Fairness guarantee in multi-class classification and regression

Christophe Denis

Joint work with:
E. Chzhen, R. Elie, M. Hebiri,
F. Hu, L. Oneto, and M. Pontil

SAMM, Université Paris1 Panthéon-Sorbonne

16/01/2025

Journée de Statistique Mathématiques, IHP

Multi-class classification: *setting*

Framework

- observation $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{1, \ldots, K\}$
- classifier $f : \mathcal{X} \to \mathcal{Y}$
- misclassification risk $R(f) = \mathbb{P}(f(X) \neq Y)$

Optimal rule

- conditional probabilities $p_k(X) = \mathbb{P}(Y = k \big| X)$
- Bayes classifier $f^*(\cdot) \in \arg\max_{k \in \mathcal{Y}} p_k(\cdot)$
- oracle risk $R^* = R(f^*) = \min_f R(f)$

Goal

- learning sample $(X_i, Y_i)_{1 \leq i \leq n}$ and new observation X_{n+1}
- empirical classification rule $\widehat{f}$ based on the learning sample
- $\hat{f}(X_{n+1})$ prediction of the associated label

Multi-class classification: *plug-in approach*

Plug-in rule

- build $\hat{p}_k$ estimators of p_k
- consider $\hat{f}(\cdot) \in \arg\max_{k\in\mathcal{Y}} \hat{p}_k(\cdot)$

Excess risk

- one can show that

$$\mathbb{E}\left[R(\hat{f})\right] - R^* \leq \sum_{k=1}^{K} \mathbb{E}\left[|\hat{p}_k(X) - p_k(X)|\right]$$

- consistency of $\hat{p}_k \Rightarrow$ consistency of $\hat{f}$
  $\hookrightarrow \ \mathbb{E}\left[R(\hat{f})\right] \to R^*$

Multi-class classification through awareness under DP constraint

Fairness in multi-class classification: *group fairness approach*

Framework

- obervation (X, S) and $Y \in \mathcal{Y}$,
- $S \in \{-1, 1\}$ sensitive attribute
- Fairness through awareness: $f \rightarrow$ prediction $f(X, S)$

Definition of fairness

- Demographic parity (DP), for each $k \in \mathcal{Y}$

$$\mathbb{P}\left(f(X,S) = k | S = 1\right) = \mathbb{P}\left(f(X,S) = k | S = -1\right)$$

- Equalized odds, for each $k \in \mathcal{Y}$

$$\mathbb{P}\left(f(X,S) = k | S = 1, Y = k\right) = \mathbb{P}\left(f(X,S) = k | S = -1, Y = k\right)$$

Main approaches to enforce fairness in classification

Pre-processing

- find a feature representation $z \mapsto \phi(z)$
- such that $\phi(Z)$ independent on S
- adversarial methods Zhang *et al* (2018), Tavker *et al* (2020)

In-processing

- given a set of predictor $\mathcal{F}$, solve

$$f \in \arg\min_{f \in \mathcal{F}} \hat{R}(f) + \lambda \hat{C}(f),$$

  with $\hat{R}(f)$ empirical risk, $\hat{C}(f)$ empirical fairness constraints
- E.R.M. with convex loss Donini *et al* (2018), Ye and Xie (2020)
- E.R.M. with randomized classifiers Agarwal *et al* (2018)

Post-processing

- given a pre-built predictor f, not necessary fair
- find $\hat{T}$ *s.t.* $\hat{T}(f)$ satisfies a desired fairness constraint
- based on optimal transport Chiapa *et al* (2020), Xian *et al* (2023)

Multi-class classification under DP constraint

Notations

- $\mathcal{S} = \{-1, 1\}$, and $\mathcal{Y} = \{1, \ldots, K\}$
- $\pi_s = \mathbb{P}(S = s) > 0$, and $p_k(X, S) = \mathbb{P}(Y = k | X, S)$
- classifier $f \to$ prediction $f(X, S) \in \mathcal{Y}$

Problem

- DP constraint, for each $k \in \mathcal{Y}$

$$\sum_{s \in \mathcal{S}} s\mathbb{P}\left(f(X, S) = k | S = s\right) = 0$$

- $f^* \in \arg\min_f \{\mathbb{P}(f(X, S) \neq Y), f \text{ satisfies DP}\}$
- lagrangian associated to the minimization problem

$$\mathcal{R}_\lambda(f) = \mathbb{P}\left(f(X, S) \neq Y\right) + \sum_{k=1}^{K} \lambda_k \sum_{s \in \mathcal{S}} s\mathbb{P}(f(X, S) = k | S = s)$$

Optimal fair classifier

Continuity assumption

- $t \mapsto \mathbb{P}(p_k(X,S) - p_j(X,S) \leq t | S = s)$ is continuous

Optimal predictor

- the optimal fair classifier f^* can be characterized as

$$f^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s} \lambda_k^* \right)$$

