

Individual Fairness and Fair Representations

Farhad Mohsin

3/29/2022

Fairness through awareness

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January).

Individual Fairness - Motivation

- Group fairness (statistical parity) focuses on group level metrics
 - Fair in average
 - $\Pr[M(x) = 1 | x \in S] = \Pr[M(x) = 1 | x \in T]$
- When is statistical parity not enough? (Assume two groups S and T.)
 - Self-fulfilling prophecy
 - When unqualified members of S are chosen to justify future discrimination against S
 - Subset targeting
 - Statistical parity does not guarantee anything about subsets. A mechanism can be fair on a high level, but discriminatory in a more granular level.

Preliminaries

- Random classifier
 - Set of individuals V
 - Set of outcomes A , e.g., $A = \{0,1\}$ for a binary classification
 - $\Delta(A)$ is a distribution over A
 - Classifier, $M: V \mapsto \Delta(A)$
 - $\mu_x = M(x)$ is prediction for $x \in V$ (e.g., 0.8 probability for 0, 0.2 prob for 1)
 - Many learning models predict a probability distribution anyway
- Running example:
 - What advertisement to show to which individual

Preliminaries (contd.)

- Metrics & Distances

- For $x, y \in V$, we assume the distance $d(x, y)$ can be computed

(The existence of such a metric d is a rather big assumption; we'll discuss more later)

- For two distributions, $P, Q \in \Delta(A)$, we can measure similarity using existing measures. E.g., Total variation distance

$$D_{TV}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

	<i>a</i>	<i>b</i>
<i>P</i>	0.8	0.2
<i>Q</i>	0.6	0.4
$ P - Q $	0.2	0.2
$D_{TV}(P, Q)$	$\frac{1}{2}(0.2 + 0.2) = 0.2$	

This allows us to measure difference between two predictions

Fairness through Awareness

- Key purpose
 - Treating similar individuals similarly
 - Similar individuals: $d(x, y)$ is low
 - Similar treatment: $D_{TV}(\mu_x, \mu_y)$ is low (reminder: $\mu_x = M(x)$)
- Example: Regardless of group membership, if two individuals have similar credit scores, they should receive similar advertisements

Fairness through Awareness

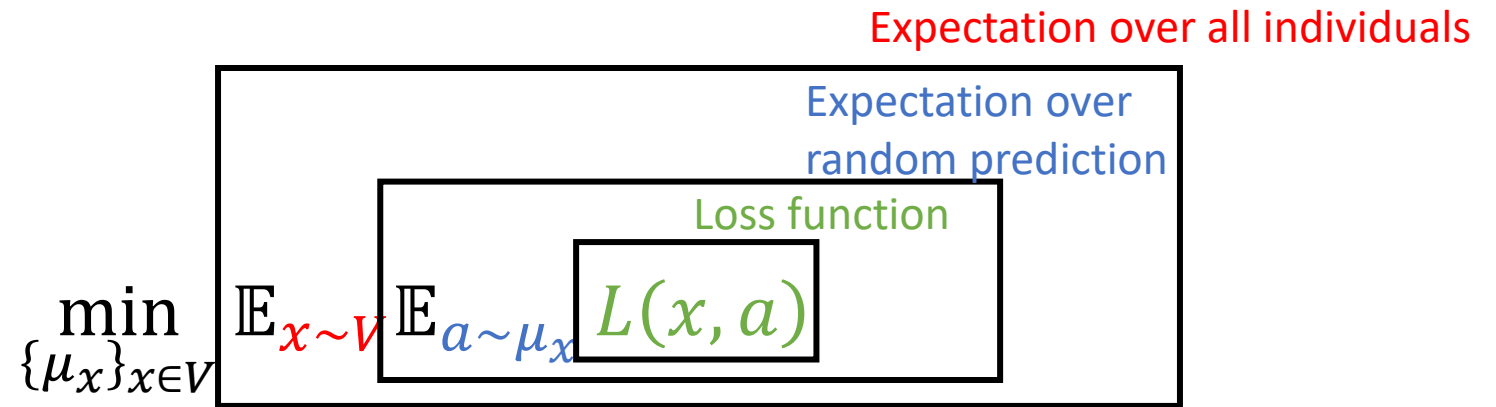
(Formal formulation)

$$\min_{\{\mu_x\}_{x \in V}} \mathbb{E}_{x \sim V} \mathbb{E}_{a \sim \mu_x} L(x, a)$$

