(Deep) Machine Learning Algorithms
Bias & Explainability Challenges for Regulation

J.-M. Loubes, Professor
&
L. Risser, AI Research Engineer
Chair : C. Benesse, L de Lara, A. Gonzalez, B. Laurent, M. Serrurier
UPS, CNRS & Chair @ Artificial & Natural Intelligence Institute of Toulouse (ANITI)
Part 1: Principles and Dangers of Machine Learning
Mathematical Guarantees in Machine Learning

Goal
• Learning the relationships between characteristic variables $X$ and a target variable $Y$.
• Being then be able to forecast new observations.

Learning Sample
I.i.d. observations with unknown distribution $\mathbb{P}$: $(Y_1, X_1), \ldots, (Y_n, X_n)$.

Machine Learning Algorithm $\hat{f}_n$ for a given risk $R(f) = \ell(y, f(x))$
Train the best model among a class of algorithms $\mathcal{F}$, based on the observations:

$$\hat{f}_n \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) \right\}$$

Unknown oracle rule. $f^* \in \arg \min_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}} \{ \ell(Y, f(X)) \}$

$\rightarrow$ Mathematical guarantees on $\hat{Y} = \hat{f}_n(X)$: Control of generalization error

$$\mathbb{E}_{\mathbb{P}} \{ \ell(Y, \hat{f}_n(X)) \} - \mathbb{E}_{\mathbb{P}} \{ \ell(Y, f^*(X)) \} \leq \varepsilon$$
Big Data paradigm
• The Data convey all the information.
• The more the data the more accurate the description of the reality.

→ From data to information: extraction of the knowledge from empirical observations

Need for Large amount of data of good quality

Principle of Machine Learning
• Learn decision rules fitting the data using a set of labeled examples (learning sample).
• The learned decision rules will be used for all the population.
• The whole population is supposed to follow same distribution as the learning sample.

→ The Machine Learning algorithm (or AI) learn the best rule from the data and then can forecast new observations with a guaranteed precision.

Need for Complex Models
Applications of Machine Learning Algorithms

Development of such algorithms for a **large number of applications in all fields of our lives** even critical ones (health, finance, justice, education, transports, resources management …)

Classified **High Risk Use Cases** by European Community AI Act

- **Credit Scoring**
- **Personalised Medicine**
- **Autonomous Vehicles**
- **Pattern Detection**
- **Time series Forecasting**
Need for Regulations and Law: first principles

Amazon, Facebook, Google, IBM, Microsoft... (2015)
Bruxelles, le 21.4.2021
COM(2021) 206 final
2021/0106 (COD)

Proposition de

RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL

ÉTABLISSANT DES RÈGLES HARMONISÉES CONCERNANT L’INTELLIGENCE ARTIFICIELLE (LÉGISLATION SUR L’INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L’UNION

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}
Artificial Intelligence Act (**April 2021**) by European Commission

- Definition of **High Risk domains** of applications (health, finance, public services, transports …)
- Performance matters but not only: notions of **equity, transparency** and robustness
- Need for **definitions of norms** to measure bias (AFNOR, IEEE, …)
- Need for **explainable & understandable** decisions
- **Primum non nocere**

Works in progress to Certify AI based systems (for cars, airplanes …)
Part 2: Bias in Machine Learning
General Data Protection Regulation (GDPR) & European AI Act (2021)

- Effective in the E.U. since 05/2018
- According to the GDPR, automatic decisions taken by an algorithm should be:
  - *un-biased*
  - *not discriminant*
  - *fair*
  - *with the same performance as regards the persons or the groups of persons*

More generally

- E.U. (GDPR, art 22-4 2018): "A decision is declared fair if it is neither based on affiliation to a protected minority group, nor based on the explicit or implicit knowledge of sensitive personal data."
- NYC Bill (Dec. 2017) : local decision
- Several Trials (USA-Canada)
- …
Bias leads to unfairness and **personal or group** discrimination.

