Pairwise Fairness for Ordinal Regression

Matthäus Kleindessner, Samira Samadi, Muhammad Bilal Zafar,
Krishnaram Kenthapadi, Chris Russell

Ordinal regression

- Multiclass classification over an ordered label set, e.g., {bad, okay, good, excellent}.
- Typically, we incur different costs for different misclassifications (e.g., incorrectly predicting "bad" for "excellent" is more critical than incorrectly predicting "good" for "excellent").
- Algorithms take order information into account; one common approach are threshold models, consisting of a real-valued scoring function and thresholds that bin the score into classes.

Formal setup: k classes $\simeq \{1,\dots,k\} =: [k]$, cost matrix $C \in \mathbb{R}^{k\times k}_{\geq 0}$, datapoints (x,y) are drawn i.i.d. from a distribution $\mathbb{P}$ on $\mathcal{X}\times[k]$
Goal: learn predictor $f:\mathcal{X}\to[k]$ with small expected cost $\mathbb{E}_{(x,y)\sim\mathbb{P}}\, C_{y,f(x)}$
Mean absolute error (MAE): expected cost for $C_{i,j}=|i-j|$

Pairwise fairness notions

- Order of labels carries fairness implications (e.g., misclassifying all "okay" job applicants from majority group as "good" and all "okay" job applicants from minority group as "bad" would be highly unfair).

We initiate the study of fairness for ordinal regression.

We now assume that datapoints come with a protected attribute $a\in\mathcal{A}$ (e.g., gender) and adapt two fairness notions from literature on fair ranking (e.g., Beutel et al. 2019):

Pairwise demographic parity (DP): f satisfies pairwise DP if for all $\tilde{a},\hat{a}\in\mathcal{A}$

$$\mathbb{P}[f(x_1) > f(x_2) | a_1=\tilde{a}, a_2=\hat{a}] = \mathbb{P}[f(x_1) < f(x_2) | a_1=\tilde{a}, a_2=\hat{a}],$$

where the probability is over (x_1,y_1,a_1) and (x_2,y_2,a_2) being independent samples from $\mathbb{P}$ and potentially the randomness of the predictor f.

Pairwise equal opportunity (EO): f satisfies pairwise EO if for all $\tilde{a},\hat{a}\in\mathcal{A}$

$$\mathbb{P}[f(x_1) > f(x_2) | a_1=\tilde{a}, a_2=\hat{a}, y_1>y_2] = \mathbb{P}[f(x_1) < f(x_2) | a_1=\tilde{a}, a_2=\hat{a}, y_1<y_2].$$

- Pairwise DP and pairwise EO can be interpreted as pairwise analogues of standard DP (requiring $\mathbb{P}[f(x)=1|a=\tilde{a}] = \mathbb{P}[f(x)=1|a=\hat{a}]$) and standard EO (requiring $\mathbb{P}[f(x)=1|a=\tilde{a},y=1] = \mathbb{P}[f(x)=1|a=\hat{a},y=1]$), respectively.
- Any constant predictor $f(x)=i$ satisfies both pairwise DP and pairwise EO; perfect predictor $f(x)=y$ satisfies pairwise EO.

We measure violation of pairwise DP / EO by DP-viol / EO-viol:

$$\text{DP-viol} = \text{DP-viol}(f;\mathbb{P}) = \max_{\tilde{a},\hat{a}\in\mathcal{A}} \big|\mathbb{P}[f(x_1)>f(x_2)|a_1=\tilde{a},a_2=\hat{a}] - \mathbb{P}[f(x_1)<f(x_2)|a_1=\tilde{a},a_2=\hat{a}]\big|$$

and EO-viol is defined analogously; on dataset $\mathcal{D}$, we can evaluate DP-viol$(f;\mathcal{D})$ in time $\mathcal{O}(|\mathcal{D}|+k|\mathcal{A}|^2)$ and EO-viol$(f;\mathcal{D})$ in time $\mathcal{O}(|\mathcal{D}|+k^2|\mathcal{A}|^2)$.

Our approach to learning a fair predictor

We propose a two-step approach to learn a fair threshold model:
1) We learn a fair scoring via a reduction to fair binary classification.
2) We choose fair thresholds via local search.

1) Learning the scoring function:

We learn a fair linear scoring function by adapting the classical approach of Herbrich et al. (2000):

- scoring function s should satisfy $s(x_1)<s(x_2) \Leftrightarrow y_1<y_2$; if $s(x)=w\cdot x$, then $s(x_1)<s(x_2)\Leftrightarrow \text{sgn}(w\cdot(x_1-x_2))=-1$
- given training data $\mathcal{D}=((x_i,y_i))_{i=1}^n$, Herbrich et al. learn w by solving binary classification problem on $\mathcal{D}'=\{(x',y')=(x_i-x_j,\text{sgn}(y_i-y_j)) : i,j\in[n], y_i\neq y_j\}$
- we adapt their approach by learning a linear classifier on $\mathcal{D}'$ that (approximately) satisfies a fairness constraint closely related to standard DP or EO

