

Empirical Risk Minimization Under Fairness Constraints

Massimiliano Pontil

Computational Statistics and Machine Learning
Istituto Italiano di Tecnologia
and
Department of Computer Science
University College London

(joint work with *Michele Donini, Luca Oneto, Shai Ben-David, & John Shawe-Taylor*)

Operator Splitting Methods in Data Analysis Workshop
Flatiron Institute, March 21, 2019

Unfair AI

- ▶ We address the question of fairness in machine learning: preventing that sensitive variables 'unfairly' influence the outcome of a classifier



Statistical Learning Setting

- ▶ Let μ be a prob. distribution on $Z \times S \times Y$, where $Y = \{-1, +1\}$ and $S = \{a, b\}$ represents group membership among two groups. Also let $X = Z \times S$
- ▶ Let $\ell : \mathbb{R} \times Y \rightarrow \mathbb{R}$ be a loss function and define the risk of a function $f : X \rightarrow Y$ as

$$L(f) = \mathbb{E}[\ell(f(x), y)]$$

- ▶ Define the risk of the positive labeled samples in group g

$$L^{+,s}(f) = \mathbb{E}[\ell(f(x), y) | y = 1, s = g]$$

- ▶ **Definition.** Let $\epsilon \in [0, 1]$. We say that function f is ϵ -fair if

$$|L^{+,a}(f) - L^{+,b}(f)| \leq \epsilon$$

Problem

- We wish to minimize the risk subject to the fairness constraint:

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & L(f) \\ \text{s.t.} \quad & |L^{+,a}(f) - L^{+,b}(f)| \leq \epsilon \end{aligned} \tag{1}$$

- If $\epsilon = 0$ and ℓ is the 01-loss function, our definition is equivalent to notion of Equal Opportunity [Hardt et al., 2016]:

$$\mathbb{P} \{f(x) > 0 \mid y=1, s=a\} = \mathbb{P} \{f(x) > 0 \mid y=1, s=b\} \tag{2}$$

Fair Empirical Risk Minimization (FERM)

- ▶ Distribution μ is unknown and only have a data sequence $(x_i, s_i, y_i)_{i=1}^n$ sampled independently from μ
- ▶ We consider the empirical problem

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & \hat{L}(f) \\ \text{s.t.} \quad & |\hat{L}^{+,a}(f) - \hat{L}^{+,b}(f)| \leq \hat{\epsilon} \end{aligned} \tag{3}$$

where $\hat{\epsilon}$ is a parameter linked to ϵ

- ▶ Empirical risk $\hat{L}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i)$
- ▶ Empirical risk for the positive samples in group g

$$\hat{L}^{+,g}(f) = \frac{1}{n^{+,g}} \sum_{i \in \mathcal{I}^{+,g}} \ell(y_i, f(x_i))$$

with $\mathcal{I}^{+,g} = \{i : y_i=1, s_i=g\}$ and $n^{+,g} = |\mathcal{I}^{+,g}|$, $g \in \{a, b\}$

Statistical Analysis of FERM

We say a class of functions $\mathcal{F}$ is learnable (wrt. loss ℓ) if:

$$\sup_{f \in \mathcal{F}} |L(f) - \hat{L}(f)| \leq B(\delta, n, \mathcal{F}), \quad \text{with} \quad \lim_{n \rightarrow \infty} B(\delta, n, \mathcal{F}) = 0$$

Proposition 1. Let $\delta \in (0, 1)$. If $\mathcal{F}$ is learnable, f^* solves (1) and $\hat{f}$ solves (3) with

$$\hat{\epsilon} = \epsilon + \sum_{g \in \{a, b\}} B(\delta, n^{+,g}, \mathcal{F})$$

then with prob. $\geq 1 - 6\delta$ it holds simultaneously:

$$L(\hat{f}) - L(f^*) \leq 2B(\delta, n, \mathcal{F})$$

$$|L^{+,a}(\hat{f}) - L^{+,b}(\hat{f})| \leq \epsilon + 2 \sum_{g \in \{a, b\}} B(\delta, n^{+,g}, \mathcal{F})$$