- August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark:
  - 98.7% for DARKER MALES
  - 68.6% for DARKER FEMALES
  - 100% for LIGHTER MALES
  - 92.9% for LIGHTER FEMALES

- Amazon Rekognition Performance on Gender Classification:
  - Skyscrapers
  - Airplanes
  - Cars
  - Bikes
  - Gorillas
  - Graduation

- COMPAS recidivism black bias:
  - DYLAN FUGETT: Prior Offense 1 attempted burglary, Subsequent Offenses 9 drug possessions, Risk 3
  - BERNARD PARKER: Prior Offense 1 resisting arrest without violence, Subsequent Offenses None, Risk 10

- Equality of Odds

- Statistical Parity
Data & Machine Learning are subjected to bias

- ML Algorithms amplify pre-existing bias
- or maintain a biased status-quo
- Auto-prophetic algorithm shape biased worlds
- Accuracy is not enough ....
An A.I. algorithm suffers from **unfairness** if its outcomes $Y$ (decisions) are fully or partly based on a **sensitive variable** $A$ that *should* not play a decisive role in the decision making process.

**Statistical Parity** : $\hat{Y} \perp A$

**Equality of Performance** : $\hat{Y} \perp A \mid Y$

Being **globally fair** is a probabilistic notion of dependency or conditional dependency.

Measures of fairness are numerous and correspond to measuring joint effects which are complex in high dimensions since « **Biases are everywhere** ».

1. **Disparate Treatment** for all $x$,
   
   $\mathbb{P}(\hat{Y} = 1 \mid X = x, A = 0) - \mathbb{P}(\hat{Y} = 1 \mid X = x, A = 1)$

2. **Avoiding Disparate Treatment** :
   
   $\mathbb{P}(\hat{Y} \neq Y \mid A = 0) - \mathbb{P}(\hat{Y} \neq Y \mid A = 1)$

3. **Predictive Parity**
   
   $\mathbb{P}(Y = 1 \mid \hat{Y} = 1, A = 0) - \mathbb{P}(Y = 1 \mid \hat{Y} = 1, A = 1)$

4. **For Quantitative case**
   
   $\min \text{Var}_A \mathbb{E}(\hat{Y} \mid A) \quad \min \text{Var}_A \mathbb{E}(\ell(\hat{Y}, Y) \mid A)$
Granting a Loan by minimising Risk « Adult Data set (UCI database) »

$\text{n} = 48842$ observations (individuals) described by $p = 14$ variables

Objective: Forecast if a credit can be given (future salary > 50k$)

Problem: Not balanced w.r.t to variable « A = Sex »
Illustration on the *Adult Income* dataset — Disparate impact and accuracy

**Disparate Impact** w.r.t variable *Sex* considered as sensitive variable A

\[
Ref = DI(Y, X, A) = \frac{\mathbb{P}(Y = 1 \mid A = 0)}{\mathbb{P}(Y = 1 \mid A = 1)}
\]

\[
DI(f, X, A) = \frac{\mathbb{P}(f(X) = 1 \mid A = 0)}{\mathbb{P}(f(X) = 1 \mid A = 1)}
\]

- Statistical increase of discrimination between A=1 (Men) et A=0 (Women)
- « Gender » variable leads to discrimination
What says the law? High quality data without discriminative variables.

GDPR or AI’s Act focus on quality of the dataset
Sensitive variables should not be used: \(A=\text{Sex is removed from the learning sample}\)

Disparate Impacts

Accuracies

Bias is not modified → comes from **correlations** and not only the \(A\) variable
L'apprentissage automatique semble renforcer les biais existants dans la société.

A Survey of Bias in Machine Learning Through the Prism of Statistical Parity

Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes & Laurent Risser

Received 01 Apr 2020, Accepted 02 Jul 2021, Accepted author version posted online: 13 Jul 2021, Published online: 25 Aug 2021
Choose a definition for fairness (mainly based on conditional independence) & pay a price for fairness

**Three main ways of obtaining fairness** according to the criterion which is chosen

1. **Pre-processing** the learning sample and removing the effect of the sensitive variable such that the algorithm does not take into account the effect of the variable that creates the biased behaviour.

   \[ X \leftrightarrow \tilde{X} \leftrightarrow f(\tilde{X}) \]

2. **Constraining the algorithm** by adding a fairness constraint

   \[ \hat{f} \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda I(f) \]

3. **Post-processing** the outcome of the algorithm to comply the fairness restrictions.

   \[ f(X) \mapsto \Phi_{\text{fair}}(f(X)) \]

\[
W_c(\mu_0, \mu_1) = \inf_{\Pi \in \mathcal{P}(\mu_0, \mu_1)} \int c(x, y) d\Pi(x, y)
\]
Fairness constraint for Deep Neural Network

