Fairness in Machine Learning : detection and correction using Optimal Transport Theory

J-M. Loubes

joint work with E. del Barrio, P. Besse, P. Gordaliza, F. Gamboa

Conference of European Statistics Stakeholders 2018

Institut Mathématiques de Toulouse
University of Toulouse

Fairness & Machine Learning

Big Data and AI

The Data convey the information and the model should be built to fit the data

- From data to information : extraction the knowledge from empirical observations
- Finding relationships or links between variables and a target
- The rule can be generalized to forecast new observations

The more the data the better the description of the reality

Principle of Machine Learning : from a set of labeled examples (learning sample), build a decision rule that fits the data that will be used for all the population which has the same distribution as the learning sample.

The Algorithm (or **AI**) learn the best rule from the data and then can forecast for new observations with a guaranteed precision.

How Machine Learning could go possibly wrong ?

Decisions are taken based on machine learning algorithms in many fields (recommendation systems, insurance, banks, human ressources, education, communication ...) and are on the way to be used in many sensitive areas (justice, medicine, police, political decisions ....)
Algorithms are more and more complex producing highly non linear outputs with non interpretable rules.

- A classifier goal is to cluster observation, to discriminate. If there is seed of discrimination it will be increased, **enforced** by the decision rule.
- AI **generalizes** the situation encountered in the learning sample to the whole population.
  It shapes the reality according to the learnt rule without questioning nor evolution.
- "*It's the mathematics, stupid*" . Difficult to argue with an expert AI to understand a decision. Mathematics can not be questioned so the decisions taken can be **justified** using a scientific argument.

What does it mean Being Fair ?

- Fair : "*treating someone in a way that is right or reasonable, or treating a group of people equally and not allowing personal opinions to influence your judgment*" (Cambridge Dictionnary)
- Every group should be treated without prejudice and the decisions should be based on **variables that makes sense**.

  Examples :
  Hiring policy driven by people's study or sex ?
  Granting a loan using religion or earnings ?
  Product Recommendations using Internet History or street address?
  Recidivism score based on judicial past of offendant or colour of skin?
  Some variables (**protected variables**) should not be used alone to build a model.
- The **efficiency** of the algorithm should not depend on these variables and should be the same for all groups.
  Examples : new treatment in medicine should be efficient for all population.

What does it mean Being Fair ?

- Fair : "*treating someone in a way that is right or reasonable, or treating a group of people equally and not allowing personal opinions to influence your judgment*" (Cambridge Dictionnary)
- Every group should be treated without prejudice and the decisions should be based on **variables that makes sense**.

  Examples :
  Hiring policy driven by people's study or ~~sex~~ ?
  Granting a loan using ~~religion~~ or earnings ?
  Product Recommendations using Internet History or ~~street address~~?
  Recidivism score based on judicial past of offendant or ~~colour of skin~~?
  Some variables (**protected variables**) should not be used alone to build a model.
- The **efficiency** of the algorithm should not depend on these variables and should be the same for all groups.
  Examples : new treatment in medicine should be efficient for all population.

What does it mean Being Fair ?

Consider the probability space $\left(\Omega \subset \mathbf{R}^d, \mathcal{B}, \shortmid\right)$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbf{R}^d$ and $d \geqslant 1$. In this space,

- $Y : \Omega \to \{0, 1\}$ **target class**

$$Y = \begin{cases} 0 & \textit{failure} \\ 1 & \textit{success} \end{cases}$$

- $X : \Omega \to \mathbf{R}^d,\ d \geqslant 1$, **visible attributes**
- $S : \Omega \to \{0, 1\}$ **protected attribute**

$$S = \begin{cases} 0 & \textit{unfavored} \\ 1 & \textit{favored} \end{cases}$$

- $\mathcal{G}$ family of **binary classifiers** $g : \mathbf{R}^d \to \{0, 1\}$
- $\hat{Y} = g(X),\ g \in \mathcal{G}$ **outcome of the classification**

Fairness deals with the relationships between Y, $\hat{Y}$ and S.

