On preserving non-discrimination when combining expert advice

Varsha Pendyala
Fall 2018

Abstract

- Interplay between sequential decision making and avoiding discrimination against protected groups when **incoming data is non i.i.d**
- Given a class of predictors, (*each one being non-discriminatory w.r.t to a specified notion of fairness)* they are combined to perform as well as the best predictor while preserving non-discrimination.
- The above task is shown to be unachievable for the notion of **equalized odds** (*equalizing FNR and FPR across all groups*)
- For **equalized error rates**, non-discrimination by composition is achieved by running separate instances of multiplicative weights algorithm for each group.
- It is proved that the algorithms with strong performance guarantees than multiplicative weights cannot preserve non-discrimination.

Sequential setting

- i.i.d assumption is often violated since the future labels can be affected by the recent events, economy etc…
- Hence, modelled as non-i.i.d setting via adversarial online learning
- Multiplicative weights algorithm is used to prove positive as well as impossibility results for attaining non-discrimination by composition of multiple given non-discriminatory predictors
- **Given fair predictors**, it becomes a **trivial task in the batch settings** where unseen examples come from same distribution as labelled data (*it just reduces choosing best predictor since they are already non-discriminatory*)

Online learning protocol with group context

- Classical online learning setting of prediction with expert advice is considered.
- Learner makes sequential decisions for T rounds based on predictions of a finite set $\mathcal{F}$ of d hypothesis
- Outcome space $\mathcal{Y}$ = $\{+, -\}$
- Set of disjoint groups by $\mathcal{G}$
- Each round t is associated with a group context $g(t) \in \mathcal{G}$, and an outcome $y(t) \in \mathcal{Y}$.
- T-length time-group-outcome sequence tuple:
$$\sigma = \left\{\left(t, g(t), y(t)\right) \epsilon\ N\ x\ \mathcal{G}\ x\ \mathcal{Y}\right\}_{t=1}^{T}$$
- $\sigma^{1:\tau} = \left\{\left(t, g(t), y(t)\right) \epsilon\ N\ x\ \mathcal{G}\ x\ \mathcal{Y}\right\}_{t=1}^{\tau}$ is the subsequence until round τ

Protocol for generating time-group-outcome sequence

1. An example with group context $g(t)$ either arrives stochastically or is adversarially selected
2. Learner $\mathcal{L}$ commits to a probability distribution $p^t \in \Delta(d)$ across d experts, p_f^t denotes probability that $\mathcal{L}$ follows advice of expert $f \in \mathcal{F}$ at round t. p^t can be function of $\sigma^{1:t-1}$ (*group-aware learner*)
3. Adversary $\mathcal{A}$ selects an outcome $y(t) \in \mathcal{Y}$. *Adaptive* adversary selects $y(t)$ as a function of $\sigma^{1:t-1}$ otherwise adversary is *oblivious.*

Simultaneously, each expert $f \in \mathcal{F}$ predicts $\hat{y}_f^t \in \hat{\mathcal{Y}}$ where $\hat{\mathcal{Y}}$ is a prediction space.(*labels or probability score or uncalibrated score*)

$l : \hat{\mathcal{Y}}\, x\, \mathcal{Y} \rightarrow [0,1]$ is the loss function that leads to $l^t \in [0,1]^d$ where $l_f^t = l(\hat{y}_f^t, y(t))$ denotes the loss learner incurs if it follows expert f.

4. On observing $y(t)$, the learner incurs expected loss $\sum_{f\in\mathcal{F}} p_f^t l_f^t$

Group fairness in online learning

- $\mathcal{M}$: metric w.r.t which non-discrimination is defined
- For any realization of time-group-outcome sequence σ and any group $g \in \mathcal{G}$, metric $\mathcal{M}$ induces a subset of the population $S_g^\sigma(\mathcal{M})$ relevant to it
- $S_g^\sigma(FNR) = \{t: g(t) = g, y(t) = +\}$
- The performance of expert $f \in \mathcal{F}$ on $S_g^\sigma(\mathcal{M})$ is given by

$$\mathcal{M}_f^\sigma(g) = \frac{1}{|S_g^\sigma(\mathcal{M})|} \sum_{t \in S_g^\sigma(\mathcal{M})} l_f^t$$

Definition 1

An expert $f \in \mathcal{F}$ is called fair in isolation with respect to metric $\mathcal{M}$ if, $\mathcal{M}_f^\sigma(g) = \mathcal{M}_f^\sigma(g')$ for all $g, g' \in \mathcal{G}$

