Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns

YooJung Choi[1]* **Golnoosh Farnadi**[2,3]* **Behrouz Babaki**[4]* **Guy Van den Broeck**[1]

[1]University of California, Los Angeles, [2]Mila, [3]Université de Montréal, [4]Polytechnique Montréal
Email: yjchoi@cs.ucla.edu

Motivation

Existing methods to ensure fairness often assume a fixed set of observable features to define individuals. However, in practice it is common for certain features to be **missing at prediction time**.

Naive Bayes classifiers can naturally handle partial observations by treating classification as (marginal) inference tasks. We study their fairness properties by explicitly taking into account predictions with missing sensitive and non-sensitive attributes.

Discrimination Patterns

- For joint assignments $\mathbf{x}$ and $\mathbf{y}$ to $\mathbf{X}$, a subset of sensitive attributes (e.g., gender, race) and $\mathbf{Y}$, a subset of remaining variables, the **degree of discrimination** of $\mathbf{xy}$ is:

$$\Delta_{P,d}(\mathbf{x}, \mathbf{y}) \triangleq P(d \mid \mathbf{xy}) - P(d \mid \mathbf{y}).$$

  "How much does the classification change by disclosing some sensitive attributes?"

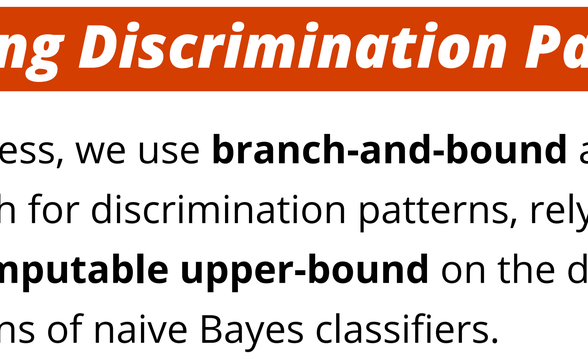
- Joint assignments $\mathbf{xy}$ form a **discrimination pattern** if $|\Delta_{P,d}(\mathbf{x}, \mathbf{y})| > \delta$.
- A distribution P is δ**-fair** if there exists no discrimination pattern w.r.t. P and δ.

Discovering Discrimination Patterns

To verify δ-fairness, we use **branch-and-bound** algorithm to efficiently search for discrimination patterns, relying on a **linear-time computable upper-bound** on the discrimination score for patterns of naive Bayes classifiers.

For more interpretable results, we want a small set of **"interesting" discrimination patterns**. In addition to discrimination score, we also rank patterns by **divergence score** to take into account the probability of a pattern:

$$\begin{aligned}\mathrm{Div}_{P,d,\delta}(\mathbf{x}, \mathbf{y}) \triangleq \min_{Q} \ & \mathrm{KL}\left(P \parallel Q\right) \\ \text{s.t.} \ & |\Delta_{Q,d}(\mathbf{x}, \mathbf{y})| \leq \delta \\ & P(d\mathbf{z}) = Q(d\mathbf{z}), \ \forall \, d\mathbf{z} \not\models \mathbf{xy}\end{aligned}$$



All discrimination patterns on COMPAS dataset. Top-10 patterns according to discrimination score, divergence score, and probability are highlighted.

- Search algorithm visits only a small fraction of patterns:

Dataset	S	N	# Pat.	k	Proportion of search space explored: Divergence $\delta = 0.01$	Divergence $\delta = 0.10$	Discrimination $\delta = 0.01$	Discrimination $\delta = 0.10$
COMPAS	4	3	15K	1	6.387e-01	3.874e-01	8.188e-03	8.188e-03
				100	8.222e-01	4.335e-01	9.914e-02	9.914e-02
Adult	4	9	11M	1	3.052e-06	1.248e-05	2.451e-04	2.451e-04
				100	1.458e-05	2.509e-05	2.600e-04	2.597e-04
German	4	16	23B	1	5.075e-07	2.374e-06	7.450e-08	7.450e-08
				100	1.454e-06	3.407e-06	5.897e-06	5.897e-06

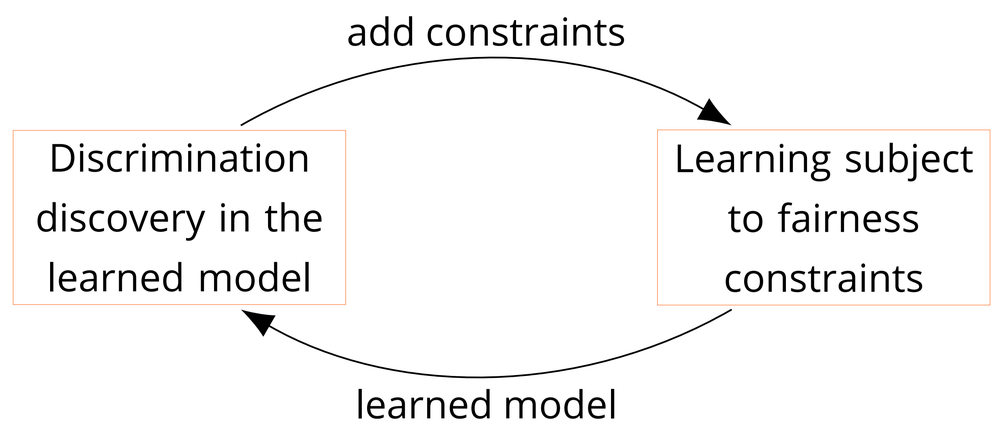
Learning δ-Fair Naive Bayes

Maximum-likelihood learning with fairness constraints:

$$\begin{aligned}\operatorname*{argmax}_{\theta} & \prod_{\mathbf{x} \in \mathcal{D}} P(\mathbf{x}; \theta) \\ \text{s.t.} \quad & \left|\Delta_{P_\theta,d}(\mathbf{x}, \mathbf{y})\right| \leq \delta \quad \forall \mathbf{x}, \mathbf{y}\end{aligned}$$

For naive Bayes classifiers, the optimization can be formulated as **signomial programs**, whose local optima can be computed efficiently.

It is highly inefficient to enumerate and solve the optimization with exponentially many fairness constraints. Instead, we take an **iterative approach**:



- Log-likelihood close to that of unconstrained, unfair model:

Dataset	Unconstrained	δ-fair	Independent
COMPAS	-207,055	-207,395	-208,639
Adult	-226,375	-228,763	-232,180
German	-12,630	-12,635	-12,649

- Higher accuracy than other fairness methods:

Dataset	Unconstrained	2NB	Repaired	δ-fair
COMPAS	0.880	0.875	0.878	0.879
Adult	0.811	0.759	0.325	0.827
German	0.690	0.679	0.688	0.696

Paper: arxiv.org/abs/1906.03843
Code: github.com/UCLA-StarAI/LearnFairNB