Examples

1. **Adult Income Data**.
   Data from a bank : Forecast from characteristics if someone has the potential to have a high income ($\geq 50k\$$) to grant a loan.
   Variables : Age, Workclass, Final weight, Education, Marital status, Occupation, Relationship, **Gender**, Race, Capital gain, Capital loss, Hours per week, Native country.
   Output $Y \in \{0, 1\}$ if predicted income is higher than the threshold or not.
   Protected Variables : Gender, Race, Native country.
   Result :
   $$\mathbb{P}(\hat{Y} = 1 | S = 1) >> \mathbb{P}(\hat{Y} = 1 | S = 0).$$

Examples

1. **Adult Income Data**.
2. **ProPublica vs Northpoint**
   Northpoint produces a score COMPAS to measure the probability of recidivism of offendants. This score has been designed using Machine Learning Algorithm from a learning sample to predict if someone will commit a crime when set free $Y = 0$.
   Variables : characteristics of people and their crime
   Protected Variable : Ethnic Origin $S = 0$ coding Afro-American.
   It is balanced
   $$\mathbb{P}(\hat{Y} = 1|S = 1) \sim \mathbb{P}(\hat{Y} = 1|S = 0).$$
   But the errors are different
   $$\mathbb{P}(\hat{Y} = 1|S = 1, Y = 0) >> \mathbb{P}(\hat{Y} = 1|S = 0, Y = 0).$$

Removing Unfair treatment : a Challenge

Two different points of view in the Machine Learning literature.

- The forecast should **not depend** on the protected variables S
  $\hat{Y}$ should be independent of S or their *correlation strength* should not be too strong.

  Several Criteria to measure the strength of the relationship between the two variables : **Disparate Impact Assessment** or **Distance between the distributions**
  $$d(\mathcal{L}(\hat{Y}|S=0), \mathcal{L}(\hat{Y}|S=1)) \leq \varepsilon.$$

  Or the **prediction errors** should not be different for the two groups $S=0$ and $S=1$.
- Given the forecast $\widehat{Y}$ and the variables X, S **should not be predictable**.

Different notions to be considered called Statistical Parity, Conditional Equity, Conditional Use Equity, Balance Error Rate ...

Removing Unfair treatment : a Challenge

- **Machine Learning Algorithm produces discrimination**.
  Automatic classifiers aims at discriminating the population so all discriminative trends are investigated.
  Removing the variable S does not help since X and S can be strongly correlated.
  => **Penalizing the algorithm to prevent dependency w.r.t** S
- **Bias in the learning sample.**
  Data is not representative of all population and there is **different treatment** between $S = 0$ and $S = 1$ groups

$$\hat{Y} = g(X, S), \quad R(g, X, S) := \mathbb{P}(g(X, S) \neq Y).$$

=> **Repairing the data set** by removing the group influence.

"*Life is unfair and we must accept it* " : The learning sample may be biased but this peculiar situation should not be taken as granted and generalized to all cases as Machine Learning does.

Mathematical Formalism for Fairness

Criteria of Fairness

Definition

A classifier $g : \mathbf{R}^d \to \{0,1\}$ achieves **Overall Accuracy Equality**, with respect to the joint distribution of (X, S, Y), if

$$\mathbb{P}(g(X) = Y \mid S = 0) = \mathbb{P}(g(X) = Y \mid S = 1).$$

But practical study of fairness must be based on $(\hat{Y} = g(X), S)$

Definition

A classifier $g : \mathbf{R}^d \to \{0,1\}$ achieves **Statistical Parity**, with respect to the joint distribution of (X, S), if

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1).$$

- The probability of a successful outcome is the same across groups

Full parity is too strong assumption so need for a quantitative criterion

Criteria of Fairness when we know $(X, S, \hat{Y} = g(X))$

Definition

The **Disparate Impact** of the classifier $g \in \mathcal{G}$, with respect to (X, S) is defined as

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)} := \frac{a(g)}{b(g)}.$$

- We consider classifiers g such that $a(g) > 0$ and $b(g) > 0$
- Ideal scenario: g achieves Statistical Parity $\Leftrightarrow DI(g, X, S) = 1$
- Statistical Parity is often unrealistic $\dashrightarrow$ relaxation

Definition

The classifier $g : \mathbf{R}^d \to \{0, 1\}$ has **Disparate Impact at level** $\tau \in [0, 1]$, with respect to (X, S), if $DI(g, X, S) \leqslant \tau$.

