Fairness-Aware Learning for Continuous Attributes and Treatments

Jérémie Mary, Criteo AI Lab
Clément Calauzènes, Criteo AI Lab
Noureddine El Karoui, Criteo AI Lab and UC, Berkeley

ICML 2019, Long Beach, CA

Fairness and independence

Setup build prediction $\hat{Y}$ of variable Y (e.g. payment default) based on available information X (credit card history); prediction may be biased/unfair wrt sensitive attribute Z (gender).
Most fairness work restricted to binary values of Y and Z.

Fairness and independence

Setup build prediction $\hat{Y}$ of variable Y (e.g. payment default) based on available information X (credit card history); prediction may be biased/unfair wrt sensitive attribute Z (gender).
Most fairness work restricted to binary values of Y and Z.

$$\mathrm{DEO} = \mathbb{P}(\hat{Y}{=}1|Z{=}1, Y{=}1) - \mathbb{P}(\hat{Y}{=}1|Z{=}0, Y{=}1) \text{Equal Opportunity}$$

$$\mathrm{DI} = \frac{\mathbb{P}(\hat{Y}{=}1|Z{=}0)}{\mathbb{P}(\hat{Y}{=}1|Z{=}1)}\,, \quad \text{disparate impact, demographic parity}$$

Fairness and independence

Setup build prediction $\hat{Y}$ of variable Y (e.g. payment default) based on available information X (credit card history); prediction may be biased/unfair wrt sensitive attribute Z (gender).
Most fairness work restricted to binary values of Y and Z.

Generalizations using independence notions

$$\text{EO} \xrightarrow{\text{generalizes to}} \hat{Y} \perp\!\!\!\perp Z | Y\,, \quad \text{even when } Z \text{ non binary}\,,$$

$$\text{Demographic Parity} \xrightarrow{\text{generalizes to}} \hat{Y} \perp\!\!\!\perp Z\,, \quad \text{even when } Z \text{ non binary}\,.$$

We propose new metrics that also easily generalize to continuous variables.

HGR: measuring independence

Definition (Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient)

Given two random variables $U \in \mathcal{U}$ and $V \in \mathcal{V}$,

$$\boxed{\mathrm{HGR}(U, V) \triangleq \sup_{f,g} \rho(f(U), g(V))} \tag{1}$$

ρ:Pearson's correlation; f, g such that $\mathbf{E}\left[f^2(U)\right], \mathbf{E}\left[g^2(V)\right] < \infty$.

HGR: measuring independence

Definition (Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient)

Given two random variables $U \in \mathcal{U}$ and $V \in \mathcal{V}$,

$$\boxed{\mathrm{HGR}(U, V) \triangleq \sup_{f,g} \rho(f(U), g(V))} \tag{1}$$

ρ:Pearson's correlation; f, g such that $\mathbf{E}\left[f^2(U)\right], \mathbf{E}\left[g^2(V)\right] < \infty$.

- $0 \leq \mathrm{HGR}(U, V) \leq 1$; $\mathrm{HGR}(U, V) = 0$ iff V and U independent.
- If f, g only linear functions, get CCA.
- Connection exploited in RDC, [8] with CCA in RKHS

Information theory and relaxation

Theorem (Witsenhausen'75)

Suppose U and V discrete and let matrix

$$Q(u,v) = \frac{\pi(u,v)}{\sqrt{\pi_U(u)}\sqrt{\pi_V(v)}}\ , \ \textit{then}\ \boxed{\mathrm{HGR}(U,V) = \sigma_2(Q)}\ .$$

$\pi(u,v)$ joint distribution of (U,V); π_U and π_V marginals. σ_2: 2nd largest singular value.

- Upper bound on HGR by χ^2-divergence
- Extends naturally to continuous variables (replace sums by integrals)

Fairness aware learning; Equalized Odds (EO)

Given expected loss $\mathcal{L}$, function class $\mathcal{H}$ and fairness tolerance $\varepsilon > 0$, solve :

$$\operatorname*{argmin}_{h \in \mathcal{H}} \mathcal{L}(h, X, Y) \;\text{ subject to }\; \mathsf{HGR}|_\infty \triangleq ||\mathsf{HGR}(\hat{Y}|Y = y, Z|Y = y)||_\infty \leq \varepsilon$$

Practicals: Relax constraint $\mathsf{HGR}|_\infty \leq \varepsilon$ to get tractable penalty : If

$\chi^2|_1 = \left\|\chi^2 \left(\hat{\pi}(\hat{y}|y, z|y), \hat{\pi}(\hat{y}|y) \otimes \hat{\pi}(z|y)\right)\right\|_1$, this yields

$$\boxed{\operatorname*{argmin}_{h \in \mathcal{H}} \mathcal{L}(h, X, Y) + \lambda \chi^2|_1}$$

Fairness aware learning; Equalized Odds (EO)

Given expected loss $\mathcal{L}$, function class $\mathcal{H}$ and fairness tolerance $\varepsilon > 0$, solve :

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(h, X, Y) \text{ subject to } \quad \text{HGR}|_\infty \triangleq ||\text{HGR}(\hat{Y}|Y = y, Z|Y = y)||_\infty \leq \varepsilon$$

Practicals: Relax constraint $\text{HGR}|_\infty \leq \varepsilon$ to get tractable penalty : If

$\chi^2|_1 = \left\| \chi^2 \left(\hat{\pi}(\hat{y}|y, z|y), \hat{\pi}(\hat{y}|y) \otimes \hat{\pi}(z|y) \right) \right\|_1$, this yields

$$\boxed{\underset{h \in \mathcal{H}}{\operatorname{argmin}} \mathcal{L}(h, X, Y) + \lambda \chi^2|_1}$$

Related work : [2], [5],[9], [4], [1], [3], [6], [11], [7, 10]

Y and Z binary valued: comparison with previous work

Test case use our proposal with neural network to train a classifier such that a binary sensitive Z does not unfairly influence an outcome $\widehat{Y}$. Reproduce and compare experiments from Donini et al. '18 [3].

