Fair Adversarial Network and explainability

Simon Grah [1] & Ilyes Mahammed Chikouche [1]& Vincent Thouvenot [1]

[1] *Thales SIX GTS France, 1 Avenue Augustin Fresnel, Palaiseau*
prenom.nom@thalesgroup.com

Résumé. Les modèles de Machine Learning ont tendance à reproduire et à amplifier les biais. Certaines techniques, s'appuyant sur la Fairness, permettent de lutter contre ces biais. La Fairness est l'un des sous-thèmes du "Trustable Machine Learning". Un autre sous-thème est l'explicabilité. En Machine Learning, nous utilisons certaines techniques souvent très fortement non linéaires. Comprendre le comportement des modèles est une tâche compliquée. Dans cet exposé, nous utilisons un réseau advseraire issu de la littérature pour limiter l'influence de certains paramètres sur un classifieur afin de respecter les contraintes de Fairness. Nous mesurons le gain en Fairness par rapport à un classifieur classique avec une métrique nommée la p-rule, avec des métriques basées sur la distance entre deux distributions et avec la Differential Fairness. Pour avoir une explication locale de la différence entre le classifieur classique et le classifieur adversaire, nous calculons les valeurs de Shapley, une technique d'explication classique issue de la théorie des jeux, dont la définition dépendra d'une population de référence, et non de la prédiction moyenne comme il est souvent fait. Enfin, nous montrons l'apport du réseau adversaire en termes de protection des données. Les travaux ont été réalisé dans le cadre du projet H2020 SPARTA.

Mots-clés. Explainabilité, Fairness, Réseau adversaire, Valeur de Shapley

Abstract. Machine Learning models tend to reproduce and amplify biases. Some techniques, relying on Fairness, allow to fight against these bias. Fairness is one of the sub-topic of Trustable Machine Learning. Another sub-topic is explainability. In Machine Learning, we use some techniques often very strongly non-linear and understanding models behavior is a hard task. In this talk, we will use an adversarial network coming from literature to limit influence of some parameters on a classifier to achieve Fairness. We will measure the gain of Fairness compared to a classifier with the base rate metric named p-rule, with metrics based on the distance between two distributions and with the Differential Fairness. To have local explanation of the difference between the classical and the fair classifier, we will compute the Shapley Value, a classical explanation technique coming from Game Theory, whose definition will depend on a reference population, and not the average prediction as often seen. Finally, we will show the contribution in terms of data protection of the fair classifier. This work is funded under the SPARTA H2020 project.

Keywords. Adversarial Network, Explainability, Fairness, Shapley Value

1 Introduction

1.1 Fairness

Machine Learning models tend to reproduce and amplify biases. These biases can come from the data: there are known biases such as selection bias when sampling is poor, historical biases, for example when a population is disadvantaged, etc. There are also biases that can come from algorithms: some recommendation algorithms lock people in bubbles instead of offering them new possibilities, when data are unbalanced, etc. There are several possible definitions of Fairness. First, we can wish that the model outputs would be similar for two subpopulations. It is to be hoped that the errors in the model will be similar for the different subpopulations. Finally, we can wish that two similar individuals except the sub-population to which they belong receive similar treatment.

1.2 Explainability

In Machine Learning (ML), and more generally in artificial intelligence, we mainly focus on performance. Highly non-linear models such as Random Forests, Gradient Boosting, Deep Learning, etc. are often used. Generally, performance are satisfactory. However, we are unable to explain how the model builds its decisions. This can be a problem, especially when dealing with critical systems. There are several levels of explanation in ML models. The first level is global: we want to understand the model general behavior and the features impact on the model's outputs. At local level, we explain why a prediction was made for an observation. The explanation position has to be defined. Do we learn a model that is interpretable by nature, like e.g. additive model ? Do we learn the model to generate an explanation at the same time as we make a prediction (see e.g. Baratt, 2017)? Does the explanation require an additional component, independent of the type of model used (see e.g. Ribeiro *et al*, 2016)? Some objectives of explainability are to verify ML functionality and increase confidence of users. Since these systems are used as decision support, the task is to design an approach to optimize the relation between performance boost provided by these systems and explainability of its decisions. This explainability can be directed to the engineer developing these systems, to improve it, and the final users of these systems who needs to understand the decisions made, for being more confident in the results and enriching the by its business knowledge.

2 Fair Adversarial Network

2.1 Fairness measure

In this subsection, we give some classical measure of Fairness on a binary classification task (see e.g. Friedler *et al.*, 2018, Zafar *et al.*, 2015). Note a binary class prediction

$\hat{Y} \in \{0,1\}$ and a binary sensitive attribute $Z \in \{0,1\}$. A base rate metrics measures the change of models' outputs according to a sensitive attribute. For instance, p%-rule is given by $\min\left(\frac{P(\hat{Y}=1|Z=1)}{P(\hat{Y}=1|Z=0)}, \frac{P(\hat{Y}=1|Z=0)}{P(\hat{Y}=1|Z=1)}\right) \geq \frac{p}{100} \in [0,1]$. Group-conditioned accuracy metrics measure the model's errors according a sensitive attribute. For instance, the Z-accuracy is given by $P(\hat{Y} = y|Z = z, Y = y) \in [0,1]$. The group-conditioned calibration measure the label repartition knowing model prediction according to an attribute. For instance, Z-calibration + is given by $P(Y = 1|Z = z, \hat{Y} = 1) \in [0,1]$. Individual-level discrimination measure how the model handle one individual comparing the most similar individuals. The consistency is given by $1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in knn(i)}|\hat{Y}_i - \hat{Y}_j| \in [0,1]$, where $knn(i)$ are the K-Nearest Neighbors of i.