$DI(g, X, S)$ measures the level of fairness of g: the smaller the value of τ, the less fair it is

Confidence Interval for Fairness

Besse, P., Del Barrio, E., Gordaliza, P., and Loubes, J-M.
Statistical tests of unfairness in algorithmic decisions. (2018).

$$DI(g,X,S) = \frac{\mathbb{P}(g(X)=1 \mid S=0)}{\mathbb{P}(g(X)=1 \mid S=1)}$$

$$\downarrow$$

Statistic:

$$T_n := \frac{\sum_{i=1}^n 1_{g(X_i)=1} 1_{S_i=0} \sum_{i=1}^n 1_{S_i=1}}{\sum_{i=1}^n 1_{g(X_i)=1} 1_{S_i=1} \sum_{i=1}^n 1_{S_i=0}},$$

$$CLT \ + \ Delta\ Method \Rightarrow \frac{\sqrt{n}}{\sigma} \left(T_n - DI(g,X,S) \right) \xrightarrow{d} N(0,1), \ as\ n \to \infty,$$

$\left(T_n \pm \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \right)$ is a CI for $DI(g,X,S)$ asymptotically of level $1-\alpha$

Achieving Fair Learning

Achieving Fairness

To ensure Fairness we may consider

- Finding Classifiers such that $\mathcal{L}(g(X)|S=0)$ **is close to** $\mathcal{L}(g(X)|S=1)$ by adding **a penalty**
  Bechavod and Ligett (2017), Zafar et al. (2017), Donini et al. (2018)
- Modify the input data ⇒ break the relationship with the protected attribute
  (Feldman et al. (2015) and large literature) Changing the data X into $\tilde{X}$ a way such that $\mathcal{L}(\tilde{X}|S=0)$ **is close to** $\mathcal{L}(\tilde{X}|S=1)$ to gain that for all possible classifiers constructed using $\tilde{X}$

$$\text{for all} \quad DI(g, \tilde{X}, S) > \tau.$$

- The more fair maybe the less predictable (or predictable in a different way)
- **Total Repair** = Achieving Statistical Parity at all cost (v.s) **Partial Repair** with a trade-off between fairness and performance

Total Repair: achieving Statistical Parity

- **Goal:**

$$X \longrightarrow \tilde{X} \text{ such that } \mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right)$$

$$\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right), \forall g \in \mathcal{G}$$

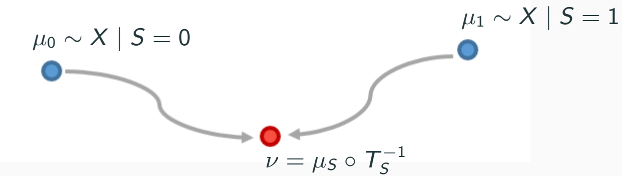
$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

- **Methodology:**

$$\begin{array}{rccc} T_S : & \mathbf{R}^d & \longrightarrow & \mathbf{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X) \end{array} \text{ s.t.}$$

$$\mathcal{L}\left(T_0(X) \mid S=0\right) = \mathcal{L}\left(T_1(X) \mid S=1\right)$$

- T_S depends on the binary random variable S



1. **Best choice for the distribution ν of the repaired variable?**
2. **Optimal way of transporting μ_1 and μ_0 to this new distribution ν?**

Total Repair: achieving Statistical Parity

- **Goal:**

$$X \longrightarrow \tilde{X} \text{ such that } \mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right)$$

$$\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right), \forall g \in \mathcal{G}$$

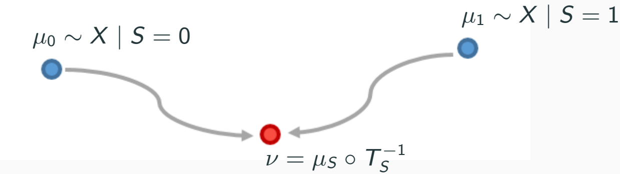
$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

- **Methodology:**

$$\begin{array}{rcccl} T_S : & \mathbf{R}^d & \longrightarrow & \mathbf{R}^d & \\ & X & \longmapsto & \tilde{X} = T_S(X) & \text{s.t.} \end{array}$$

$$\mathcal{L}\left(T_0(X) \mid S=0\right) = \mathcal{L}\left(T_1(X) \mid S=1\right)$$

- T_S depends on the binary random variable S



1. **Best choice for the distribution ν of the repaired variable?**

⇒ **Wasserstein barycenter** proposed in Fair Learning litterature

2. **Optimal way of transporting μ_1 and μ_0 to this new distribution ν?**

Total Repair: achieving Statistical Parity

- **Goal:**

$$X \longrightarrow \tilde{X} \text{ such that } \mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right)$$

$$\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right), \forall g \in \mathcal{G}$$

