FATE P03, P04

Algorithmic Fairness II

Ashwin S they/them

ashwin.singh01@estudiant.upf.edu

Programming Sessions

1. **6 Labs =** 3 Modules x 2 Sessions

2. **Grading Policy**

~~M1: Data Anonymization~~	~~**(35%)**~~	~~Submit by 11:59 PM, Feb 03, 2025~~
~~M2: Algorithmic Fairness I~~	~~**(30%)**~~	~~Submit by 11:59 PM, Feb 20, 2025~~
M3: Algorithmic Fairness II	**(35%)**	Submit by 11:59 PM, Mar 11, 2025

You must attend at least one session of each module to be eligible for grading.

3. **Queries / Discussion**

   Please post on the Aula Global Forum for everyone's benefit.

4. **Solved Practices:** To be available after each deadline.

Outline

In **Module II**, we will focus on **Classification**

1. Evaluating Classifiers: Confusion Matrices, ROC & EDC Curves.
2. Pre / In / Post - Processing Interventions to Improve Fairness Properties.
3. How to do (2) using AI Fairness 360 by IBM Research.
4. Project [**You!**]

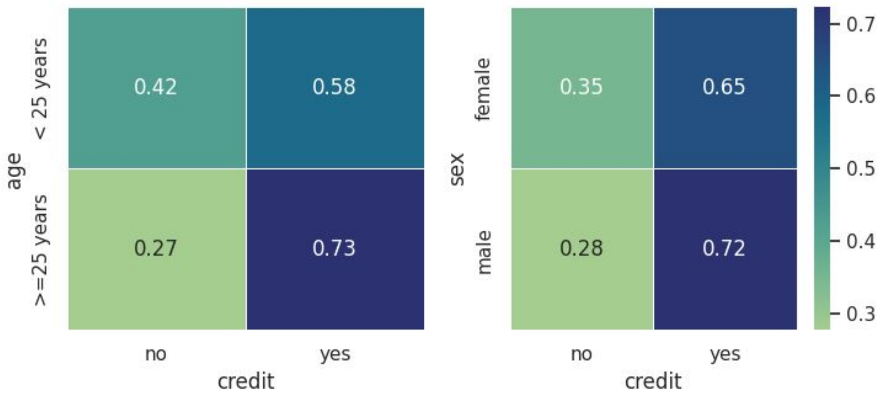
Datasets

1. German Credit Scoring

German Credit Scoring Data

age	sex	X_1	X_2	.....	X_n	credit	%
<=25	male	...	...	...	...	no	38
		...	...	...	...	yes	61
	female	...	...	...	...	no	45
		...	...	...	...	yes	55
>=25	male	...	...	...	...	no	26
		...	...	...	...	yes	74
	female	...	...	...	...	no	30
		...	...	...	...	yes	70

$A = \{ \text{sex, age} \}$ $X = \{ X_1, X_2, ..., X_n \}$ $Y = \text{credit}$



What are the privileged and unprivileged groups?