- λ_k^* are lagrange multiplier defined as

$$\lambda^* \in \arg\min_{\lambda \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max_k (\pi_s p_k(X,s) - s\lambda_k) \right]$$

Proposition

Under the continuity assumption, we have

$$f^* \in \arg\min_f \mathcal{R}_{\lambda^*}(f)$$

Optimal predictor: *sketch of the proof (1/2)*

- for each $\lambda \in \mathbb{R}^K$, consider the Lagrangian $\mathcal{R}_\lambda(f)$ defined as

$$\mathbb{P}(f(X,S) \neq Y) + \sum_{k=1}^{K} \lambda_k \sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s}\left(f(X,S) = k\right)$$

- we have that $\mathcal{R}_\lambda(f)$ can be expressed as

$$1 - \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\left(\pi_s p_k(X,S) - s\lambda_k\right) \mathbb{1}_{\{f(X,S)=k\}}\right]$$

- we deduce that $f_\lambda^* \in \arg\min_f \mathcal{R}_\lambda(f)$ is characterized as

$$f_\lambda^*(x,s) = \arg\max_{k \in \{1,\dots,K\}} \left(p_k(X,S) - \frac{s\lambda_k}{\pi_s}\right),$$

and

$$\mathcal{R}_\lambda(f_\lambda^*) = 1 - \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max\left(\pi_s p_k(X,S) - s\lambda_k\right)\right]$$

Optimal predictor: *sketch of the proof (2/2)*

- consider $\lambda^* \arg\min_\lambda H(\lambda)$ with

$$H(\lambda) = \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max\left(\pi_s p_k(X, S) - s\lambda_k\right)\right]$$

- observe that $\lambda^* \in \arg\max_\lambda \mathcal{R}_\lambda\left(f^*_\lambda\right)$
- under continuity assumption, $\lambda \mapsto H(\lambda)$ is differentiable and the first order condition shows that $f^*_{\lambda^*}$ satisfies DP
- therefore, with the weak duality, we obtain that $f^* = f^*_{\lambda^*}$

Data driven procedure: *post-processing approach*

Objective

- estimate $f^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s}\lambda_k^* \right)$

Plug-in approach

- labeled sample $\to$ estimate p_k
- unlabeled sample $(X_1, S_1), \ldots, (X_N, S_N)$
- $\{S_1, \ldots, S_N\}$ $\to$ estimate π_s by their empirical frequencies
- $\{X_1, \ldots, X_N\}$ $\to$ estimate parameter λ_k^*

Randomization

- fairness guarantee requires continuity assumption
- introduce $\zeta \sim \mathcal{U}_{[0,u]}$ independent of (X,S), $u \to 0$
- $\bar{p}_k(X,S,\zeta) = \hat{p}_k(X,S) + \zeta$

Post-processing estimator: *randomized classifier*

Randomized fair classifier

- $(X_1, \ldots, X_N) \to (X_1^s, \ldots, X_{N_s}^s)$ i.i.d. from $X|S = s$
- estimator $\hat{\lambda}$

$$\hat{\lambda} \in \arg\min_{\lambda \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \frac{1}{N_s} \sum_{i=1}^{N_s} \max_k \left(\hat{\pi}_s \bar{p}_k(X_i^s, s, \zeta_{k,i}^s) - s\lambda_k \right)$$

- resulting classifier

$$\hat{f}(x, s) \in \arg\max_{k \in \mathcal{Y}} \left(\bar{p}_k(x, s, \zeta_k) - \frac{s}{\hat{\pi}_s} \hat{\lambda}_k \right)$$

Theoretical guarantees: *fairness guarantee*

Unfairness measure

$$\mathcal{U}(f) = \max_k \left|\mathbb{P}\left(f(X,S) = k | S = 1\right) - \mathbb{P}\left(f(X,S) = k | S = -1\right)\right|$$

Distribution free-result

There exists C depending only on K and π_s such that for any estimator $\hat{p}_k$

$$\mathbb{E}\left[\mathcal{U}(\hat{f})\right] \leq C N^{-1/2}$$

Theoretical guarantees: *consistency*

Measure of performance

- $f^* \in \arg\min_f \mathcal{R}_{\lambda^*}(f)$

$$\mathcal{R}_{\lambda^*}(f) = \mathbb{P}\left(f(X,S) \neq Y\right) + \sum_{k=1}^{K} \lambda_k^* \sum_{s \in \mathcal{S}} s\mathbb{P}(f(X,S) = k | S = s)$$

- $\|\hat{\mathbf{p}} - \mathbf{p}\|_1 = \sum_{k=1}^{K} |\hat{p}_k(X,S) - p_k(X,S)|$

