Distributionally Robust Optimization to Improve Fairness Generalization in Machine Learning

Julien Ferry, PhD Student

Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet and Mohamed Siala

jferry@laas.fr

April 30, 2021

ROADEF 2021

Table of Contents

1. Background
2. Distributionally Robust Fair Learning
3. Experimental Evaluation
4. Conclusion

1 Background
- Supervised Learning
- Fairness in Machine Learning
- Distributionally Robust Optimization

2 Distributionally Robust Fair Learning

3 Experimental Evaluation

4 Conclusion

Notations - Classification

Let $\mathcal{D} = (X, Y)$ be a dataset. Each example $e_{i, i \in [1..|\mathcal{D}|]} = (x_i, y_i) \in \mathcal{D}$, where:

- x_i is the vector of attributes
- y_i is the label associated to e_i

Classification: Problem Formulation

Given training dataset $\mathcal{D}$ drawn from an (unknown) underlying distribution $\mathcal{P}$, and hypothesis class $\mathcal{H}$, the objective of a supervised learning algorithm is to build a model $h \in \mathcal{H}$ solution to the following optimization problem:

$$\arg\min_{h \in \mathcal{H}} \quad f_{obj}(h, \mathcal{D}) \tag{1}$$

Let $\hat{y}_i$ be the prediction of classifier h for example e_i

The problem of dataset bias

- Supervised learning models learn correlations contained in the training data
- What if some correlations are undesirable or not relevant ?

Gender	Education	Age	Income >50K$
Male	Master	25	No
Female	Master	25	No
Female	Dropout	50	No
Male	Dropout	50	No
Male	Master	50	Yes
Female	Master	50	No

Table: Example of biased dataset

Group/Statistical Fairness in Machine Learning

- Features space is partitioned into *sensitive and unsensitive attributes*: each example $e_{i, i \in [1..|\mathcal{D}|]} = (x_i, a_i, y_i) \in \mathcal{D}$, where:
  - x_i is the vector of unsensitive attributes
  - a_i is the vector of sensitive attributes, defining e_i's membership to *protected groups*
  - y_i is the label associated à e_i
- Main principle: ensure that some measure *differs by no more than* ϵ between several *protected groups*
- Many metrics proposed, depending on the measure to be equalized
  - e.g., Statistical Parity: Equalize probability of being assigned to the positive class:

$$\forall j, \forall k : |P(\hat{y} = 1 | a = j) - P(\hat{y} = 1 | a = k)| \leq \epsilon$$

  - e.g., Equal Opportunity: Equalize false negative rates:

$$\forall j, \forall k : |P(\hat{y} = 0 | y = 1, a = j) - P(\hat{y} = 0 | y = 1, a = k)| \leq \epsilon$$

Supervised Fair Learning: A Bi-Objective Optimization Problem

- Let $\text{unf}(\cdot)$ be an unfairness oracle. A common formulation of the Fair Learning problem is:

$$\begin{aligned} \underset{h \in \mathcal{H}}{\arg\min} \quad & f_{obj}(h, \mathcal{D}) \\ \text{s.t.} \quad & \text{unf}(h, \mathcal{D}) \leq \epsilon \end{aligned} \tag{2}$$

where one wants to build model h minimizing objective function f_{obj} and exhibiting unfairness at most ϵ (on training dataset $\mathcal{D}$)

Distributionally Robust Optimization (DRO)

- Instead of minimizing objective function f_{obj} for a given distribution $\mathcal{P}$, DRO aims at minimizing f_{obj} for a worst-case distribution among a set of perturbations of $\mathcal{P}$ [Sagawa et al., 2019]
- Such neighbouring distributions define a *perturbation set* $\mathcal{B}(\mathcal{P})$
- The problem of distributionally robust supervised learning can be rewritten as:

$$\underset{h \in \mathcal{H}}{\arg\min} \max_{\mathcal{Q} \sim \mathcal{B}(\mathcal{P})} f_{obj}(h, \mathcal{Q}) \tag{3}$$

1 Background

2 Distributionally Robust Fair Learning
- Fairness Generalization - Related Work
- Proposed Formulation

3 Experimental Evaluation

4 Conclusion

Fairness Generalization

- Does fairness on training data imply fairness on unseen data?
  - In practice, it is often not the case, and fairness constraints *overfitting* can occur [Cotter et al., 2018, 2019]

Related Work

- Methods have been proposed recently to address this issue: [Cotter et al., 2018, 2019; Chuang and Mroueh, 2021; Huang and Vishnoi, 2019; Mandal et al., 2020; Sagawa et al., 2019; Taskesen et al., 2020; Wang et al., 2021]
- Such methods often present applicability and/or scalability limits

Intuition

- Each subset of $\mathcal{D}$ with sufficiently important size has a distribution slightly different to that of $\mathcal{D}$
- Hence, ensuring fairness on $\mathcal{D}$, but also on some of its subsets is a form of distributionally robust optimization