Learner's performance on this subpopulation is

$$\mathcal{M}_\mathcal{L}^\sigma(g) = \frac{1}{|S_g^\sigma(\mathcal{M})|} \sum_{t \in S_g^\sigma(\mathcal{M})} \sum_{f \in \mathcal{F}} p_f^t l_f^t$$

Definition 2

- Consider a set of experts $\mathcal{F}$ such that each one is fair in isolation w.r.t metric $\mathcal{M}$. Learner $\mathcal{L}$ is called α $-$fair in comparison w.r.t metric $\mathcal{M}$ if, for all adversaries that produce $E_\sigma\left[\min\left(\left|S_g^\sigma(\mathcal{M})\right|, \left|S_{g'}^\sigma(\mathcal{M})\right|\right)\right] = \Omega(T)$ for all g, g', it holds that:

$$|E_\sigma[\mathcal{M}_\mathcal{L}^\sigma(g)] - E_\sigma[\mathcal{M}_\mathcal{L}^\sigma(g')]| \leq \alpha$$

Regret notions for online learning setting

- $Reg_T = \sum_{t=1}^{T} \sum_{f\in\mathcal{F}} p_f^t l_f^t - \min_{f^* \in \mathcal{F}} \sum_{t=1}^{T} l_{f^*}^t$
- $ApxReg_{\epsilon,T}(f^*) = \sum_{t=1}^{T} \sum_{f\in\mathcal{F}} p_f^t l_f^t - (1+\epsilon) \sum_{t=1}^{T} l_{f^*}^t$

Goal:

Develop online learning algorithms that combine fair in isolation experts to achieve both vanishing expected ϵ-approximate regret, i.e. for any fixed $\epsilon > 0$ and $f^* \in \mathcal{F}$, $E_\sigma\left[ApxReg_{\epsilon,T}(f^*)\right] = o(T)$, and also non-discrimination w.r.t fairness metrics of interest.

Impossibility results for equalized odds

- Is it possible to achieve vanishing regret property while guaranteeing α-fairness in composition w.r.t FNR for arbitrarily small α?
- This is a trivial task when the input is i.i.d, since best predictor can be learnt in O($\log d$) rounds.
- Impossibility result is proved for non-i.i.d case with the only adversarial component being the order in which examples arrive.

Group-unaware algorithms

Theorem 1:

For all $\alpha < 3/8$, there exists $\epsilon > 0$ such that any group-unaware algorithm that satisfies $E_\sigma\left[ApxReg_{\epsilon,T}(f)\right] = o(T)$ for all $f \in \mathcal{F}$ is α-unfair in composition w.r.t FNR even for perfectly balanced sequences.

In particular, for any group-unaware algorithm that ensures vanishing approximate regret, there exists an oblivious adversary for assigning labels such that:

- In expectation, half of the population corresponds to each group
- For each group, in expectation half of its labels are positive and the other half are negative
- The FNRs of the two groups differ by α

- Consider an instance with two groups $\mathcal{G} = \{A, B\}$, two experts $\mathcal{F} = \{h_n, h_u\}$ and two phases: Phase I and Phase II
- $\hat{y}_f^t \in \hat{\mathcal{Y}} = [0,1]$: probability that expert f assigns to label being positive in round t
- $l(\hat{y}, y) = \hat{y}.\mathbf{1}\{y = -\} + (1 - \hat{y}).\mathbf{1}\{y = +\}$
- $\hat{y}_{h_n}^t = 0, \hat{y}_{h_u}^t = \beta.\mathbf{1}\{y(t) = -\} + (1 - \beta).\mathbf{1}\{y(t) = +\}$ for all t

Phase I: (T/2 rounds)

Adversary assigns negative labels for group B examples

Assigns a label uniformly at random for group A examples

Phase II: (two plausible worlds)

a) If $E_\sigma\left[\sum_{t=1}^{\frac{T}{2}} p_{h_u}^t\right] > \sqrt{\epsilon}.T$, then adversary assigns negative labels for both groups
b) Else, group B examples are assigned positive labels while group A keeps receiving positive and negative labels with equal probability

For any algorithm with vanishing approximate regret property, first world condition is never triggered and hence the above sequence is balanced.

It is shown that this instance is unfair in composition w.r.t FNR by showing following two claims:

1. In Phase I, any ϵ-approximate regret algorithm needs to select negative expert most of the times to ensure small approximate regret w.r.t h_n. Since we encounter half of positive examples from group A and none from group B, FNR of the algorithm is close to 1
2. In Phase II, any ϵ-approximate regret algorithm should quickly catch up to ensure small approximate regret w.r.t h_u and hence FNR of the algorithm is close to β.