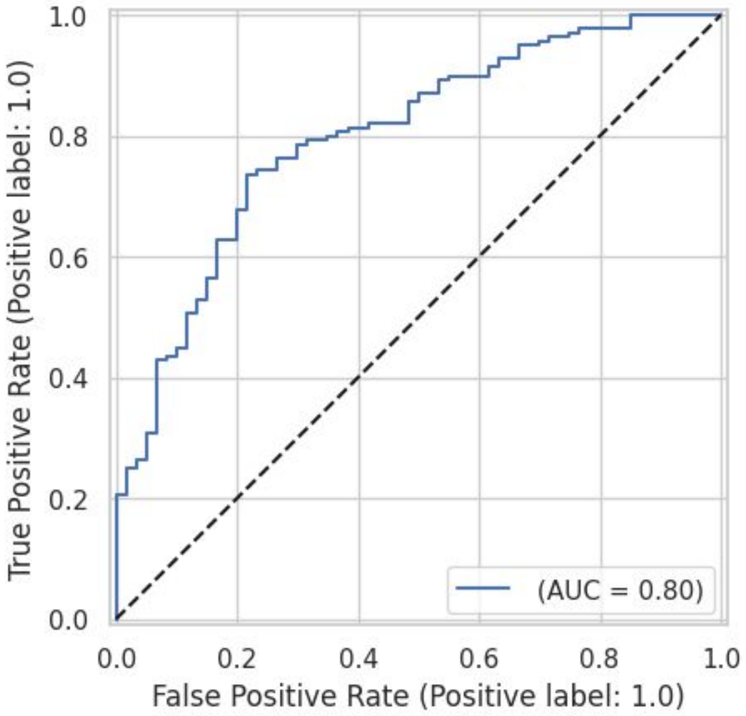
Evaluating Classifiers ▶ ROC Curves

		Predictions	
		$\hat{Y} = 0$	$\hat{Y} = 1$
Ground Truth	$Y = 0$	TN	FP
	$Y = 1$	FN	TP

False Positive Rate = FP / (FP + TN)

True Positive Rate = TP / (TP + FN)

Each point in the **ROC Curve** corresponds to a classification threshold θ such that $\mathbf{h(x) > \theta \Rightarrow \hat{Y} = 1}$.



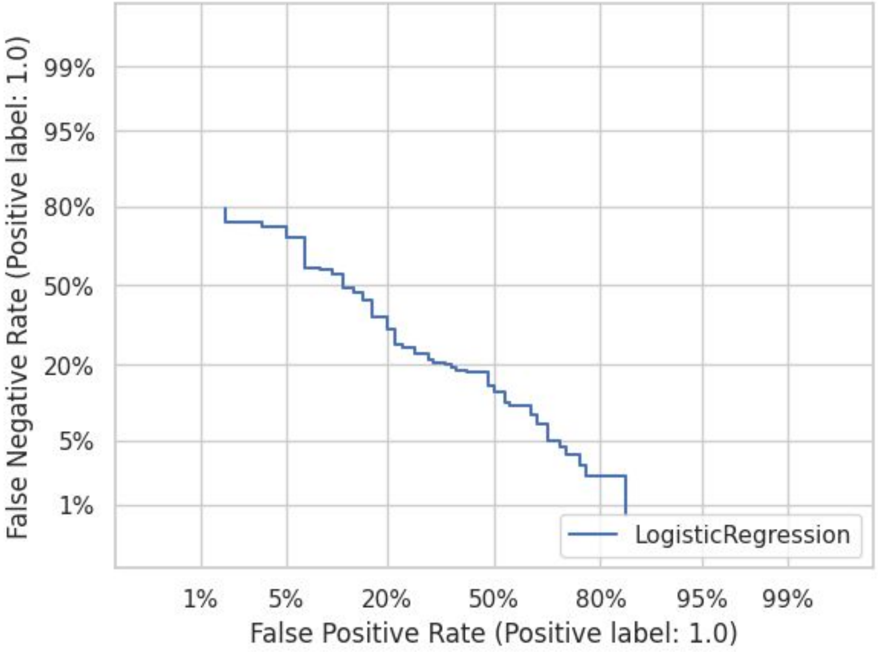
Evaluating Classifiers ▶ EDC Curves

		Predictions	
		$\hat{Y} = 0$	$\hat{Y} = 1$
Ground Truth	$Y = 0$	TN	FP
	$Y = 1$	FN	TP

False Positive Rate = FP / (FP + TN)

False Negative Rate = FN / (TP + FN)

Each point in the **EDC Curve** corresponds to a classification threshold θ such that **h(x) >** $\boldsymbol{\theta}$ **=>** $\boldsymbol{\hat{Y} = 1}$.



