ANITI

ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Robust and Fair Artificial Intelligence

Jean-Michel Loubes

September 9-10 2019

Université Fédérale
Toulouse Midi-Pyrénées

Agenda

Chair members

Context

Objective

Outline

Chair members

- Jean-Michel Loubes : Professor at Mathematical Institute of Toulouse.
  Research : Mathematical statistics, Machine Learning
- Mathieu Serrurier : Assistant Professor at Institut de Recherche en Informatique de Toulouse.
  Research: Causality, Deep Learning, Adversarial Networks.
- Beatrice Laurent : Professor at Mathematical Institute of Toulouse.
  Research: Mathematical Statistics, Tests.

Context

Machine Learning methods aim at learning the relationships between characteristic variables X and a target variable Y to be able to forecast new observations.

The **distribution** of the learning sample is used to make inference to future observations.

The chair aims at answering the question : what happens when the distribution of the target does not correspond to the one of the target sample

- because the learning sample is biased : **Fairness**
- because the distribution of the target is different : **Robustness**
- because some information should be removed or hidden **Differential Privacy**

Fairness

•An algorithm suffers from **unfairness** if its outcome (decisions) is (partly) based on a variable S that *should* not play a decisive role in the decision making process.

•S is called **sensitive attribute** and is a variable that divides the observations into subgroups while the algorithm should not show a different behaviour over these subsets. If the algorithm does *not depend* on S, it will be **fair**.

•We assume that the algorithm is not meant to be unfair (not unfair by design) but the possible unfairness comes from the **learning process** in a machine learning framework.
•Privacy means that an observer seeing its output cannot tell if a particular individual's information was used in the computation.

Mathematical Formalism

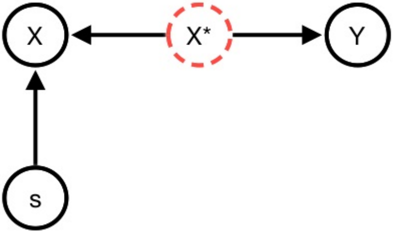
- Y **target**
- $X : \Omega \to \mathbb{R}^d,\ d \geqslant 1$, **visible attributes**
- $S : \Omega \to \{0, 1\}$ which induces a bias **protected attribute**

$$S = \begin{cases} 0 \quad minority \quad (unfavored) \\ 1 \quad majority \quad (favored) \end{cases}$$

Fairness deals with the relationships between Y, $\hat{Y}$ and S : S is related to (X, Y) and produces bias that may not be desired between the two groups driven by S.

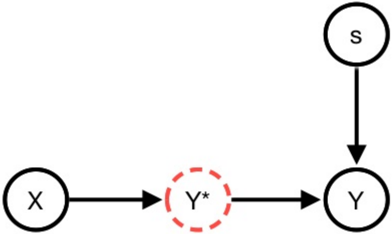
S is chosen by the practitioner and its choice is driven by legal, ethic or technical issues.

Mathematical Model for Fairness



The attributes X are a biased version of unobserved fair attributes $X^\star$, while the target variable Y depends only on $X^\star$ and is fair. Learning from X induces biases while fairness enables a most accurate forecast.

Mathematical Model for Fairness



The decision Y observed is the result of a fair score $Y^\star$ which has been biased by the uses giving rise to Y.

Objective

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Removing or controlling the *bias effect* in the Machine Learning Process enables

1. to obtain **Fair procedures** (legal or societal compliance for acceptability)
2. to obtain **Robust procedures** to be generalized without the effect of a variable that can affect proxies leading to **Transfert Learning**.

Outline :

- How to measure fairness and then detect it ?
- Achieving Fairness by correction on the database.
- Achieving Fairness by controlling the algorithm.
- Achieving Fairness by changing the post-processing the output of the algorithm.