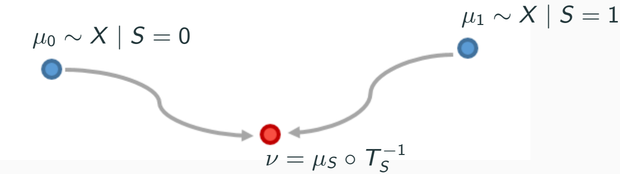
$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

- **Methodology:**

$$\begin{array}{rccc} T_S : & \mathbf{R}^d & \longrightarrow & \mathbf{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X) \end{array} \quad \text{s.t.}$$

$$\mathcal{L}\left(T_0(X) \mid S=0\right) = \mathcal{L}\left(T_1(X) \mid S=1\right)$$

- T_S depends on the binary random variable S



1. **Best choice for the distribution ν of the repaired variable?**

⇒ **Wasserstein barycenter** proposed in Fair Learning litterature

2. **Optimal way of transporting μ_1 and μ_0 to this new distribution ν?**

⇒ **Optimal Transport Maps**

Partial Repair : Random Repair (del Barrio et al. 2018)

Instead of moving all distributions , shift randomly a sufficient part of it

- Z target variable with general distribution μ
- $B \sim \mathcal{B}(\lambda)$, independent of (X, S, Y)
- $R_s = T_s^{-1}$, where $\mu = \mu_{s\sharp} T_s,\ s = 0, 1$,
  $R_0(Z) \sim \mu_0$ and $R_1(Z) \sim \mu_1$

$$\tilde{X}_\lambda = BT_s(X) + (1 - B)X$$

New Variable with λ as a trade-off

$$\tilde{\mu}_{s,\lambda} = \mathcal{L}(BZ + (1 - B)R_s(Z)) = \mathcal{L}(BT_s(X) + (1 - B)X \mid S = s),\ s \in \{0, 1\}$$

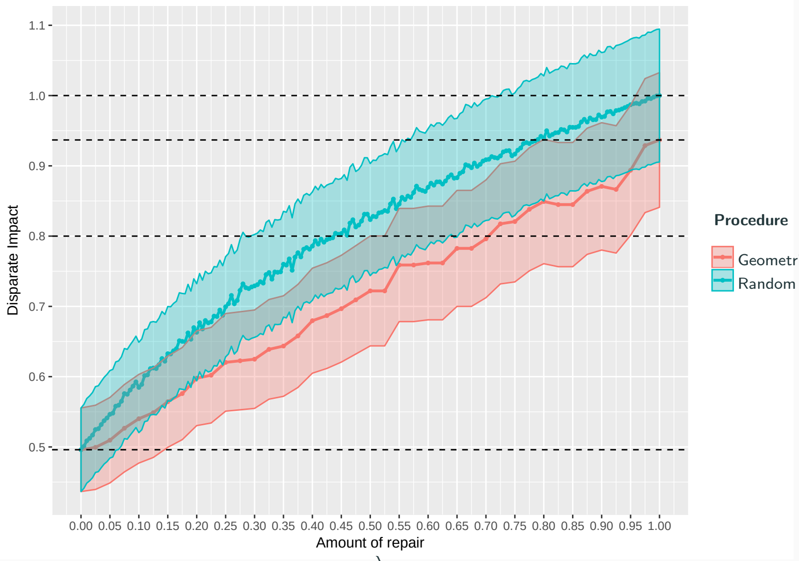
$$\lambda = 0 \quad \Rightarrow \quad \tilde{\mu}_{s,0} = \mathcal{L}(X \mid S = s) \qquad \text{Unmodified variable}$$
$$\lambda = 1 \quad \Rightarrow \quad \tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu \qquad \text{Totally repaired version}$$

