(+|e_i, f)$$

$$P_c(+|e_i) = P^{\star}(+|e_i).$$

LM&LPS
Algorithms

Massaging

Learn an internal ranker, to identify the instances that are close to the decision boundary

(a classifier that outputs the posterior probabilities)

Change the values of their labels to the opposite.

Algorithm 1: Local massaging

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  rank males using $\mathcal{H}_i$;
  `relabel DELTA` (male) males that are the closest to the decision boundary from $+$ to $-$ (Algorithm 4);
  rank females using $\mathcal{H}_i$;
  `relabel DELTA` (female) females that are the closest to the decision boundary from $-$ to $+$
end

Algorithm 4: subroutine DELTA(gender)

return $G_i |p(+|e_i, \text{gender}) - p^\star(+|e_i)|$,
where $p^\star(+|e_i)$ comes from (Eq. (4)),
G_i is the number of gender people in $X^{(i)}$;

Algorithm 2: Local preferential sampling

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: resampled dataset (a list of instances)

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (see Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  `rank` males using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (male) (see Algorithm 4) males $+$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (male) males $-$ that are the closest to the decision boundary;
  `rank` females using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (female) females $-$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (female) females $+$ that are the closest to the decision boundary;
end

Evaluation and Conclusion

```
recidivated rate for Caucasian = 41.04%
recidivated rate for African-American = 50.03%
```

```
recidivated rate for Caucasian = 40.06%
recidivated rate for African-American = 50.57%
```

Conclusions:

Based on the reuslts above, we can observe that the difference between the recidivated rate for two groups of people becomes smaller when we using the method of local massaging. Thus, local massaging may be a better choice in this case.

Comparing two pivotal methods in the field

Focus:

DM focuses on the model's learning process, ensuring fairness in error rates
LM focuses on adjusting the data or model outputs to reduce bias

Mechanism:

DM integrate fairness directly into the model's training algorithm
LM involves post-hoc adjustments to the data or decisions

Trade-offs:

In DM, the trade-off is often between fairness (in terms of error rates) and overall model accuracy
In LM, the trade-off is between reducing bias and maintaining the utility and accuracy of the data

Applicability:

DM approaches are more holistic and integrated into the model's training, making them potentially more robust but also more complex to implement.

LM is more flexible and can be applied to various models but might require careful tuning to avoid introducing new biases.

References

[1] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web (pp. 1171-1180).

[2] Žliobaite, F. Kamiran and T. Calders, "Handling Conditional Discrimination," 2011 IEEE 11th International Conference on Data Mining, Vancouver, BC, Canada, 2011, pp. 992-1001, doi: 10.1109/ICDM.2011.72.