Targeted Removal of Unwanted Biases in Probabilistic Circuits

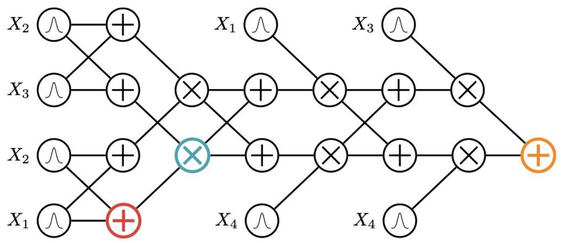
Marko Jojic, Computer Science
Mentor: YooJung Choi, Assistant Professor
Arizona State University

Objective

To learn a Probabilistic Model which **minimizes** the unwanted **bias of decisions** with respect to protected attributes and the ability to reconstruct those attributes.

Background & Motivation

Probabilistic Circuits (PCs) are a class of Neural Networks (NN) that represent probability distributions. PCs offer **tractable exact marginalization.** Sum Product Networks fall into this category, with sum nodes representing mixtures of distributions, and product nodes representing factorizations of joint distributions across subsets of variables. [1]



Unreasonable biases in training data **skew decisions**, especially when evidence is missing. Correlations involving protected attributes can be used to **reconstruct sensitive information**.

Tractable marginalization enables us to:

- **Find dependencies** between variables
- **Measure** the contribution of **biases** in a decision
- Use this measure of bias as a learning objective to **enforce fairness constraints** or domain specific knowledge

Methodology

Given some protected information *x*, and non-protected information *y*, we define *x*'s **Discrimination Score:** [2]

$$|p(d|x,y) - p(d|y)|$$

Preprocessing:

- Discrimination score misses cases where **y is a proxy for x**
- Use a NN to project non-protected attributes into a latent space
- Objective: **preserve predictive power** over decision label, while **hiding** information about **protected attributes**

Training:

- Learn initial PC using Maximum Likelihood Estimation (MLE)
- Find high scoring patterns (*x*,*y*)
- **Minimize discrimination score**, minimize Cross Entropy loss over decision label, maximize likelihood of data given the PC
- Randomly select some patterns to retain for the next round
- **Repeat** finding/constraining patterns **until** the maximum discrimination **score** falls **under some threshold**

Results on COMPAS recidivism dataset

Reconstructing Protected Data from Non Protected Data (NP):

Model	i (Protected)	σ^2 of $P(X_i = 1 \mid NP)$
MLE PC	4	0.029
Constrained PC	4	0.000
MLE PC	5	0.020
Constrained PC	5	0.000
MLE PC	6	0.039
Constrained PC	6	0.000
MLE PC	7	0.020
Constrained PC	7	0.000

$P(X_i \mid NP)$ is a **uniform distribution** after constraints are applied

Results (continued)

Classification Power:

Model	Accuracy	AUC
Naive Bayes	0.8778	0.8988
Logistic Regression	0.8826	0.9190
MLE PC	0.8776	0.9000
Constrained PC	0.8851	0.9130

Equalized Odds (lower is better):

Model	TPR variance	FPR variance
MLE PC	0.003921	0.163241
Constrained PC	0.002446	0.072093

Statistical Parity (lower is better):

Model	Variance of $P(D = 1 \mid S)$
MLE PC	0.009269
Constrained PC	0.005925

Future Work

- Extension to continuous data
- Auditing/**debiasing larger models** using a PC

References

[1] Y. Choi, A. Vergari, G. Van den Broeck. "Probabilistic circuits: A unifying framework for tractable probabilistic modeling." 2020.
[2] N. Selvam, G. Van den Broeck, and Y. Choi. "Certifying fairness of probabilistic circuits" AAAI 2023

MORE Masters Opportunity for Research in Engineering

Ira A. Fulton Schools of Engineering Arizona State University