Project 4: Machine Learning Fairness Algorithms Evaluation

Group 9:
Micol Clement, Vaishak Naik, Yinan Shi, Yvonne Zha, Ran Zhang

Introduction

In this project, we implemented, evaluated and compared algorithms for Machine Learning Fairness. Our study compared two different algorithms, Learning Classification without Disparate Mistreatment (DM and DM-sen) and Information Theoretic Measures for Fairness-aware Feature selection (FFS). Implementation and evaluation was based on the COMPAS dataset, which contains the criminal history, jail and prison time, demographics, and COMPAS risk scores for defendants from Broward County from 2013 and 2014. Our goal is to compare the performance and efficiency of the above algorithms.

Algorithm: DM & DM-sen

Issue: The difference in misclassification rate for different groups.
Solution: Quantify disparate mistreatment in following three ways:

- Overall Missclassification (OMR):

$$P[\hat{y} \neq y | z = 0] = P[\hat{y} \neq y | z = 1]$$

- False Negative Rate (FNR):

$$P[\hat{y} \neq y | z = 0, y = 0] = P[\hat{y} \neq y | z = 1, y = 0]$$

- False Positive Rate (FPR):

$$P[\hat{y} \neq y | z = 0, y = 1] = P[\hat{y} \neq y | z = 1, y = 1]$$

minimize loss function subject to constraints on disparate mistreatment

Solving optimization problem

For a logistic regression we end up solving the following Convex-Concave optimization problem which can be solved using DCCP:

$$\min_{w \in \mathcal{R}^d} \quad \frac{1}{N} \sum_{X,y \in \mathcal{D}} \log(1 + e^{Xw}) - yXw$$

$$\text{subject to} \quad -\frac{N_1}{N} \sum_{X,y \in \mathcal{D}_0} g_w(X,y) + \frac{N_0}{N} \sum_{X,y \in \mathcal{D}_1} g_w(X,y) \leq c$$

$$-\frac{N_1}{N} \sum_{X,y \in \mathcal{D}_0} g_w(X,y) + \frac{N_0}{N} \sum_{X,y \in \mathcal{D}_1} g_w(X,y) \geq -c$$

Implementation

DM & DM-sen, respectively with four types of constraints

- DM: sensitive feature z is not used while making decisions
- DM-sen: sensitive feature z is used as a learnable feature
- Features: Age Category, Gender, Priors Count, Charge Degree, Length of Stay, (Race)
- Constraints: OMR, FNR, FPR, FNR+FPR

Results of COMPAS Dataset

	Baseline logisitc	DM				DM-sen			
		OMR	FNR	FPR	FNR+FPR	OMR	FNR	FPR	FNR+FPR
Accuracy (%)	62.7642	61.9512	53.2520	56.3415	56.0976	62.7642	62.1951	61.4634	61.2195
DFNR(%)	36.6570	16.5646	4.0324	3.4928	2.7116	23.2764	12.4087	1.2860	0.07651
DFPR(%)	-25.8211	-8.9024	-14.0562	-3.6200	-3.6200	-21.9117	-7.6437	-0.0709	0.8757
Calibration(%)	3.0596	0.3722	-3.2624	11.4738	11.8830	5.0854	1.9887	2.5413	2.2753

Algorithm: Fairness-aware Feature Selection

Issue: Features that are relevant for accurate decisions may however lead to either explicit or implicit forms of discrimination against unprivileged groups, such as those of certain race or gender.

Solution: This model tries to tackle it by using information theoretic measures which quantify the impact of different subsets of features on the accuracy and discrimination on the dependent variable(Outcome Variable) Then use Shapley value function to quantify the marginal impact of each feature. This method does not focus on classifier design.

Goal: Select features that optimally satisfy **accuracy and fairness requirements**

Quantifying accuracy effect: A good accuracy measure for a subset of features

$$v^{Acc}(X_S) = I(Y; X_S|\{A, X_{S^c}\}) = UI(Y; X_S \backslash \{A, X_{S^c}\}) + CI(Y; X_S, \{A, X_{S^c}\}).$$

- Non-negativity: $v^{Acc}(X_S) \geq 0, \ \forall X_S \subseteq X^n$.
- Monotinicity: $v^{Acc}(X_{S_1}) \leq v^{Acc}(X_{S_2}), \ \forall S_1 \subseteq S_2$.
- Blocking: $Y \perp\!\!\!\perp X_S|\{A, X_{S^c}\} \iff v^{Acc}(X_S) = 0$.

Quantifying discriminatory effect: Discriminatory impact of Xs, should satisfy the following properties.

$$v^D(X_S) \triangleq SI(Y; X_S, A) \times I(X_S; A) \times I(X_S; A|Y).$$

- Non-negativity: $v^D(X_S) \geq 0$
- Monotinicity: $v^D(X_{S_1}) \leq v^D(X_{S_2})$ for $S_1 \subseteq S_2$
- Y-independence: $Y \perp\!\!\!\perp X_S \implies v^D(X_S) = 0$.
- A-independence: $A \perp\!\!\!\perp X_S \implies v^D(X_S) = 0$.
- AY-independence: $A \perp\!\!\!\perp X_S|Y \implies v^D(X_S) = 0$.

Shapley value function: Measures the marginal accuracy and discrimination impacts of a single feature.

Process:

- Find all subsets of feature and its impact excluding feature X(i)
- Calculate marginal(accuracy and discrimination) impacts including X(i) for each subset.
- Calculate weighted average with all subsets.
- Repeat the process for each feature(i).

Results of COMPAS Dataset

	Feature	Shapley Accuracy	Shapley Discrimination
0	Prior Count	1.26E+00	5.44E+04
1	Age Categorical	1.23E+00	5.38E+04
2	Length of Stay	1.09E+00	5.32E+04
3	Charge Degree	1.07E+00	4.36E+04
4	Gender	9.91E-01	4.29E+04

	Eliminating Feature	Accuracy (%)	Calibration (%)
0	None	65.77	2.47
1	Age Categorical	61.54	2.43
2	Prior Count	59.76	5.48
3	Gender	63.06	-1.57
4	Charge Degree	66.02	1.69
5	Length of Stay	65.85	2.33

Result comparison between DM-sen & FFS

DM-sen:

	FNR+FPR
Accuracy (%)	61.2195
DFNR(%)	0.07651
DFPR(%)	0.8757
Calibration(%)	2.2753

FFS:

	Eliminating Feature	Accuracy (%)	Calibration (%)
0	None	65.77	2.47
1	Age Categorical	61.54	2.43
2	Prior Count	59.76	5.48
3	Gender	63.06	-1.57
4	Charge Degree	66.02	1.69
5	Length of Stay	65.85	2.33