Theorem

Under continuity assumption

$$\mathbb{E}\left[\mathcal{R}_{\lambda^*}(\hat{f}) - \mathcal{R}_{\lambda^*}(f^*)\right] \lesssim \mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right] + u + N^{-1/2}$$

- assume that $\hat{p}_k$ are consistent and $u \to 0$
  $\hookrightarrow$ $\hat{f}$ is consistent

Extension to ε-fairness (1/2)

Approximate fairness: ε-DP

- f is ε-fair iff $\mathcal{U}(f) \leq \varepsilon$

Optimal ε-fair classifier

- $f_\varepsilon^* \in \arg\min_f \{\mathbb{P}(f(X,S) \neq Y), f$ satisfies $\varepsilon -$ DP$\}$
- $\left(\lambda^{*(1)}, \lambda^{*(2)}\right)$ minimizer of

$$\sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_k \left(\pi_s p_k(X,s) - s\left(\lambda_k^{(1)} - \lambda_k^{(2)}\right)\right)\right] + \varepsilon \sum_{k=1}^{K} \left(\lambda_k^{(1)} + \lambda_k^{(2)}\right)$$

- $f_\varepsilon^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s}\left(\lambda_k^{*(1)} - \lambda_k^{*(2)}\right)\right)$

Extension to ε-fairness (2/2)

Optimal ε fair predictor: properties

- $\lambda_k^{*(1)}\lambda_k^{*(2)} = 0$ and $\lambda_k^{*(1)} + \lambda_k^{*(2)} \geq 0$, $k \in [K]$
- if $\mathcal{U}\left(f^*_{\text{Bayes}}\right) \leq \varepsilon$ then $f^*_\varepsilon = f^*_{\text{Bayes}}$, and $\lambda^{*(1)} = \lambda^{*(2)} = 0$
- else $\mathcal{U}\left(f^*_\varepsilon\right) = \varepsilon$

Estimation

- same procedure as for exact fairness
- $\mathbb{E}\left[\mathcal{U}(\hat{f}_\varepsilon)\right] \leq \varepsilon + CN^{-1/2}$
- fairness and risk guarantees

ε-fairness: results

Unfairness

Let $N_{\min} = \min(N_{-1}, N_1)$, under mild assumptions with probability larger than 1-δ, we have that $\hat{\lambda}_k^{(1)}\hat{\lambda}_k^{(2)} = 0$ and either

$$\left|\mathcal{U}(\hat{f}_\varepsilon) - \varepsilon\right| \leq C_0 \frac{\log(1/\delta)}{\sqrt{N_{\min}}},$$

or

$$\mathcal{U}(\hat{f}_\varepsilon) < \varepsilon - C_0 \frac{\log(1/\delta)}{\sqrt{N_{\min}}}, \quad \text{and} \quad \hat{\lambda}^{(1)} = \hat{\lambda}^{(2)} = 0$$

Fast rates

- if $p_k(X,S) - p_j(X,S)$ admits a bounded density

$$\mathbb{E}\left[\mathcal{R}_{\lambda^*}(\hat{f}) - \mathcal{R}_{\lambda^*}(f^*)\right] \lesssim \mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_\infty^{2}\right] + u + N^{-1/2}$$

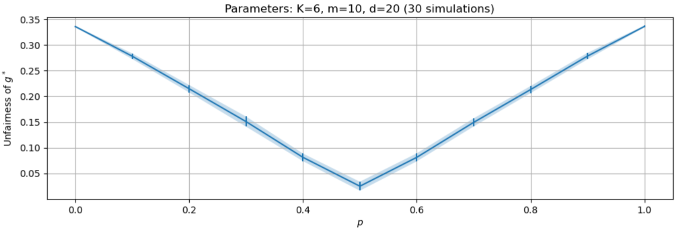
Numerical illustration: *model*

Synthetic data: *Gaussian mixture*

- let $c^k \sim \mathcal{U}_d(-1,1)$, and $\mu_1^k, \ldots, \mu_m^k \sim \mathcal{N}_d(0, I_d)$
- covariates: $(X|Y=k) \sim \frac{1}{m}\sum_{i=1}^{m} \mathcal{N}_d(c^k + \mu_i^k, I_d)$
- sensitive feature:

$$\begin{aligned}(S|Y=k) &\sim 2\cdot\mathcal{B}(p) - 1, k \le \lfloor K/2 \rfloor \\ (S|Y=k) &\sim 2\cdot\mathcal{B}(1-p) - 1, k > \lfloor K/2 \rfloor\end{aligned}$$