- Goal: maintain good accuracy while having a smaller DEO.
- Results comparable to state of the art
- Smaller datasets difficult for our proposal. NN effect.

	Arrhythmia		COMPAS		Adult		German		Drug	
Method	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Naïve SVM	75 ± 4	11 ± 3	72 ± 1	14 ± 2	80	9	74 ± 5	12 ± 5	81 ± 2	22 ± 4
SVM	71 ± 5	10 ± 3	73 ± 1	11 ± 2	79	8	74 ± 3	10 ± 6	81 ± 2	22 ± 3
FERM	75 ± 5	5 ± 2	96 ± 1	9 ± 2	77	1	73 ± 4	5 ± 3	79 ± 3	10 ± 5
NN	74 ± 7	19 ± 14	97 ± 0	1 ± 0	84	14	74 ± 4	47 ± 19	79 ± 3	15 ± 16
NN + χ^2	75 ± 6	15 ± 9	96 ± 0	0 ± 0	83	3	73 ± 3	25 ± 14	78 ± 5	0 ± 0

Continuous Case: Criminality Rates

Dataset : UCI `Communities+and+Crime`. 2 sets of experiments, 3 fairness penalties :

- Linear regression (LR), full batches of data
- Deep neural nets (DNN) with mini-batches ($n = 200$; Adam as optimizer)
- Regularization parameter λ varies 2^{-4} to 2^{6}

Continuous Case: Criminality Rates

Dataset : UCI `Communities+and+Crime`. 2 sets of experiments, 3 fairness penalties :

We find :

- DNN improves fairness at lower price than linear models in terms of MSE. Important that fairness penalty be compatible with DNNs
- $\chi^2|_1$ and KL$|_1$ work smoothly with mini-batched stochastic optimization; contrast with baseline $L_2^{\hat{Y}|Z,Y}$ penalty which suffers from mini-batching

Continuous Case: Criminality Rates

Dataset : UCI `Communities+and+Crime`. 2 sets of experiments, 3 fairness penalties :

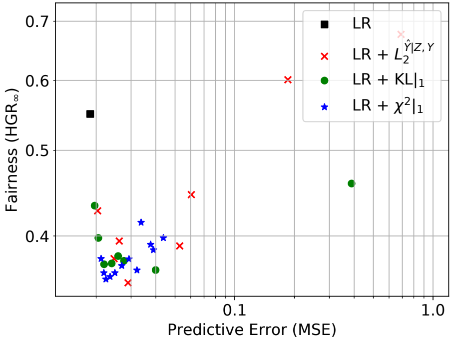

Figure: Equalized odds with Linear Regression: for $\text{KL}|_1$ and $L_2^{\hat{Y}|Z,Y}$ some points out of graph to the right.

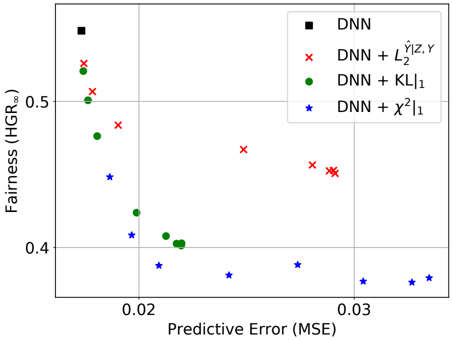

Figure: Equalized odds with DNN

Bechavod, Y., and Ligett, K.
Learning fair classifiers: A regularization-inspired approach.
arXiv pre-print abs/1707.00044 (2017).

Calders, T., and Verwer, S.
Three naive bayes approaches for discrimination-free classification.
Data Mining and Knowledge Discovery 21, 2 (Sep 2010), 277--292.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M.
Empirical risk minimization under fairness constraints.
In Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 2796--2806.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.
Fairness through awareness.
In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (New York, NY, USA, 2012), ITCS '12, ACM, pp. 214--226.

Hardt, M., Price, E., , and Srebro, N.

Equality of opportunity in supervised learning.
In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3315--3323.

KAMISHIMA, T., AKAHO, S., AND SAKUMA, J.
Fairness-aware learning through regularization approach.
In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (Washington, DC, USA, 2011), ICDMW '11, IEEE Computer Society, pp. 643--650.

KOMIYAMA, J., TAKEDA, A., HONDA, J., AND SHIMAO, H.
Nonconvex optimization for regression with fairness constraints.
In Proceedings of the 35th International Conference on Machine Learning (Stockholmsmässan, Stockholm Sweden, 10--15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of Proceedings of Machine Learning Research, PMLR, pp. 2737--2746.

LOPEZ-PAZ, D., HENNIG, P., AND SCHÖLKOPF, B.
The randomized dependence coefficient.

In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 (USA, 2013), NIPS'13, Curran Associates Inc., pp. 1--9.

MENON, A. K., AND WILLIAMSON, R. C.
The cost of fairness in binary classification.
In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (New York, NY, USA, 23--24 Feb 2018), S. A. Friedler and C. Wilson, Eds., vol. 81 of Proceedings of Machine Learning Research, PMLR, pp. 107--118.

SPEICHER, T., HEIDARI, H., GRGIC-HLACA, N., GUMMADI, K. P., SINGLA, A., WELLER, A., AND ZAFAR, M. B.
A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices.
In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (New York, NY, USA, 2018), KDD '18, ACM, pp. 2239--2248.

ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G., AND GUMMADI, K. P.

Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.
arXiv (März 2017).