Expectation over all individuals

Expectation over random prediction

Loss function



$$\text{subject to } D(\mu_x, \mu_y) \leq d(x, y)$$
$$\mu_x \in \Delta(A)$$

Lipschitz condition

Fig: The fairness LP: Loss minimization subject to fairness constraint

Let's discuss the metric

- How exactly do we compute the similarity between two individuals?
- Ideal: some unbiased oracle knows how to define a useful metric
 - Ensuring fairness is up-to this oracle now
- Compromise
 - Learn a metric
 - Approximations
 - Efficient queries
 - We still need an oracle though

Link to group fairness

- Consider two groups S, T
 - W.l.o.g., assume S, T are distributions over set of individuals
 - (Definition) Bias: A classifier satisfies statistical parity up to bias ϵ if
$$D_{TV}(\mu_S, \mu_T) \leq \epsilon, \text{ where } \mu_S = \sum_{x \in S} S(x) \mu_x$$
 - (Theorem) Bias is limited by Earthmover's distance* between S, T .
$$\epsilon \leq d_{EM}(S, T)$$
- Key point: When S, T are similar, individual fairness ensures group fairness

*Earthmover's distance is yet another way to measure distances of distributions

Fair Affirmative Action

aka “What if we want Statistical Parity more than Individual Fairness?”

- Compute a mapping $\nu: S \mapsto \Delta(T)$ to represent elements in S with distributions in T that maintains Lipschitz condition
- Now, optimize for joint loss function for $y \in T$:

$$L'(y, a) = \underbrace{\sum_{x \in S} \nu_x(y) L(x, a)}_{\text{Indirect loss}} + \underbrace{L(y, a)}_{\text{Direct Loss}}$$

- Affirmative action towards group S
- Maintains individual fairness within groups
- Guarantees ‘best outcome’ to the ‘best people’ in group S
 - Opposes self-fulfilling policy

Learning fair representations.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May).

Fair Representations

- Philosophically follows the previous paper
 - Task: Learn fair binary classifier
 - Goal: Learn a fair representation on data and then train model on fair representation
- Before we start:
 - What is a **fair representation**
 - Represents non-protected features
 - Obfuscates protected feature

Why is this important?

- A representation automatically gives a metric
- (As opposed to the previous paper) This is actually a learning problem
 - So we can generalize
- Can consider both group and individual fairness

Defining a Representation

- Let X be set of all data $X \subseteq \mathbb{R}^d$
 - $X = X^+ \cup X^-$, X^+ is the protected set
- Label $Y \in \{0,1\}$
- K prototypes $v_k \in X$ to represent all $x \in X$, using random multinomial variable Z
 - $x = \sum_{k \leq K} \Pr(Z = k|x) v_k$
- For a fair representation
 - $\Pr(Z = k|x \in X^+) = \Pr(Z = k|x \in X^-)$