Formalization

- We consider n random binary masks, defining n random subsets of the training set
- Each mask $\mathcal{M}_i$ is a vector of size $|\mathcal{D}|$, where each coordinate $\mathcal{M}_{ij} \in \{0, 1\}$ indicates whether example e_j belongs to the i^{th} subset
- We define our perturbation set as:
$\mathcal{B}(\mathcal{D}, n) = \{\mathcal{D}\} \cup \{\mathcal{D}_{i, i\in[1..n]} | \, \forall e_j \in \mathcal{D}_i, e_j \in \mathcal{D} \wedge \mathcal{M}_{ij} = 1\}$
- Our formulation of the Distributionally Robust Fair Learning problem is:

$$\begin{aligned} \underset{h \in \mathcal{H}}{\arg\min} \quad & f_{obj}(h, \mathcal{D}) \\ \text{s.t.} \quad & \max_{\forall \mathcal{D}' \in \mathcal{B}(\mathcal{D}, d)} \text{unf}(h, \mathcal{D}') \leq \epsilon \end{aligned} \tag{4}$$

Table of Contents

1. Background
2. Distributionally Robust Fair Learning
3. Experimental Evaluation
    - Integration into an Existing Method
    - Experimental Setup
    - Results
4. Conclusion

Distributionally Robust FairCORELS

- Based on the source code of FairCORELS[a] [Aïvodji et al., 2019]
- Finds model r solution to the following problem:

$$\begin{aligned} \arg\min_{r \in \mathcal{R}} \quad & f_{obj\texttt{FairCORELS}}(r, \mathcal{D}) \\ \text{s.t.} \quad & \text{unf}(h, \mathcal{D}) \leq \epsilon \\ & \max_{\forall \mathcal{D}' \in \mathcal{B}(\mathcal{D}, n)} \text{unf}(h, \mathcal{D}') \leq \epsilon \end{aligned}$$

- Can be implemented without significant running time overhead

[a] https://github.com/ferryjul/fairCORELS

Setup description

- We compare:
  - The original `FairCORELS` [Aïvodji et al., 2019]
  - Our Distributionally Robust `FairCORELS` for $n = 10$ masks
  - Our Distributionally Robust `FairCORELS` for $n = 30$ masks
- For each method, we generate sets of solutions with different accuracy/fairness tradeoffs, by varying the fairness constraint
- We repeat the experiment for:
  - Five fairness metrics:
    - Statistical Parity [Dwork et al., 2012]
    - Predictive Parity [Chouldechova, 2017]
    - Predictive Equality [Chouldechova, 2017]
    - Equal Opportunity [Hardt et al., 2016]
    - Equalized Odds [Hardt et al., 2016]
  - Four biased datasets:
    - Adult Income dataset [Frank and Asuncion, 2010]
    - COMPAS dataset [Angwin et al., 2016]
    - Default Credit dataset [Yeh and Lien, 2009]
    - Bank Marketing dataset [Moro et al., 2014]

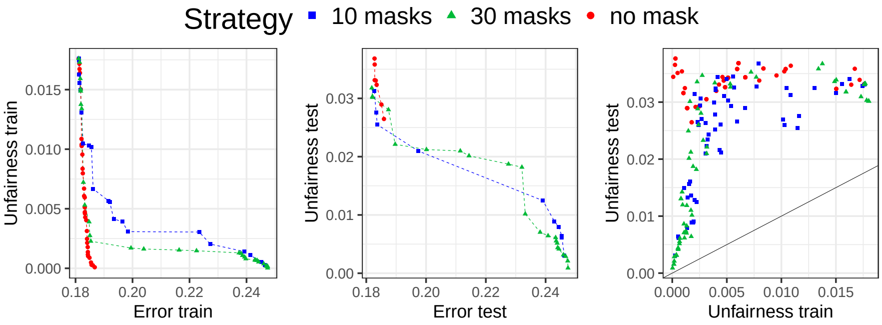

Figure: Results obtained on the Adult Income dataset, for the Equal Opportunity metric

Table of Contents

1. Background
2. Distributionally Robust Fair Learning
3. Experimental Evaluation
4. Conclusion

We propose a heuristic approach to Distributionally Robust and Fair Learning that:

- Benefits from its simplicity in terms of
  - Integrability
  - Scalability
  - Genericity
- Practically improves fairness generalization

Perspectives

- Study the effect on fairness generalization of:
  - the number of masks n
  - the size of the random subsets
- Integration into other existing fair learning algorithms

Aïvodji, U., Ferry, J., Gambs, S., Huguet, M.-J., and Siala, M. (2019). Learning fair rule lists. *arXiv preprint arXiv:1909.03977*.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017a). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44. ACM.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017b). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1):8753–8830.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chuang, C.-Y. and Mroueh, Y. (2021). Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*.

Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. (2018). Training fairness-constrained classifiers to generalize.

Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. (2019). Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR.

References II

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California. *School of information and computer science*, 213:2–2.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Huang, L. and Vishnoi, N. (2019). Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR.

