Fairness In Machine Learning
With Tabular Structured Data

Wrap Presentation

Student: Karim Kallich,
Mohamed Eltayeb
Mentor: Narayanan Venkit,
Pranav
Professor: Dr. Ahmed Rashed

Introduction

- **Fairness in machine learning with tabular structured datasets.**
- **Libraries: Fairlearn, AI Fairness 360 (AIF360), and What-If Tool.**

Fairlearn Results

Mitigation Approach	Performance Change Using Accuracy score	Fairness Improvement Using Demographic Parity Diff
Original Model	Baseline (0.86) (Ideal output is 1.00)	Baseline (0.3) (Ideal output is 0.00)
Correlation Remover Pre-Processing	0.86 Same as Original	0.22 Better by 26%
Exponentiated Gradient In-Processing	0.85 Worse by 1.1%	0.05 Better by 83%
Correlation Remover + Exponentiated Gradient	0.84 Worse by 2.3%	0.06 Better by 80%
Grid Search In-Processing	0.79 Worse by 8.1%	0.07 Better by 76%
Correlation Remover + Grid Search	0.64 Worse by 25%	0.5 Worse 66%
Exponentiated Gradient + Grid Search	0.79 Worse by 8.1%	0.06 Better by 80%
Threshold Optimizer Post-Processing	0.81 Worse by 5.8%	0.06 Better by 80%
Correlation Remover + Threshold Optimizer	0.80 Worse by 7%	0.1 Better by 66%
Exponentiated Gradient + Threshold Optimizer	0.85 Worse by 1.1%	0.04 Better by 86%
Grid Search + Threshold Optimizer	0.80 Worse by 7%	0.05 Better by 83%

AIF360 Results

	Mitigation Approach	Performance Change Average Odds Difference (AOD)	Fairness Improvement Statistical Parity Difference (SPD)
		Baseline (0.09) (Ideal output is 0.00)	Baseline (0.13) (Ideal output is 0.00)
Preprocessing	Reweighing	0.007 Better by 8.3%	0.01 Better by 12 %
	Disparate Impact Remover	0.10 Worse by 1%	0.20 Worse by 7%
In-Processing	Adversarial Debiasing	0.01 Better by 8%	0.05 Better by 8%
	Prejudice Remover	0.10 Worse by 1%	0.20 Worse by 2%
Postprocessing	Equalized Odds	0.0001 Better by 9%	0.0 Better by 13%
	Calibrated EqOdds	Error	Error

What-If-Tool Results

Thresholds	Demographic Parity Male (DPM)	Demographic Parity Female (DPF)	Performance Change Using Average Accuracy score	Fairness Improvement Using Demographic Parity Diff = DPM-DPF
Baseline 0	0.303	0.112	0.33	0.19
0.2	0.566	0.551	0.63 30% better	0.01 18% better
0.4	0.628	0.611	0.61 28% better	0.01 18% better
0.6	0.655	0.627	0.55 22% better	0.02 17% better
0.8	0.716	0.686	0.50 17% better	0.03 16% better
0.9	0.702	0.689	0.45 12% better	0.01 18% better
1.0	0	0	0 33% worse	0 19% better