Proposition 1 (Reduction to fair binary classification—informal)**.** *Let* $\mathcal{D}=((x_i,y_i,a_i))_{i=1}^n\subseteq\mathbb{R}^d\times[k]\times\mathcal{A}$ *and* $\mathcal{D}'=\{(x',y',a')=(x_i-x_j,\text{sgn}(y_i-y_j),(a_i,a_j)) : i,j\in[n],y_i\neq y_j\}\subseteq\mathbb{R}^d\times\{-1,1\}\times\mathcal{A}^2$. *For* $w\in\mathbb{R}^d$, *let* c_w *be the binary classifier* $c_w(x)=\text{sgn}(w\cdot x)$ *and* s_w *be the scoring function* $s_w(x)=w\cdot x$. *We have*

$$c_w \text{ satisfies variant of standard DP (EO) on } \mathcal{D}' \;\Leftrightarrow\; s_w \text{ satisfies pairwise DP (EO) on } \mathcal{D}.$$

2) Learning the thresholds:

Let $f(\cdot\,;s,\boldsymbol{\theta})$ denote the threshold model-predictor with scoring function s and thresholds $\boldsymbol{\theta}=(\theta_1,\dots,\theta_{k-1})$; on training data $\mathcal{D}=((x_i,y_i,a_i))_{i=1}^n$, we aim to solve

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^n C_{y_i,f(x_i;s,\boldsymbol{\theta})} + \lambda\cdot\text{Fair-viol}(f(\cdot\,;s,\boldsymbol{\theta});\mathcal{D}). \tag{1}$$

Problem (1) can be solved in polynomial time using dynamic programming, but running time is prohibitively high in practice ⇒ we perform a local search, moving one threshold θ_i at a time, to find a local minimum of (1); can be implemented with running time $\mathcal{O}(n|\mathcal{A}|^2)$ (pairwise DP) or $\mathcal{O}(nk|\mathcal{A}|^2)$ (pairwise EO) per iteration.

Do we need to be fair in both steps?

Yes! Intuitively: scoring function more fair ⇒ easier to choose fair thresholds

Lemma 1 (Enforcing fairness in both steps can be necessary—informal)**.**

1. *Choosing most accurate scoring function (not necessarily fair) and subsequently fair thresholds (as accurate as possible) can result in predictor with arbitrary higher cost compared to enforcing fairness in both steps.*
2. *Choosing fair scoring function (as accurate as possible) and subsequently most accurate thresholds (not necessarily fair) can result in most unfair predictor.*

Proof of the lemma uses worst-case examples, but we also observe the phenomenon in our experiments.

Generalization guarantees

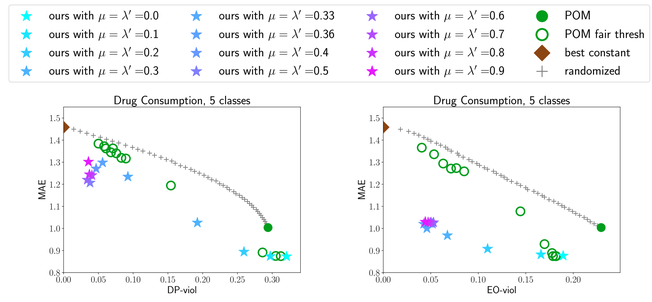
Theorem 1 (Generalization bounds—informal)**.** *Assume the data comes from a distribution* $\mathbb{P}$ *on* $\{x\in\mathbb{R}^d : \|x\|_2\leq R\}\times[k]\times\mathcal{A}$ *with* $\mathbb{P}[a=\tilde{a}]\geq\beta>0$ *for all* $\tilde{a}\in\mathcal{A}$. *Assume we learn* $s(x)=w\cdot x$ *with* $\|w\|_2^2\leq\nu$. *There exists* $M>0$ *such that for any* $\gamma>0$ *and* $0<\delta<1$, *our learned predictor* f *satisfies with probability* $1-\delta$ *over the training sample of size* n,

$$\text{MAE}(f;\mathbb{P}) \leq \widehat{L}^{\gamma}_{\mathcal{D}}(s,\boldsymbol{\theta}) + Mk\sqrt{\frac{1}{n}\left(\frac{R^2\nu}{\gamma^2}\ln(n)+\ln\left(2\frac{(k-1)R\sqrt{\nu}}{\gamma\delta}\right)\right)}$$

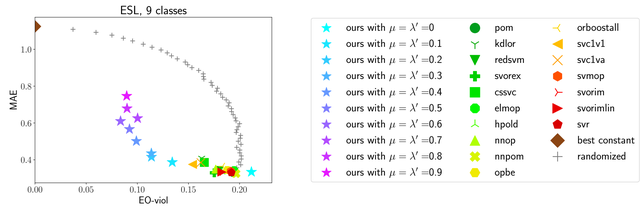
and $\text{DP-viol}(f;\mathbb{P}) \leq \text{DP-viol}(f;\mathcal{D}) + Mk^2\sqrt{\left(d+\log\frac{4|\mathcal{A}|}{\delta}\right)/(n\beta)}$.

$\widehat{L}^{\gamma}_{\mathcal{D}}(s,\boldsymbol{\theta})$ … empirical γ-margin loss; a similar statement holds for pairwise EO.

Experiments



MAE *vs* DP-viol ***(left)*** / EO-viol ***(right)*** *for various predictors — stars: our approach; brown diamond: best constant predictor; green filled circle: POM algorithm (McCullagh, 1980); green circles: second step of our approach applied to POM scoring function (supports Lemma 1, Claim 1); grey crosses: randomly mixing the POM predictor with the best constant one.*



MAE *vs* EO-viol *for our approach and various state-of-the-art methods.*

References

A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.