Fair Densities via Boosting the Sufficient Statistics of Exponential Families

Australian National University · HMI · amazon | science · Google Research

Summary

We propose a boosting algorithm for fair pre-processing of datasets with strong theoretical guarantees.

↪ two modes of fairness ↪ use in continuous domain ↪ boosting-style convergence | Others in paper: ↪ characterization of fair densities ↪ update interpretability

Fair Pre-Processing Setting

- $\mathcal{X}$ domain of observations (can be continuous) ;
- $\mathcal{Y}$ domain of labels (finite);
- $\mathcal{S}$ domain of sensitive attributes (finite, separate from $\mathcal{X}$); such as age, race, *etc.*;
- P an input distribution which is unfair.

Goal: ① find a fair density Q; but ② also close to P (*e.g.* KL).

Definitions of Data Fairness

Given a density $\mathrm{P} \in \mathcal{D}(\mathcal{X} \times \mathcal{Y} \times \mathcal{S})$, we consider two forms of data fairness. The following properties hold, if $\forall s, s' \in \mathcal{S}$:

ρ-Representation Rate:
$\mathrm{P}(s)/\mathrm{P}(s') \geq \rho.$
Representation across $\mathcal{S}$.

τ-Statistical Rate:
$\mathrm{P}(y \mid s)/\mathrm{P}(y \mid s') \geq \tau$
Positive outcomes across $\mathcal{S}$.

Denote these rates of a density as $\mathsf{RR}(\mathrm{P})$ and $\mathsf{SR}(\mathrm{P})$, resp.

Fair Boosted Density Estimation

Start with an initially fair (but not accurate) density Q_0.
Let $\mathsf{RR}_0 := \mathsf{RR}(\mathrm{Q}_0)$ and $\mathsf{SR}_0 := \mathsf{SR}(\mathrm{Q}_0)$.

$$\mathrm{Q}_t(x, y, s) \overset{\text{Update}}{\propto} \mathrm{Q}_{t-1}(x, y, s) \cdot \exp(\vartheta_t c_t(x, y, s))$$

Weak Learner $c_t(x, y, s)$:
Classifies the 'realness' of samples of P vs Q_{t-1}.

Leverage Coefficient ϑ_t:
Controls the 'step size' of each update.

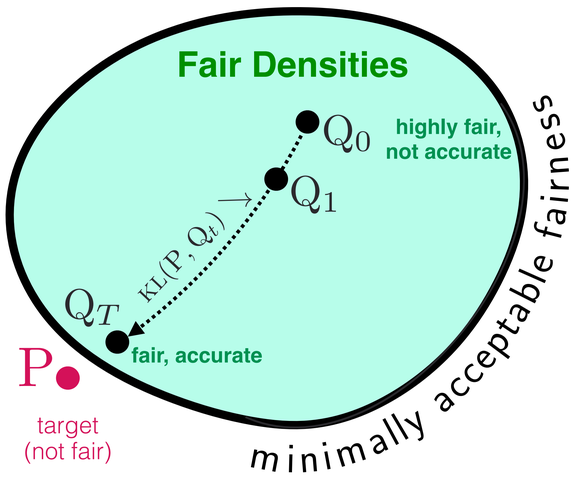


Figure: *Fair Boosted Density Estimation.*
Summary:
① Pick a fair initial distribution;
② Specify a leveraging scheme;
③ Learn weak learners to update distribution.

Fairness Guarantees [Abridged]

Given that Q_0 has high SR_0, we present two modes of fairness by changing the leverage ϑ_t. Let $\tau \in (0, 1]$ be a 'target fairness'. Then,

Exact Fairness:
$\vartheta_t^{\mathrm{E}} := -(C2^{t+1})^{-1} \log(\tau/\mathsf{SR}_0)$

Theorem 3.2:
Suppose $\vartheta_t = \vartheta_t^{\mathrm{E}}$ hold, then
$\mathsf{SR}(\mathrm{Q}_T) > \tau.$

Holds when we forever-boost.

Relative Fairness:
$\vartheta_t^{\mathrm{R}} := -(4Ct)^{-1} \log(\tau/\mathsf{SR}_0)$

Theorem 3.3:
Suppose $\vartheta_t = \vartheta_t^{\mathrm{R}}$ hold, then
$\mathsf{SR}(\mathrm{Q}_T) > \tau \cdot \tau^{\log T}.$

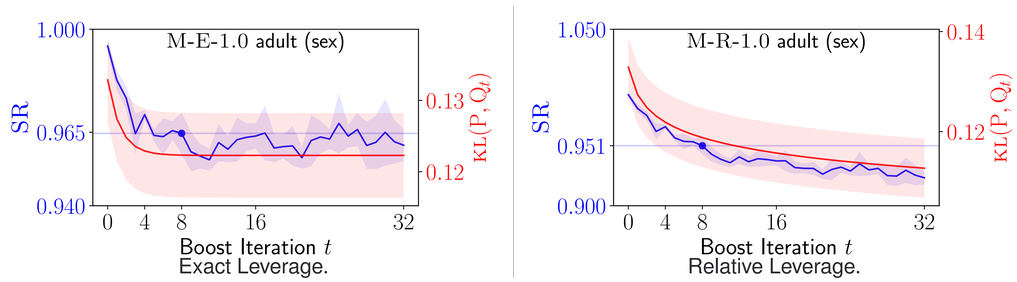
Decays with boosting iterations.

Convergence [Informal]

Theorem 3.5: If the weak learners c_t satisfies a *Weak Learning Assumption* (in predicting samples of P against Q_{t-1}), then

$\mathrm{KL}(\mathrm{P}, \mathrm{Q}_{t-1}) - \mathrm{KL}(\mathrm{P}, \mathrm{Q}_t) \geq \vartheta_t \cdot \Lambda_t, \quad \mathrm{RHS} \geq 0$ with light assumptions.

Using $\mathsf{RR}_0 = 1$, $\mathsf{SR}_0 = 1$, and $\tau = 0.8$:



		Data	M-E-1.0	M-R-1.0	MaxEnt	TabFair
Data	RR	$.496 \pm .002$	$.958 \pm .003$	$.919 \pm .004$	$.995 \pm .002$	$.516 \pm .023$
	SR	$.360 \pm .005$	$.961 \pm .005$	$.924 \pm .006$	$.979 \pm .004$	$.862 \pm .101$
	KL	–	$.122 \pm .006$	$.113 \pm .006$	$.182 \pm .005$	$1.68 \pm .538$
Pred	SR_C	$.360 \pm .003$	$.818 \pm .010$	$.793 \pm .011$	$.919 \pm .011$	$.823 \pm .118$
	EO	$.471 \pm .008$	$.959 \pm .016$	$.935 \pm .016$	$.981 \pm .010$	$.867 \pm .105$
	Acc	$.803 \pm .003$	$.785 \pm .002$	$.787 \pm .002$	$.773 \pm .005$	$.781 \pm .006$

Table: Data and prediction results (Adult Dataset)

MaxEnt: "Data preprocessing to mitigate bias: A maximum entropy based approach. 2020."
TabFair: "TabFairGAN: Fair tabular data generation with generative adversarial networks. 2022."

Alexander Soen[1,2], **Hisham Husain**[2], **Richard Nock**[3,1]
Australian National University[1] Amazon[2] Google Research[3]