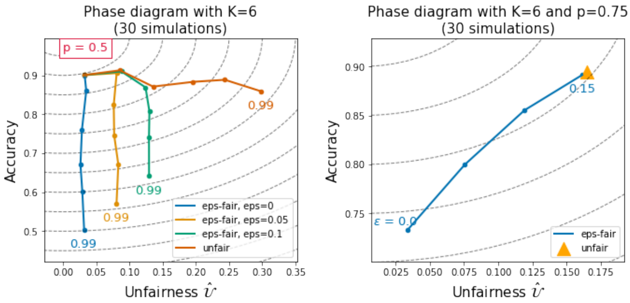
- fair data $p = 0.5$ / unfair data $p \in \{0, 1\}$



Numerical illustration: *results*

Scheme

- generate 5000 examples
- train/test/unlabeled $= 60\%/20\%/20\%$
- estimate p_k on *train* dataset using random forests
- build $\hat{f}$ using *unlabeled* dataset
- evaluated $\mathrm{Acc}(\hat{f})$ and $\mathcal{U}(\hat{f})$ using *test* dataset



Regression through awareness under DP constraint

Regression under DP constraint

Regression framework

- observation (X, S, Y), $Y \in \mathbb{R}$
- $Y = \eta(X, S) + \varepsilon$ with $\mathbb{E}\left[\varepsilon | X, S\right] = 0$
- prediction rule: $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$
- L_2 risk $R(f) = \mathbb{E}\left[(Y - f((X, S))^2\right]$
- optimal rule $\mathbb{E}\left[Y | X, S\right] = \eta(X, S)$

DP constraint

- *exact* DP constraint

$$\sup_{t \in \mathbb{R}} \left|\mathbb{P}(f(X, S) \leq t | S = 1) - \mathbb{P}(f(X, S) \leq t | S = -1)\right| = 0$$

- *optimal fair* predictor f^* defined as

$$f^* \in \arg\min_{f} \{R(f), f \ \text{ satisfies } \ \text{DP}\}$$

Related works

Regression under DP constraint

- two main approaches
- approach that relies on optimal transport
  Chzhen and Schreuder, (2020), Chzhen *et al.* (2020), Le Gouic *et al.* (2020)
- approach that relies on discretization
  Agarwall (2019), Chzhen *et al.* (2020), Chzhen *et al.* (2024)

Discretization (Chzhen *et al.* (2020))

Discretization

- assume that $|Y| \leq 1$
- consider a grid $\mathcal{G}_L = \{\frac{l}{L},\ l = -L, \ldots, L\},\ L > 0$
- discretized predictor $f_L(x, s) \in \mathcal{G}_L$

DP constraint for discretized predictor

- f_L^* satisfies DP *iff*

$$\max_{l \in \{-L, \ldots, L\}} \sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s}\left(f_L(X, S) = \frac{l}{L}\right) = 0$$

- $f_L^* \arg\min_{f_L} \{R(f_L),\ \ f_L \ \text{ satisfies } \ \text{DP}\}$
- proposal : estimate f_L^* rather than f^*

Approximation property

$$R(f^*) \leq R(f_L^*) \leq R(f^*) + 2\frac{\sqrt{\mathrm{Var}(Y)}}{L} + \frac{1}{L^2}$$

Optimal discretized fair predictor

Continuity assumption

- $t \mapsto \mathbb{P}(\eta(X,s) \leq t | S = s)$ is continuous

Optimal predictor

- f_L^* can be characterized as

$$f_L^*(x,s) \in \arg\min_l \left(\pi_s \left(\eta(x,s) - \frac{l}{L} \right)^2 - s\lambda_l^* \right) \frac{1}{L},$$

with $\lambda^* = (\lambda^*_{-L}, \dots \lambda^*_L)$

$$\lambda^* \in \arg \min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \max_l \left(s\lambda - \pi_s \left(\eta(x,s) - \frac{l}{L} \right) \right)$$

Estimation

- similar to the post-processing procedure in classification

Data driven procedure

Plug-in approach

- similar to the post-processing procedure in classification
- estimate η (*with randomization*) $\to \hat{\eta}$
- estimate $\pi_s \to \hat{\pi}_s$ and $\lambda \to \hat{\lambda}$

$$\hat{f}_L \in \arg\min_{l} \left(\hat{\pi}_s \left(\hat{\eta}(x,s) - \frac{l}{L} \right)^2 - s\hat{\lambda}_l \right) \frac{1}{L}$$

Properties

- $\mathbb{E}\left[\mathcal{U}(\hat{f}_L)\right] \leq C\sqrt{\frac{L}{N}}$
- $L = N^{-1/4}$, and $\mathbb{E}\left[R(\hat{\eta}) - R(\eta)\right] \to 0$, then

$$\mathbb{E}\left[R(\hat{f}_L)\right] \to R(f^*)$$

Perspective

DP multiclass classification

- exact and ε-fairness
- plug-in approach
- extension to multiple sensitive attributes
- fairness and risk guarantee

Some extension

- extension to prediction without sensitive attribute
- extension to other fairness measures
- study of optimal rates of convergence