Implications of the Bound

- ▶ Bound implies $\hat{f}$ is close to f^* both in term of the risk and fairness constraint
- ▶ But how do we find $\hat{f}$? We would like to solve problem (3) for the hard (misclassification) loss:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n 1\{f(x_i) \neq y_i\} \tag{4}$$
$$\left| \hat{\mathbb{P}}\{f(x) > 0 | y=1, s=a\} - \hat{\mathbb{P}}\{f(x) > 0 | y=1, s=b\} \right| \leq \epsilon$$

- ▶ The fairness constraint extends the EO constraint (2) which we call Difference of Equal Opportunity (DEO)

Convex Relaxation

- Replace the hard loss in the risk with the (larger) hinge loss, and the hard loss in the fairness constraint with a linear loss

$$\min \left\{ \hat{L}_c(f) : f \in \mathcal{F}, \quad |\hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f)| \leq \hat{\epsilon} \right\} \quad (5)$$

- The constraints with the hard and linear loss are related by

$$|\hat{L}_h^{+,a}(f) - \hat{L}_h^{+,b}(f)| \leq |\hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f)| + \hat{\Delta}(f)$$

$$\text{where } \hat{\Delta}(f) = \frac{1}{2} \sum_{g \in \{a,b\}} \left| \hat{\mathbb{E}} [\text{sign}(f(x)) - f(x) | y=1, s=g] \right|$$

- Measuring $\hat{\Delta}(\hat{f}_c)$ quantifies how well $\hat{f}_c^*$ approximates $\hat{f}_h$, but does $\hat{f}_c$ approximate well a solution of problem (4)?

Fair Learning with Kernels

- ▶ Consider the model $f(\cdot) = \langle w, \phi(\cdot) \rangle$ where $\phi : \mathcal{X} \rightarrow \mathbb{H}$ is a kernel-induced feature mapping
- ▶ For the linear loss, the fairness constraint takes the form $|\langle w, u_a - u_b \rangle| \leq \hat{\epsilon}$, where u_g is the barycenter of the positive points in group g :

$$u_g = \frac{1}{n^{+,g}} \sum_{i \in \mathcal{I}^{+,g}} \phi(x_i), \quad g \in \{a, b\}$$

- ▶ We consider the regularization problem (for some $\lambda > 0$)

$$\min_{w \in \mathbb{H}} \sum_{i=1}^n \ell(y_i \langle w, \phi(x_i) \rangle) + \lambda \|w\|^2 \quad \text{s.t.} \quad |\langle w, u_a - u_b \rangle| \leq \hat{\epsilon}$$

Fair Learning with Kernels

- By the representer theorem

$$f(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$$

where $K(x, x_j) = \langle \phi(x), \phi(x_j) \rangle$ and $\alpha = (\alpha_i)_{i=1}^n$ solves

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^n} \quad & \sum_{i=1}^n \ell \left(\sum_{j=1}^n K(x_i, x_j) \alpha_j, y_i \right) + \lambda \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) \\ \text{s.t.} \quad & \left| \sum_{i=1}^n \alpha_i \left[\frac{1}{n^{+,a}} \sum_{j \in \mathcal{I}^{+,a}} K(x_i, x_j) - \frac{1}{n^{+,b}} \sum_{j \in \mathcal{I}^{+,b}} K(x_i, x_j) \right] \right| \leq \hat{\epsilon} \end{aligned}$$

Continuous Sensitive Variables and Outputs

$$\begin{aligned} \min_{w \in \mathbb{H}} \quad & \sum_{i=1}^n \ell(\langle w, \phi(x_i) \rangle, y_i) + \lambda \|w\|^2 \\ \text{s.t.} \quad & \|A^\top w\|_1 \leq \hat{\epsilon}, \end{aligned}$$

The matrix A is built from the data (see 2019 arxiv paper), and it approximates the constraint

$$\frac{1}{KQ^2} \sum_{k=1}^K \sum_{p,q=1}^Q \left| L_k^{k,p}(f) - L_k^{k,q}(f) \right| \leq \epsilon.$$

where

$$L_k^{k,q}(f) = \mathbb{E}[\ell(f(x), y) | y \in [t_k, t_{k+1}), s \in [\sigma_q, \sigma_{q+1})], \quad \forall 1 \leq k \leq K, 1 \leq q \leq Q$$

