Probabilistic Reasoning for Robust and Fair Decision Making

YooJung Choi
yjchoi@cs.ucla.edu

Motivation

- Trustworthy AI/ML:
  Interpretability, robustness, fairness…
  The behavior of decision-making systems, not just the performance of black-box models
- Reasoning about classifier behavior in the real-world entails uncertainty, missing data…
  ⇒ *Probabilistic reasoning*
- Current approaches (e.g. for fair learning):
  - Fix data ⇒ models can introduce bias
  - Simplifying assumptions or approximations
    ⇒ not good "models of the real world"

Probabilistic Circuits

Why not exact inference/learning?

P_{model}("robust to missing value")
("makes favoring decision")
⋮
⇒ *computationally hard*

Probabilistic Circuits:
expressive and *tractable* probabilistic models

Robust Trimming

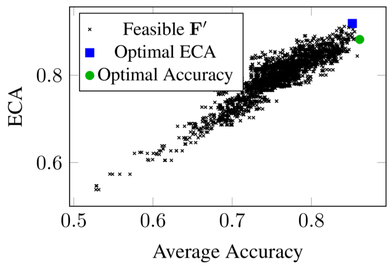
"Make classifications robust to missing features"

Observing features is expensive
⇒ Simplify (trim features) while keeping similar behavior as the original classifier.

$$\text{ECA}(\alpha, \beta) = \sum_{\mathbf{f}} [C_T(\mathbf{f}) = C_{T'}(\mathbf{f}')] \cdot \Pr(\mathbf{f})$$

Expected Classification Agreement:

- Probability of classifier α and trimming β agreeing on instances
- PP$^{\text{PP}}$-complete query, but tractable with probabilistic circuit



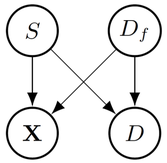
		Agreement	Accuracy
pima	Opt. ECA	0.9863	0.7123
	Opt. Acc.	0.9452	0.7260
heart	Opt. ECA	0.9245	0.8491
	Opt. Acc.	0.9057	0.7925

(left) ECA and accuracy of all possible trimmings. (right) test ECA and accuracy of optimal trimmings based on each metric.

YooJung Choi and Guy Van den Broeck. "On Robust Trimming of Bayesian Network Classifiers". In IJCAI 2018

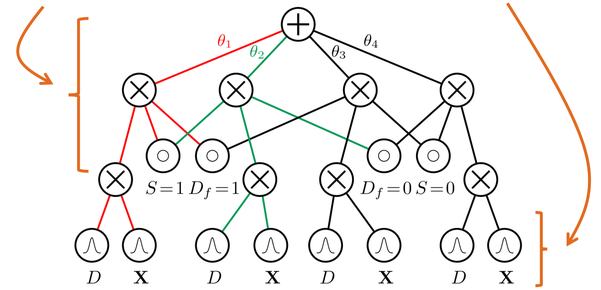
Fair PC

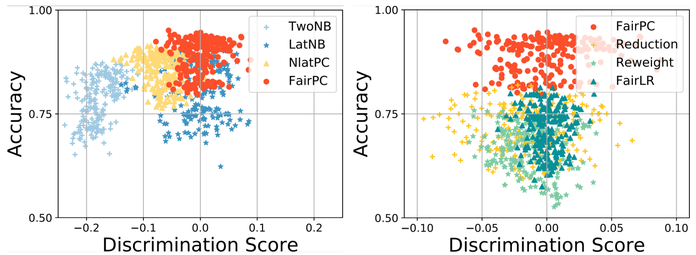
"Learn fair classifiers from biased labels"



Observed labels (D) come from latent fair labels (D_f) with sensitive attribute (S) dependent noise

Learn a probabilistic distribution that models the ***bias mechanism*** and ***best fits the data***





Accuracy and demographic parity violation of FairPC and fair probabilistic (left) and classification (right) methods.

YooJung Choi, Meihua Dang, and Guy Van den Broeck. "Group Fairness by Probabilistic Modeling with Latent Fair Decisions". arXiv:2009.09031