Fairness Metrics ▶ Adapted to Loan Approval

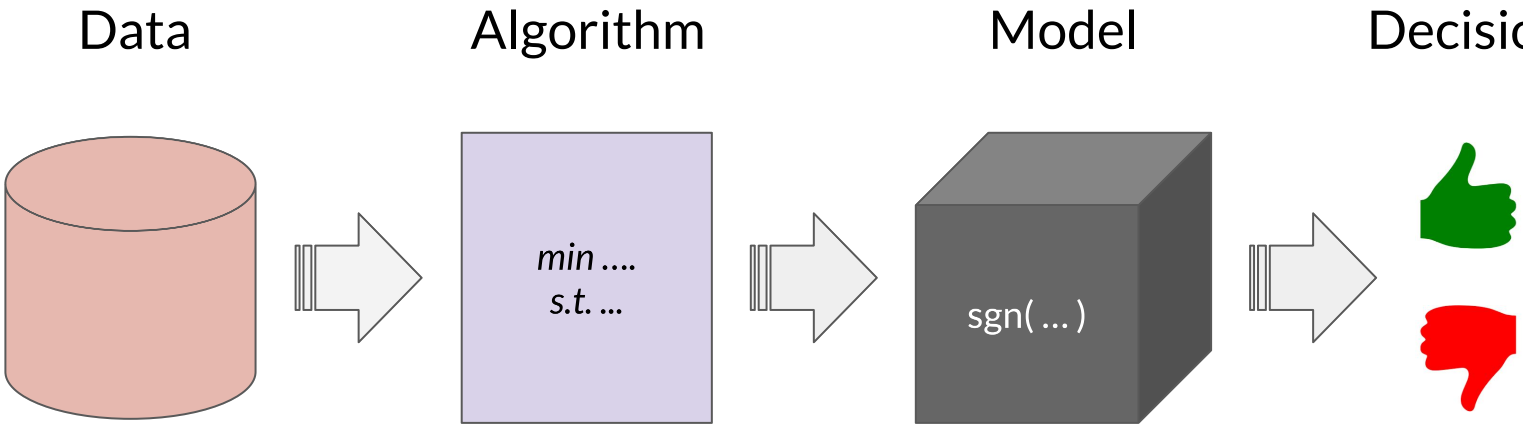
Loan approval is an **assistive** intervention.

Therefore, focus is on protected category **false negatives.**

At the same time, the **approval rate** across privileged and unprivileged groups by must be similar.

Disparate Impact	False Negative Rate Ratio	False Negative Rate Difference
$\dfrac{P(\ \widehat{Y} = 1 \mid A = \text{unprivileged}\ )}{P(\ \widehat{Y} = 1 \mid A = \text{privileged}\ )}$	$\dfrac{FNR_{A=\text{unprivileged}}}{FNR_{A=\text{privileged}}}$	$FNR_{A=\text{unprivileged}} - FNR_{A=\text{privileged}}$

Interventions for Fairness



Pre-Processing

Optimized Pre-Processing
Instance Reweighing
..
..

In-Processing

Exponentiated Gradient Red.
Grid Search Reduction
..
..

Post-Processing

Calibrated Eq. Odds.
Equalized Odds
..
..

Interventions for Fairness ▶ Pre-Processing

Optimized Pre-Processing transforms Data A, X, Y => A, X', Y' such that

1. The dependence of Y' on A is bounded. For any two groups $a_1, a_2 \in A$, the difference in dependencies of Y' on a_1 and a_2 is also bounded.
2. Expected pointwise distortions in the transformation are minimal.
3. Differences (KL-divergence) in the probability distribution underlying the data before and after the transformation are bounded.

Instance Re-Weighing re-weighs each sample in data to ensure $Y \perp A$.

$P(Y = 1, A = a) = P(Y = 1) \times P(A = a)$ and vice versa for $Y = 0$.

Calmon et. al, Optimized pre-processing for discrimination prevention. Neural Information Processing Systems (NeurIPS'17).
Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. Knowledge & Information Systems (2012).

Interventions for Fairness ▶ In-Processing

Exponentiated Gradient Reduction

Trains the classifier subject to constraints towards ensuring one of **demographic parity / equalized odds** for all combinations of protected attributes in data.