Modified Validation Procedure

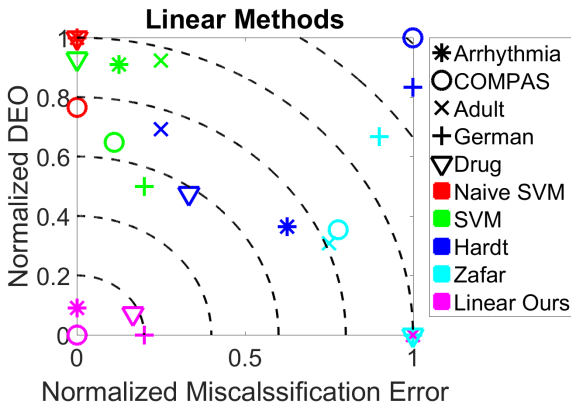
- ▶ In experiments, we employ a two steps 10-fold CV procedure:
- ▶ Step 1: we shortlist the hyperparameter with accuracy above a certain percentage (we choose 90%) of the best accuracy
- ▶ Step 2, from this list, we select the hyperparameter with the highest fairness (i.e. lowest DEO)

Benchmark Experiments

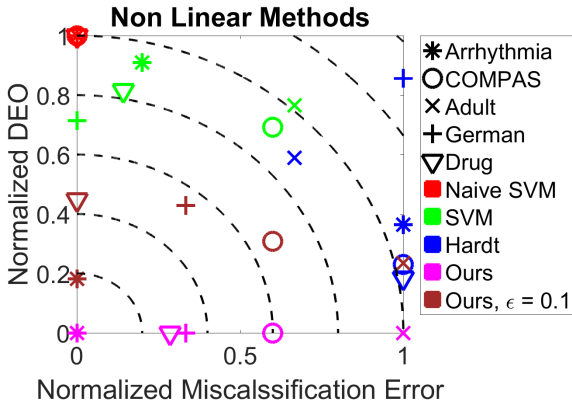
- ▶ Naïve SVM, validated with a standard nested 10-fold cross validation (only optimize accuracy ignoring fairness)
- ▶ SVM with the novel validation procedure. This baseline is the simplest way to inject the fairness into the model
- ▶ The method by Hardt et al. [2016] applied to the best SVM
- ▶ The method Zafar et al. [2017], implemented with the code provided by the authors for the linear case¹

¹Python code: <https://github.com/mbilalzafar/fair-classification>

Results



Results (cont.)



Using Multitask Regularizers

- ▶ Consider that $f(x, s) = \langle w_s, x \rangle$ and formulate the problem as a multitask (matrix) learning one

$$\min_{w_1, \dots, w_k \in \mathbb{H}} \sum_{i=1}^n \ell(\langle w_{s_i}, x_i \rangle, y_i) + \lambda \Omega(w_1, \dots, w_k)$$

with additional fairness constraints

- ▶ Different regularizers from the MTL literature can be used to encourage similarities between the groups:
 - variance or quadratic
 - group lasso /structure sparsity
 - trace norm or other spectral functions

Regularization Around a Common Mean

We build on regularization "around a common mean" to jointly learn a shared and group specific models

$$\min_{w_0, w_1, \dots, w_k \in \mathbb{H}} \frac{1}{k} \sum_{s=1}^k \hat{L}_s(w_s) + \rho \left[\frac{\lambda}{k} \sum_{s=1}^k \|w_s - w_0\|^2 + (1 - \lambda) \|w_0\|^2 \right]$$

$$\text{s.t. } \|Aw_0\|_1 \leq \epsilon$$

$$\|A_k(w_k - w_1)\|_1 \leq \epsilon, \quad k \in \{2, \dots, k\}$$

$\hat{L}_s$ is the empirical risk on the s -th group

Can also add a term $\hat{L}(w_0)$

Experiments (1)