**Back-propagation of Fairness constraints in Neural Networks:**

\[
\hat{\theta} = \arg \min_{\theta} R(\theta) + \lambda W^2_2(\mu_{\theta,0}^n, \mu_{\theta,1}^n)
\]

Optimal Transport distance (Wasserstein distance) to enforce both distributions to be the same

Risk

Fairness Constraint

Loubes et al. (ICML 2019)
Bias and Robustness w.r.t change of context

EuroSAT dataset ([https://madm.dfki.de/downloads](https://madm.dfki.de/downloads)) : 27,000 remote sensing images / 10 classes

Blue shade effect ($\approx 3\%$)

Automatic Classification between Roads and Rivers is hampered by « Blue shade » variable
Examples of Applications in Econometry

• Gender Effect in microfinance

• Finding Instruments in Instrumental Variable Regression without using some variables (protected variables)

• Constraining the IV regression to be independent from a sensitive attribute
Part 3: 3.1 Explainability in Machine Learning
**Explanability techniques in M.L.**

**Need for explainability**

**Emergence of a *Right to explanation***

- E.U. (RGPD, art 22 — 2018) : « Right not to be subject to a decision solely based on automated processing, including profiling »
- Fr (Loi Informatique et Libertés) : « Right to understand the rules of automatic treatments and their main characteristics »
- NYC Bill (Dec. 2017) : Local laws related to automatic decision systems
- E.U (AI Act - 2021) : « Necessity to be able to correctly interpret and understand the high-risk AI system’s output » (Art 13) « sufficiently transparent to enable users to interpret the system’s output and use it appropriately. »

**Exemples of recent works**

- Besse, Castet-Renard, Garivier, Loubes : L’I.A. du quotidien peut-elle être éthique? Statistique et société 6(3), 2018 — [https://www.youtube.com/watch?v=RwsMy0ILXos](https://www.youtube.com/watch?v=RwsMy0ILXos)
- Packages Grad-Cam, Lime, GEMS-AI
- …
1) Introduction — Unexplainable prediction model

Example of clearly unexplainable model → convolutional neural network:

class basicCNN(nn.Module):
    def __init__(self):
        super(basicCNN, self).__init__()
        #Convolution/ReLU/MaxPooling layers
        self.conv1 = nn.Conv2d(1, 2, kernel_size=2, stride=1, padding=1) #1 to
        self.pool1 = nn.MaxPool2d(kernel_size=2, stride=2) #32x32 to 16x16
        self.conv2 = nn.Conv2d(2, 4, kernel_size=2, stride=1, padding=1) #2 to
        self.pool2 = nn.MaxPool2d(kernel_size=2, stride=2) #16x16 to 8x8
        self.conv3 = nn.Conv2d(4, 8, kernel_size=2, stride=1, padding=1) #4 to
        self.pool3 = nn.MaxPool2d(kernel_size=2, stride=2) #8x8 to 4x4

        #Dense layers
        self.fc1 = nn.Linear(8 * 4 * 4, 32)
        self.fc2 = nn.Linear(32, 10)

    def forward(self, x):
        x = F.relu(self.conv1(x))
        x = self.pool1(x)
        x = F.relu(self.conv2(x))
        x = self.pool2(x)
        x = F.relu(self.conv3(x))
        x = self.pool3(x)
        x = x.view(-1, 8*4*4) #flatten the data
        x = F.relu(self.fc1(x))
        x = self.fc2(x)
        return(x)

Mnist: predicting Digits

https://github.com/gwding/draw_convnet
Need for explainability to trust the model

Strong interest to certify algorithmic decisions → robust decision making + towards certifiable IA

Example:

Suppose that the predictions are generally accurate:
- Which features were used to take the decision?
- If inadequate features were used, the NN is likely to generalise poorly!
Part 3: 3.2 Explainability in Machine Learning

Solutions & Research
Surrogate Models → LIME (Local interpretable model-agnostic explanations)

“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington, USA
marcotcr@cs.washington.edu

Sameer Singh
University of Washington,
Seattle, WA 98195, USA
sameer@cs.washington.edu