Recall that a classifier *h* satisfies:

Demographic Parity $[\, h \perp A \,]$

$$P[\, h(x) = \hat{y} \mid A = a \,] = P[\, h(x) = \hat{y} \,] \;\forall\; a \in A$$

Equalized Odds $[\, h \perp A \mid Y \,]$

$$P[\, h(x) = \hat{y} \mid A = a, Y = y \,] = P[\, h(x) = \hat{y} \mid Y = y \,] \;\forall\; a \in A, y \in Y$$

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J. & Wallach, H.. (2018). A Reductions Approach to Fair Classification. Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research 80:60-69 https://proceedings.mlr.press/v80/agarwal18a.html.

Interventions for Fairness ▶ Post-Processing

Calibrated Odds-Equalizing

Optimizes over calibrated classifier (h) score outputs using a linear program to find probabilities with which to change output labels to satisfy:

Equalized Odds [$h \perp A \mid Y$]

$$P[\, h(x) = \hat{y} \mid A = a, Y = y \,] = P[\, h(x) = \hat{y} \mid Y = y \,] \;\forall\; a \in A, y \in Y$$

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'17).

AI Fairness 360

These are ten state-of-the-art bias mitigation algorithms that can address bias throughout AI systems. Add more!

Optimized Pre-processing
Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing
Use to mitigate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing
Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification
Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover
Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations
Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover
Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing
Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing
Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier
Use to mitigate bias in classifier. Meta algorithm

Algorithms
- aif360.algorithms.preprocessing
- aif360.algorithms.inprocessing
- aif360.algorithms.postprocessing

Project

1. Select a dataset, and a protected attribute (*sex*, *race*, *etc.*).
2. Identify privileged and unprivileged groups, and their base rates.
3. Determine the nature of intervention (*assistive / punitive*) and fairness metrics.
4. Train a classifier and measure its performance on fairness metrics.
5. Use one intervention to improve your model's performance on fairness metrics.

Be descriptive in your answers, justify your choice of metrics, models, and interventions.

If the performance does not improve that is ok! The focus of this exercise is more on how you analyze and make sense of your results.

Project ▶ Bonus [2 points]

How do the fairness properties of your classifier change when you k-anonymize the data?

Report and analyze your results for different values of k.

Does using only a subset of features lead to better performance on fairness metrics?

Explore why - using correlations, or understanding what those features represent.

Add an explainability component to the intervention using LIME / SHAP / Counterfactuals.

What do the differences in model explanations for errors pre and post-debiasing represent?

You only need to do one.

Project ▶ Datasets

<table>
<tr><th>Dataset</th><th>Protected Category</th><th>Group</th></tr>
<tr><td rowspan="3">Adult
- aif360.datasets.AdultDataset

Task: Predict whether annual income of an individual exceeds $50K/year based on census data.</td><td>Sex
{ male / female }</td><td></td></tr>
<tr><td>Race
{ white / non-white }</td><td></td></tr>
<tr><td>Native-Country
{ national / immigrant }</td><td></td></tr>
<tr><td rowspan="2">ProPublica Compas
- aif360.datasets.Compas

Task: Predict whether an individual will re-offend</td><td>Sex
{ male / female }</td><td></td></tr>
<tr><td>Race
{ caucasian / other }</td><td></td></tr>
</table>

Project ▶ Datasets

Dataset	Protected Category	Group
Default of Credit Card Clients - [ucimlrepo (id=350)] ***Task:*** *Predict whether a customer will face the default situation in the next month.*	Sex { male / female }	
	Marital Status { single / married }	
Law School Bar Exam Passage - [Course Github] ***Task:*** *Predict whether an applicant will pass the bar exam in first try (as a proxy for admission).*	Sex { male / female }	
	Race { white / non-white }	

Project ▶ Datasets

Dataset	Protected Category	Group
Covid-19 Open Data Mexico - Course Github ***Task 1:*** *Predict whether an intubated patient will get admitted into intensive care (Intubated patients in ICU have a higher mortality rate).*	Sex { male / female }	
	Age { < 60 / >= 60 }	

Tip: if your dataset is too large for processing, you can take a smaller (stratified) sample