Mandal, D., Deng, S., Jana, S., Wing, J., and Hsu, D. J. (2020). Ensuring fairness beyond the training data. *Advances in Neural Information Processing Systems*, 33.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Rivest, R. L. (1987). Learning decision lists. *Machine learning*, 2(3):229–246.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. (2020). A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*.

Wang, Y., Nguyen, V. A., and Hanasusanto, G. A. (2021). Wasserstein robust support vector machines with fairness constraints. *arXiv preprint arXiv:2103.06828*.

Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480.

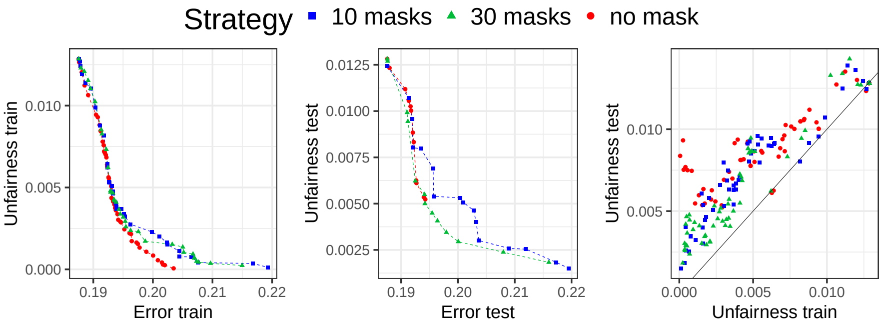

Figure: Results obtained on the Default Credit dataset, for the Predictive Equality metric

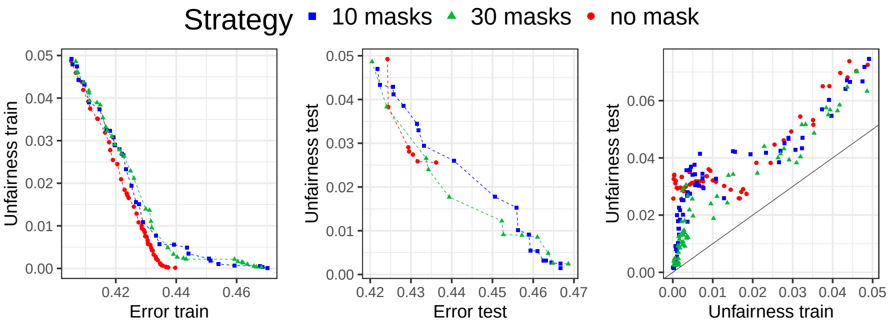

Figure: Results obtained on the COMPAS dataset, for the Statistical Parity metric

Rule Lists: Definition

Rule lists [Rivest, 1987] are classifiers formed by an ordered list of *if-then* rules with antecedents in the *if* clauses and predictions in the *then* clauses. More precisely, a rule list $r = (\{p_{k,k \in \{1..K\}}\}, \{q_{k,k \in \{1..K\}}\}, q_0)$ consists of K distinct association rules $p_k \rightarrow q_k$, in which p_k is the antecedent of the association rule and q_k its associated consequent, followed by a default prediction q_0.

A possible rule list for the example dataset of slide 3 (with 100% accuracy)

```
if [Education:Dropout] then [low]
else if [Gender:Male AND Age>45] then [high]
else [low]
```

FairCORELS Problem Formulation

- Based on the CORELS algorithm [Angelino et al., 2017a,b]
- FairCORELS [Aïvodji et al., 2019] returns rule list r^* that is a solution to the following problem:

$$\begin{aligned} \arg\min_{r \in \mathcal{R}} \quad & \text{misc}(h, \mathcal{D}) + \lambda . K_r \\ \text{s.t.} \quad & \text{unf}(h, \mathcal{D}) \leq \epsilon \end{aligned}$$

where:

- $\mathcal{R}$ is the space of rule lists
- $\mathcal{D}$ denotes the training dataset
- K_r is the length of rule list r
- λ is a regularization parameter balancing sparsity and accuracy
- misc$(\cdot)$ is the misclassification error and unf$(\cdot)$ measures unfairness

FairCORELS search space

- FairCORELS represents the search space of rule lists as a prefix tree (trie)
- FairCORELS leverages several bounds to efficiently explore this search space (including CORELS' original bounds)

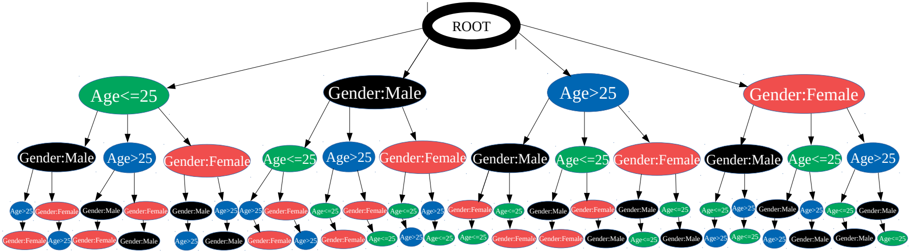

Figure: Example prefix tree with 4 attributes