Carlos Guestrin
University of Washington,
Seattle, WA 98195, USA
guestrin@cs.washington.edu

https://homes.cs.washington.edu/~marcotcr/blog/lime/
https://github.com/marcotcr/lime

Training a **local surrogate models** to explain the prediction of $X_i$ with $f_\theta$

**Drawbacks**: NN are highly non linear and local models can be very different

Our neural-network prediction model $f_\theta$ …

\[ f_\theta(X_i) = (0.97, 0.02, \ldots, 0.07) \]

… can become a linear, and straightforwardly interpretable, model $g_\theta$ for images close to $X_i$:

Chosen model can be linear regression or decision tree (interpretable models)

Weighted sum of the intensities with weights:

\[ g_\theta(X_i) = (0.95, 0.03, \ldots, 0.05) \]

(followed by logistic function)
Instead of back-propagating the derivatives of the risk \( R \), it is possible to back-propagate the derivatives of a specific value in the N.N. outputs.

Sensitivity to the input → Grad-CAM

Represented how \( y^c \) is sensitive to the N.N. inputs (for the tested image)

Guided Backpropagation

Slightly modified back-propagation

[Springenberg et al. 2014]

http://gradcam.cloudcv.org/
https://github.com/ramprs/grad-cam/
3) Three explainability solutions → Grad-CAM

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA

Results

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>#1 boxer</th>
<th>#2 bull mastiff</th>
<th>#3 tiger cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grad-CAM [1]</td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
</tr>
<tr>
<td>Guided backpropagation [2]</td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
</tr>
<tr>
<td>Guided Grad-CAM [1]</td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
<td><img src="https://github.com/ramprs/grad-cam/" alt="Image" /></td>
</tr>
</tbody>
</table>
**Sensitivity Analysis** for AI Algorithms. used to certify computer code

(Used in nuclear safety for instance)

Quantification of the dependency of an output w.r.t changes of input parameters

**Sobol indices or Shapley values methods** …. (Also to quantify the variability of a bias criterion and understand the root of the bias) Fairness seen as Global Sensitivity Analysis work by Benesse et al. https://arxiv.org/abs/2103.04613

Sobol indices when Prediction Myocardial Infarction
3) Three explainability solutions → Gems-AI : explanation under stress

Explaining Machine Learning Models using Entropic Variable Projection
François Bachoc¹, Fabrice Gamboa¹,², Max Halford², Jean-Michel Loubes¹,³ and Laurent Risser¹,³

1Institut de Mathématiques de Toulouse
2 Institut de recherche en informatique de Toulouse
3 Artificial and Natural Intelligence Toulouse Institute (3IA ANITI)

https://www.gems-ai.com/
https://github.com/XAI-ANITI/ethik

« What-if machine » for group-explainability : Explaining models under stress

Intuition : Re-weighting the observations \( \{X_i, Y_i\}_{i=1,\ldots,n} \) to stress the distributions of the data transform a specific property of the test set in average.

Test set
\[
\{X_i, Y_i\}_{i=1,\ldots,n}
\]

\[
P_n = \frac{1}{n} \sum_{i=1}^{n} \delta(X_i, Y_i)
\]

« Black-box » decision rules

\[
\begin{align*}
X_i^1 & \quad \rightarrow \\
X_i^2 & \quad \rightarrow \\
X_i^3 & \quad \rightarrow \\
& \quad \ldots \\
X_i^p & \quad \rightarrow \\
\end{align*}
\]

\[\hat{Y}_i := f(X_i)\]

Modify Input Distribution under constraint:

\[
\arg \min_{Q} \left\{ KL(Q \mid P_n), \text{s.t. } \int \Phi(X, Y) dQ = \lambda \right\}
\]
Theorem 2.1. Let \( t \in \mathbb{R}^k \) and \( \Phi : \mathbb{R}^{p+2} \to \mathbb{R}^k \) be measurable. Assume that \( t \) can be written as a convex combination of \( \Phi(X_1, \hat{Y}_1, Y_1), \ldots, \Phi(X_n, \hat{Y}_n, Y_n) \), with positive weights. Assume also that the empirical covariance matrix \( \mathbb{E}_{Q_n}(\Phi \Phi^\top) - \mathbb{E}_{Q_n}(\Phi)(\Phi)^\top \) is invertible.