This creates a mismatch between the FNR of B (that only receives false negatives in this phase) and A (that has also received many false negatives before)

- **Upper bound on probability of playing $\boldsymbol{h_u}$ in Phase I**

$$E_\sigma\left[\sum_{t=1}^{\frac{T}{2}} p_{h_u}^t\right] < \sqrt{\epsilon}.T$$

- **Upper bound on probability of playing $\boldsymbol{h_n}$ in Phase II**

$$E_\sigma\left[\sum_{t=\frac{T}{2}+1}^{T} p_{h_n}^t\right] < 16\sqrt{\epsilon}.T$$

- **Gap in FNRs between groups A and B**

$$E_\sigma[FNR_{\mathcal{L}}^\sigma(B)] = E_\sigma\left[\frac{\sum_{t=\frac{T}{2}+1}^{T} G_{B,+}^t \,.\left(p_{h_u}^t.\beta + p_{h_n}^t.1\right)}{\sum_{t=\frac{T}{2}+1}^{T} G_{B,+}^t}\right]$$

$$E_\sigma[FNR_{\mathcal{L}}^\sigma(A)] = E_\sigma\left[\frac{\sum_{t=1}^{T} G_{A,+}^t \,.\left(p_{h_u}^t.\beta + p_{h_n}^t.1\right)}{\sum_{t=1}^{T} G_{A,+}^t}\right]$$

Group-aware algorithms

Theorem 2:

For any group imbalance b < 0.49 and $0 < \alpha < \frac{0.49 - 0.99b}{1-b}$ there exists $\epsilon_0 > 0$ such that for all $0 < \epsilon < \epsilon_0$ any algorithm that satisfies $E_\sigma\left[ApxReg_{\epsilon,T}(f)\right] = o(T)$ for all f ∈ $\mathcal{F}$, is α-unfair in composition.

The instance has two groups: $\mathcal{G} = \{A, B\}$. Group A examples arrive randomly with probability $b < 0.49$, while B examples arrive with remaining probability . There are two experts $\mathcal{F} = \left\{h_n, h_p\right\}$. $\hat{y}_{h_n} = 0, \hat{y}_{h_p} = 1$. $c = \frac{1}{101^2}$ percentage of input that is about positive examples for group A to ensure $\left|S_g^\sigma(FNR)\right| = \Omega(T)$. It has two phases.

Phase I ($\mathbf{\Theta.T}$ rounds): (for $\mathbf{\Theta = 101c}$)

Adversary assigns negative labels to group B examples

For group A examples, adversary acts as following:

- If algorithm assigns probability on negative expert below $\gamma(\epsilon) = \frac{99-2\epsilon}{100}$, i.e $p_{h_n}^t(\sigma^{1:t-1}) < \gamma(\epsilon)$, then the adversary assigns negative label.
- Otherwise, the adversary assigns positive labels

Phase II, (two plausible worlds):

a) adversary assigns negative labels to both groups if $E_\sigma\left[\mathbf{1}\left\{t \leq \Theta.T\colon g(t) = A, p_{h_n}^t \geq \gamma(\epsilon)\right\}\right] < c.b.T$

b) Otherwise, it assigns positive labels to examples from group B and negative labels to group A examples

Proof is based on following claims:

1. Any vanishing approximate regret algorithm enters the second world of Phase II
2. This implies the lower bound: FNR(A) $\geq \gamma(\epsilon) = \frac{99-2\epsilon}{100}$
3. In phase II, any ϵ-approximate regret algorithm assigns large enough probability to the positive expert h_p for group B. This implies the upper bound: $FNR(B) \leq \frac{1}{2(1-b)}$. Therefore this leads to a gap in FNRs of atleast α.

Fairness in composition w.r.t equalized error rates

- The relevant subset induced by this metric $S_g^\sigma(EER)$ is the set of all examples coming from group $g \in \mathcal{G}$.
- Running one instance of multiplicative weights for each group achieves fairness in composition
- The results holds for general loss functions (beyond pure classification) and is robust to experts only being approximately fair in isolation
- Algorithms that ignore group identity can be unboundedly fair even with this notion of fairness (*Thus, actively discriminate based on groups to achieve fairness w.r.t EER*)

Positive result

- Separate multiplicative weights instances with a fixed learning rate η, one for each group
- For each pair of expert $f \in \mathcal{F}$ and group context $g \in \mathcal{G}$, initialize $w_{f,g}^{1} = 1$
- At round t, select probability distribution $p_f^t = \frac{w_{f,g(t)}^t}{\sum_{j \in \mathcal{F}} w_{j,g(t)}^t}$.
- Then weights are corresponding to group $g(t)$ are updated exponentially: $w_{f,g}^{t+1} = w_{f,g}^{t}. (1-\eta)^{l_f^t . \mathbf{1}\{g(t)=g\}}$