Application to a real data example: Adult Income data

Disparate Impact and Accuracy of the classification with the repaired data set

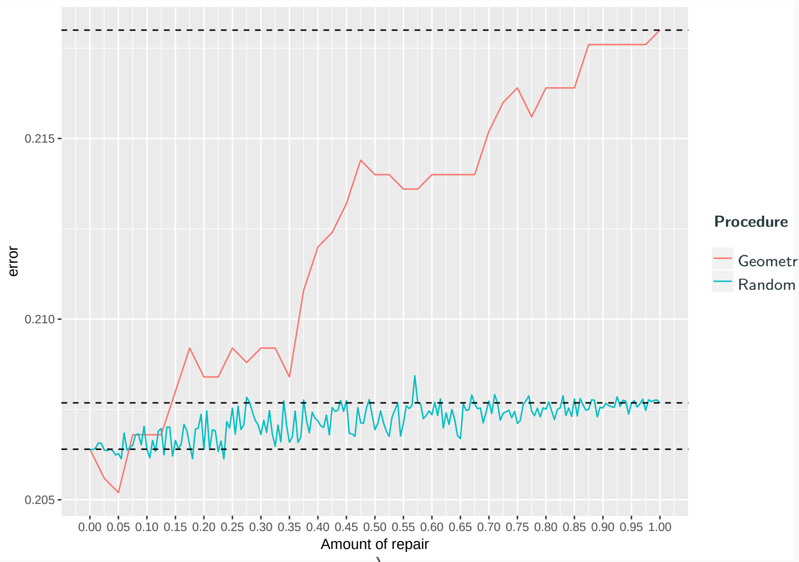
Statistical Model	Repair	Error	Difference	$\hat{DI}$	CI 95%
Logit	**(A)**	0.218	0.0116	0.937	$(0.841, 1.033)$
Logit	**(B)**	0.2077	0.00128	1	$(0.905, 1.095)$
Random Forests	**(A)**	0.2272	0.0592	1.1	$(0.976, 1.223)$
Random Forests	**(B)**	0.2045	0.0365	1	$(0.886, 1.114)$

95% Confident Interval for DI of logit



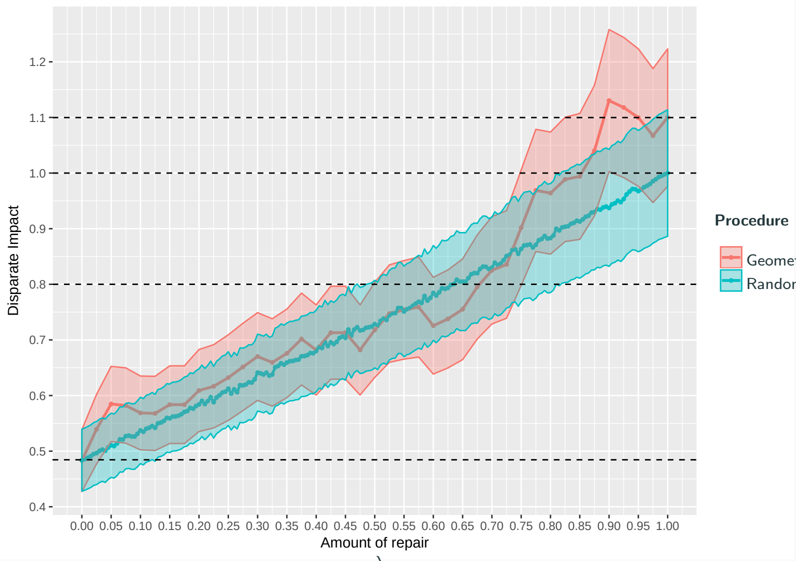
λ

Classification error



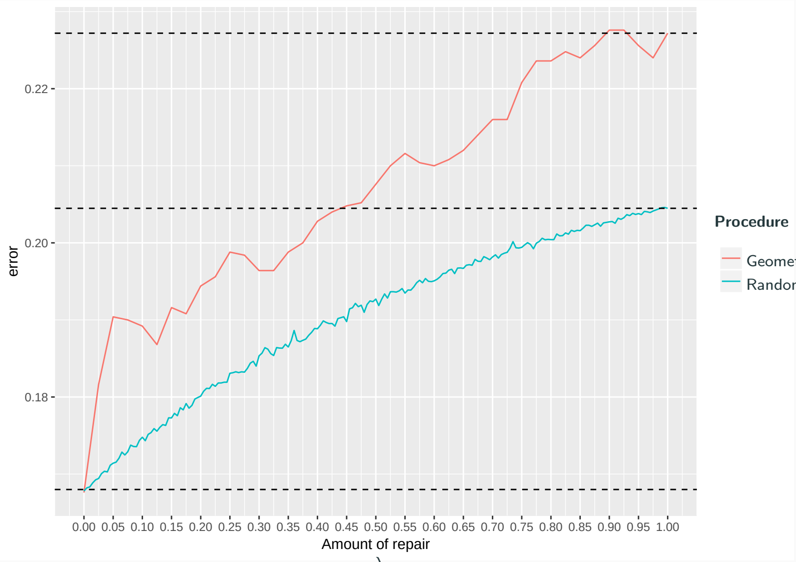
λ

95% Confident Interval for DI of random forests



λ

Classification error



λ