Let \( \mathbb{P}_{\Phi, t} \) be the set of all probability measures \( P \) on \( \mathbb{R}^{p+2} \) such that \( \int_{\mathbb{R}^{p+2}} \Phi(x) \, dP(x) = t \). For a vector \( \xi \in \mathbb{R}^k \), let \( Z(\xi) := \frac{1}{n} \sum_{i=1}^n e^{\langle \Phi(X_i, \hat{Y}_i, Y_i), \xi \rangle} \). Define now \( \xi(t) \) as the unique minimizer of the strictly convex function \( H(\xi) := \log Z(\xi) - \langle \xi, t \rangle \). Then,

\[
Q_t := \underset{P \in \mathbb{P}_{\Phi, t}}{\text{arg inf}} \, \text{KL}(P, Q_n)
\]

exists and is unique. Furthermore, we have

\[
Q_t = \frac{1}{n} \sum_{i=1}^n \lambda_i^{(t)} \delta_{X_i, \hat{Y}_i, Y_i},
\]

with, for \( i = 1, \ldots, n \),

\[
\lambda_i^{(t)} = \exp \left( \langle \xi(t), \Phi(X_i, \hat{Y}_i, Y_i) \rangle - \log Z(\xi(t)) \right).
\]

Consistent Estimation:

\[
\mathcal{W}_1(Q_t, Q^*) = O_p \left( n^{-1/(p+2)} \right).
\]
3) Three explainability solutions → Entropic Variable Projection

Explainability techniques in M.L.

Explaining Machine Learning Models using Entropic Variable Projection
François Bachoc¹, Fabrice Gamboa¹,³, Max Halford², Jean-Michel Loubes¹,³ and Laurent Risser¹,³

¹Institut de Mathématiques de Toulouse
² Institut de recherche en informatique de Toulouse
³ Artificial and Natural Intelligence Toulouse Institute (3IA ANITI)

Example: Automatic decision to grant a loan.
What-if the average age is 50 instead of 42 in the test set?

<table>
<thead>
<tr>
<th>Age (X¹)</th>
<th>Education.num (X²)</th>
<th>Marital.status (X³)</th>
<th>Hours.per.week (X⁴)</th>
<th>...</th>
<th>Loan granted — True (Y)</th>
<th>Loan granted — Predicted (ŷ = f_θ(X))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.05</td>
<td>54</td>
<td>Divorced</td>
<td>40</td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>0.83</td>
<td>41</td>
<td>Never-married</td>
<td>60</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1.15</td>
<td>51</td>
<td>Married-civ</td>
<td>40</td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>0.81</td>
<td>39</td>
<td>Married-civ</td>
<td>65</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1.15</td>
<td>49</td>
<td>Divorced</td>
<td>50</td>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Compute optimal weights

Advantages:
- Small Algorithmic cost in high-dimension
- Evaluate Robustness and Resiliency w.r.t realistic stress conditions
- Explain effects on decision and risks
- Mathematical guarantees on convergence.

Explain how the outputs vary
3) Three explainability solutions → Entropic Variable Projection

Explaining Machine Learning Models using Entropic Variable Projection
François Bachoc¹, Fabrice Gamboa¹, Max Halford², Jean-Michel Loubes¹,³ and Laurent Risser¹,³

¹Institut de Mathématiques de Toulouse
²Institut de recherche en informatique de Toulouse
³Artificial and Natural Intelligence Toulouse Institute (3IA ANITI)

What-if the average […] is […] instead of [original average value] in the test set?

[Graphs showing logistic regression, XGBoost, and Random forest results for different variables like age, education-num, capital-gain, capital-loss, and hours-per-week.]
When Interpretability and Bias collide

The background bias

S the confounding variable is here the **snow** but it is hidden since not encoded in the data base. Need to **unveil the bias with explainability**.
Main Question:

How to certify the behaviour of a Neural Network?

Regulations require a better understanding of Deep Networks:

1. Need for Quantification of Biases in the dataset but also of its propagation by the algorithm
2. Explainability & Transparency of Algorithmic Decisions
3. Need for proper definitions and norms
4. Need for sandboxes, and use-cases

Need to work together between designers of algorithms and regulators
Not complete at all Bibliography ...


- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil, Fair Regression Wasserstein barycenter, Neurips 2020.


Toolbox : https://www.gems-ai.com/