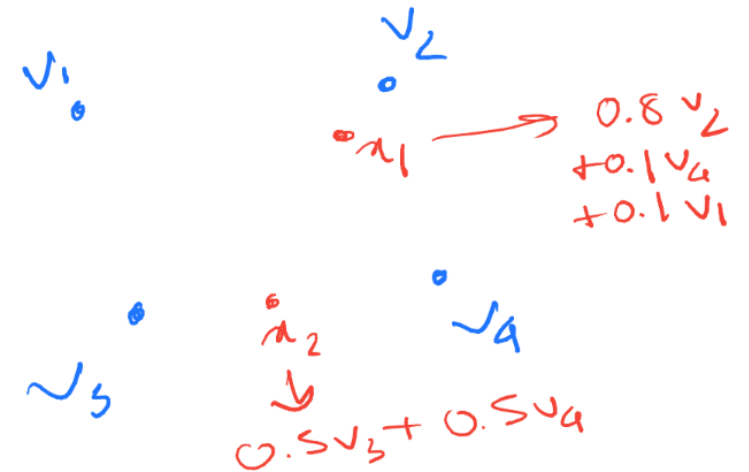


Fig: Example representation in \mathbb{R}^2

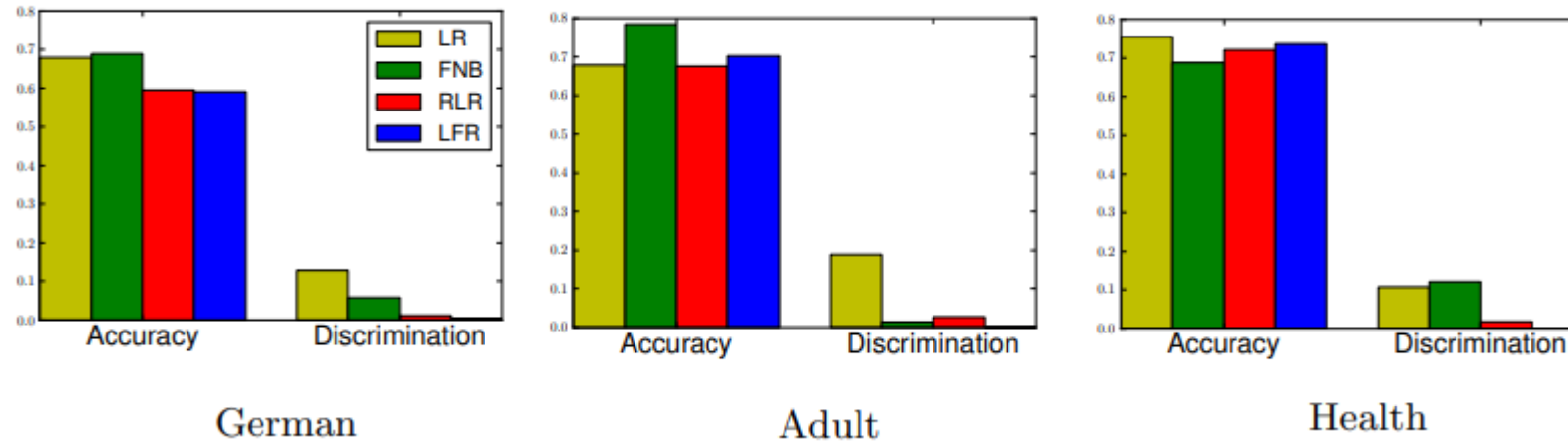
Learning Fair Representation

- Joint Loss function

$$L = A_z \overset{\text{representation}}{L_z} + A_x \overset{\text{fairness}}{L_x} + A_y \overset{\text{prediction}}{L_y}$$

- $L_z = \sum_k |\mathbb{E}_{X^+} \Pr(Z = k|x) - \mathbb{E}_{X^-} \Pr(Z = k|x)|$
 - $L_x = \sum_n (x_n - \widehat{x}_n)^2$, the reconstruction error
 - L_y is some predictive loss, e.g., cross-entropy loss
-
- Prediction
 - $\widehat{y}_n = \sum_k \Pr(Z = k|v_k) w_k$
 - What we're learning:
 - Prototypes $\{v_k\}$, weights $\{w_k\}$

Results



Baseline
LR: Logistic Regression

Fair Baselines:
FNB: Fair Naïve Bayes
RLR: Regularized Logistic Regression

LFR: Learned Fair Representation

Fig: Accuracy and Discrimination (group unfairness measure) for three tasks

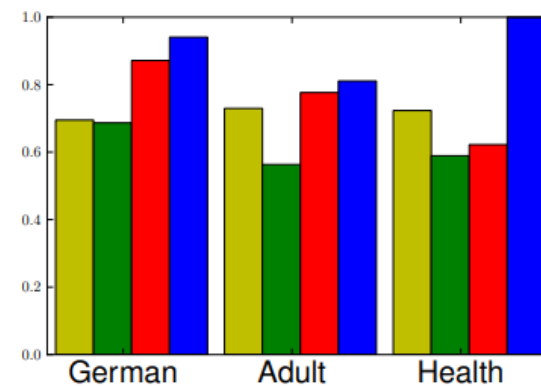


Fig: Individual Fairness for three tasks

Topics of interest

- Defining/learning similarity metric
 - Biggest hurdle against wide adoption of individual fairness notion
 - E.g., Kim et.al. 2018, *Fairness Through Computationally-Bounded Awareness*, Ilvento 2020, *Metric Learning for Individual Fairness*
- Relation with Differential Privacy
 - Applying fairness through awareness is similar to exponential mechanism
 - Study compatibility of fairness and privacy
- Learning fair representations for more complex models