Aims on Fairness

- Mathematical foundations of Fairness : develop theoretical framework and feel gaps between theory and algorithms
- New Measures of Fairness : independency of the conditional distributions $\mu_s = \mathcal{L}(f(X)|S = s)$ using Optimal Transport theory and Monge-Kantorovich distance and using causality.
- New algorithms to repair the data or to build fair classifiers with provable guarantees (using gradient descent or adversarial networks).
- Applications to robust and transfert learning for critical systems.
- Fairness for Natural Language Processing.

Aims on Robustness & Privcacy

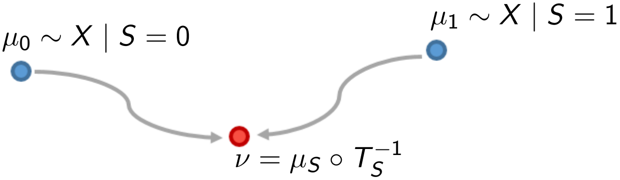
- Creation of stress models to test algorithms by imposing deformations on the distributions. (directly using optimal transport, using entropic methods, using variational auto-encoders)
- Sensitivity Analysis to understand uncertainty propagation of the algorithm or the learning process (provides explainability of black box models).
- Link between differential privacy and fairness.
- Bounds for statistical learning under privacy assumptions.

Achieving Statistical Parity (Proceedings ICML 2019)

Methodology: Find a transformation for each group such that

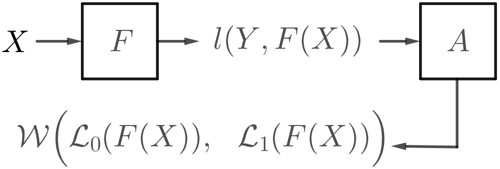
$$\begin{array}{rccc} T_S : & \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X) \end{array}$$

s.t. $\nu := \mathcal{L}\left(T_0(X) \mid S = 0\right) = \mathcal{L}\left(T_1(X) \mid S = 1\right)$

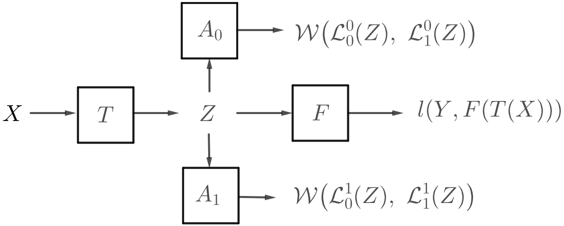


- The target distribution ν has to be chosen in order to convey *enough* information on the link between X and Y.

Fairness penalty : Adversarial networks

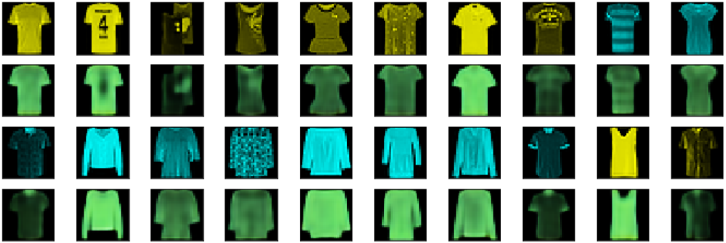


Networks for equality of odds



Networks for equality of opportunities

Fairness penalty : Illustration



Planned PhD / post doc proposals

- Phd on robustness of sensivity analysis of learning process starting (november 2019 joint with E. Pauwels)
- Phd on optimal transport and regularization for robust machine learning (joint with M. Serrurier and E. del Barrio) proposal for 2020
- Propositions of industrial Phd (Continental, Liebherr, Quantmetry...)
- Post-doc : 1/2020 Transfert Learning
- Post-doc : 09/2020 Robust Tests with differential privacy and applications to supervised classification (supervised by B. Laurent)

Interaction with other chairs / industrial

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

- Existing Links with geometrical methods (F. Gamboa) due to manifold structure of set of distributions
- Existing Links with Law and Ethics (C. Castets-Renard)
- Links to be drawn with AI by argumentation and persuasion (Leila Amgoud) and all chairs related to certification (Daniel Delahaye) and optimisation (numerous chairs)
- Industrial Applications (DEEL project)