Comparison between

- ($S = 0$) the shared model trained with MTL, with fairness constraint, and no sensitive feature in the predictors
- ($S = 1$) the group specific models trained with MTL, with fairness constraint, the sensitive feature exploited as predictor
- **BUMP IN ACCURACY ($S = 1$)**

	S	MTL		MTL		MTL	
		ACC	DEOp ⁺	ACC	DEOp ⁻	ACC	DEOd

Adult Dataset

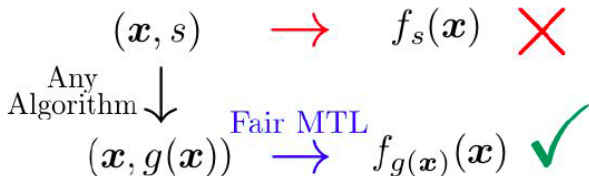
G	0	81.8	0.06	82.7	0.05	82.0	0.06
	1	88.1	0.03	89.1	0.03	88.3	0.03
R	0	82.6	0.01	83.5	0.01	82.8	0.01
	1	90.4	0.03	91.3	0.03	90.6	0.03
G+R	0	83.2	0.04	83.9	0.04	83.5	0.04
	1	90.0	0.05	90.8	0.05	90.3	0.05

COMPAS Dataset

G	0	76.5	0.03	76.4	0.03	75.7	0.03
	1	82.9	0.07	82.8	0.06	82.1	0.06
R	0	82.4	0.03	83.3	0.03	82.6	0.03
	1	90.0	0.03	91.0	0.03	90.2	0.03
G+R	0	83.1	0.05	83.8	0.05	83.4	0.05
	1	89.9	0.05	90.7	0.05	90.3	0.05

A Two-Stage Approach

- Optimize model accuracy and fairness without explicitly using the sensitive feature in the functional form of the model
- Our method is based on two key ideas
 - Use MTL enhanced with fairness constraints to jointly learn group specific classifiers that leverage information between sensitive groups
 - Since learning group specific models might not be permitted, we propose to first predict the sensitive features by any learning method and then to use the predicted sensitive feature



Experiments (2)

Comparison between

- The group specific models trained with MTL, with fairness constraint, and the true sensitive feature exploited as a predictor ($P = 0$)
- Against the same model when the predicted sensitive feature exploited as predictor ($P = 1$)
- BUMP IN FAIRNESS ($P = 1$)**
- MILD DECREASE IN ACCURACY ($P=1$)**

	P	MTL		MTL		MTL	
		ACC	DEOp ⁺	ACC	DEOp ⁻	ACC	DEOd
Adult Dataset							
G	0	88.1	0.03	89.1	0.03	88.3	0.03
	1	87.4	0.01	88.3	0.01	87.6	0.01
R	0	90.4	0.03	91.3	0.03	90.6	0.03
	1	89.2	0.01	90.2	0.01	89.4	0.01
G+R	0	90.0	0.05	90.8	0.05	90.3	0.05
	1	89.0	0.01	89.8	0.01	89.3	0.01
COMPAS Dataset							
G	0	82.9	0.07	82.8	0.06	82.1	0.06
	1	82.1	0.01	82.0	0.01	81.3	0.01
R	0	90.0	0.03	91.0	0.03	90.2	0.03
	1	89.0	0.01	89.9	0.01	89.2	0.01
G+R	0	89.9	0.05	90.7	0.05	90.3	0.05
	1	89.0	0.01	89.8	0.01	89.3	0.01

References and Code

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, & M. P. Empirical risk minimization under fairness constraints. NIPS 2018.

L. Oneto, M. Donini, A. Elders, & M. P. Taking advantage of multitask learning for fair classification. AIES 2019.

L. Oneto, M. Donini, A. & M. P. General Fair Empirical Risk Minimization. arXiv:1901.10080.

Code: https://github.com/jmikko/fair_ERM

Previous work:

C. Dwork, N. Immorlica, A. T. Kalai, and M. D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 2018.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

Open Problems / Research Directions

- ▶ Improve the convex relaxation
- ▶ Extension to continuous sensitive and output variables
- ▶ Multitask learning plus fairness constraint
- ▶ What if the data is observed sequentially?