Theorem 3:

For any $\alpha > 0$ and any $\epsilon < \alpha$ such that running separate instances of multiplicative weights for each group with learning rate $\eta = \min\left(\epsilon, \frac{\alpha}{6}\right)$ guarantees α-fairness in composition and ϵ-approximate regret of at most $O\left(|\mathcal{G}| \log\left(\frac{d}{\epsilon}\right)\right)$

- Since multiplicative weights performs not only no worse than the best expert in hindsight but also no better.
- Thus the average performance of multiplicative weights at each group is approximately equal to the average performance of the best expert in that group.
- Since the experts are fair in isolation, the average performance of the best expert in all groups is the same which guarantees the equalized error rate fairness notion.

Cumulative loss of expert f in examples with group context g:

$$L_{f,g} = \sum_{t:g(t)=g} l_f^t . \mathbf{1}\{g(t) = g\}$$

Expected loss of the algorithm on group g:

$$\hat{L}_g = \sum_{t=1}^{T} \sum_{f \in \mathcal{F}} p_f^t l_f^t . \mathbf{1}\{g(t) = g\}$$

Best in hindsight expert $f^*(g) = argmin_{f \in \mathcal{F}} L_{f,g}$

$$\boldsymbol{\hat{L}_g \ \le (1+\eta). L_{f,g} + \frac{ln(d)}{\eta}} \text{ for all } \boldsymbol{f} \in \boldsymbol{\mathcal{F}}$$

$$\boldsymbol{\hat{L}_g \ \ge (1-4\eta). L_{f^*(g),g}}$$

$$\boldsymbol{\frac{\hat{L}_g}{|\{t: g(t) = g\}|} - \frac{\hat{L}_{g^*}}{|\{t: g(t) = g^*\}|} \le \frac{(1+\eta). L_{f^*(g),g}}{|\{t: g(t) = g\}|} + \frac{ln(d)}{\eta. |\{t: g(t) = g\}|} - \frac{(1-4\eta). L_{f^*(g^*),g^*}}{|\{t: g(t) = g\}|}}$$

$$\boldsymbol{\le 5\eta. \frac{L_{f^*(g^*),g^*}}{|\{t: g(t) = g^*\}|} + \frac{ln(d)}{\eta . T_0}} \text{ (number of examples from each group is at least } T_0\text{)}$$

Impossibility results

Group-unaware algorithms:

- If single multiplicative weights instance is run in a group-unaware way, fairness in composition cannot be achieved
- This impossibility result holds for any algorithm with vanishing ϵ-approximate regret where learning dynamic (p^t at each round t) is a deterministic function of the difference between the cumulative losses of the experts (without taking into consideration their identity)

Theorem 4:

For any $\alpha > 0$ and for any $\epsilon > 0$, running a single algorithm from the above class in a group-unaware way is α-unfair in composition with respect to equalized error rate

Two groups $\mathcal{G} = \{A, B\}$ that come in adversarial order, two experts $\mathcal{F} = \{f_1, f_2\}$ and four phases of equal size.

Phase I {1,…,T/4}

Group A examples arrive and first predictor is correct: $l_{f1}^{t} = 0$ and $l_{f2}^{t} = 1$

Phase II {T/4+1,…,T/2}

Group B examples arrive: $l_{f1}^{t} = 1$ and $l_{f2}^{t} = 0$

Phase III {T/2+1,…,3T/4}

Group A examples arrive: $l_{f1}^{t} = 1$ and $l_{f2}^{t} = 0$

Phase IV{3T/4,…,T}

Group B examples arrive: $l_{f1}^{t} = 0$ and $l_{f2}^{t} = 1$

- Since the loss of first expert is 0 in first quarter of the setting: $\sum_{t=1}^{T/4} l_{f1} = 0$. To achieve vanishing approximate regret

$$\sum_{t=1}^{T/4} p_{f2} < \frac{1-\alpha}{8} T$$

- Hence, error rate on group A is at most $EER(A) \leq \frac{1-\alpha}{2}$ in Phase I, this also applies to Phase III
- For group B, the cumulative probability of the correct expert is upper bounded by $\frac{1-\alpha}{2}$
- Hence, group B incurs an equalized error rate $EER(B) \geq \frac{1+\alpha}{2}$
- Thus, the algorithm is α-unfair in composition