Foulds and Pan (2018) proposes the Differential Fairness which states that a mechanism $M(x)$ is ε-differentially fair in a framework (A, Θ), where A is the ensemble of attributes to protect, if for all $\theta \in \Theta$ with $x \sim \theta$ and $y \in Range(M)$,

$$\exp(-\varepsilon) \leq \frac{P_{M,\theta}(M(x) = y|\mathbf{s_i}, \theta)}{P_{M,\theta}(M(x) = y|\mathbf{s_j}, \theta)} \leq \exp(\varepsilon),$$

for all $(\mathbf{s_i}, \mathbf{s_j}) \in A \times A$, where $P(\mathbf{s_i}|\theta) > 0$, $P(\mathbf{s_j}|\theta) > 0$.

2.2 Adversarial Network

Several techniques have been proposed to achieve Fairness. Some are based on regularization methods: the direct and indirect information which rely to the sensitive attribute(s) are penalized (see e.g. Raff *et al.*, 2017). Some others techniques will use new representations that hide attribute, e.g. by Deep Learning (see e.g. Louizos *et al.*, 2015). Some others techniques introduce fairness constraint by modifying the inputs or the outputs of the algorithms (see e.g. Kamiran *et al.*, 2018).

We consider the adversarial network proposed by Louppe *et al.*, 2017. They use this architecture in context where they want to introduce independence between the outputs of a classifier and some nuisance parameters. The overview of the architecture is given by Figure 1. Our objective will be to use this architecture to make independent the outputs of a classifier of constituent elements to a sub-population to be protected or to elements which we wish that they cannot be revealed by the model's outputs.

Note $\hat{\mathbf{Y}}$ classifier prediction based on input X, $\mathbf{Y}$ the true value and A the sensitive attributes, $\theta_{\mathbf{clf}}$ and $\theta_{\mathbf{ad}}$ parameters of respectively the classifier and the adversarial networks, $Loss_Y(\theta_{\mathbf{clf}})$ and $Loss_A(\theta_{\mathbf{clf}}, \theta_{\mathbf{ad}})$ the loss of a pre-trained classifier and adversarial networks.

During the iteration, the objective function given by Equation (1) is considered:

$$\theta_{\mathbf{clf}}, \theta_{\mathbf{ad}} = \arg\min_{\theta_{\mathbf{clf}}}\max_{\theta_{\mathbf{ad}}}\left(Loss_Y(\theta_{\mathbf{clf}}) - \lambda Loss_A(\theta_{\mathbf{clf}}, \theta_{\mathbf{ad}})\right). \quad (1)$$

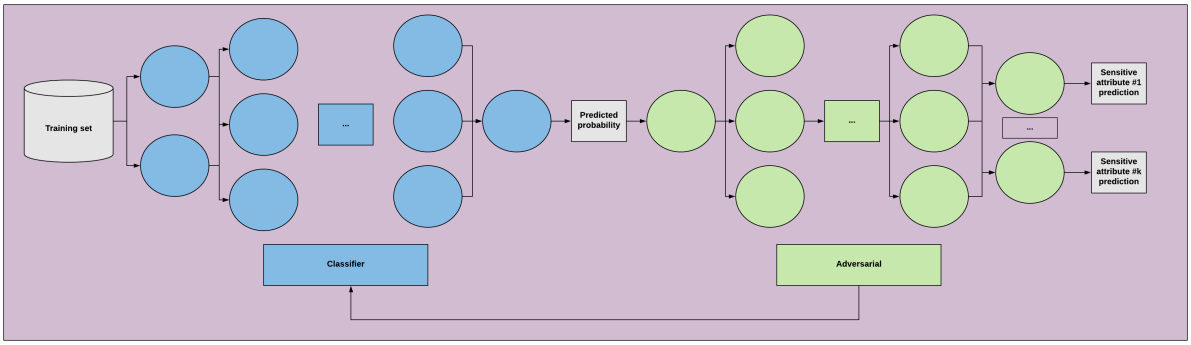

Figure 1: Overview of adversarial network architecture of Louppe *et al.* (2017)

3 Shapley Value

3.1 Definition

Note $v : 2^N \to \Re$ such as $v(\emptyset) = 0$ and N a set of players. If $S \subset N$, $v(S)$ is the amount of wealth produced by S when they cooperate. The Shapley Value (see Shapley, 1953) is a fair share of the global wealth $v(N)$ produced by all players together, among themselves.

$$\Phi_i(N, v) = \sum_{S \subset N \setminus i} \frac{(card(N) - card(S) - 1)!card(S)!}{card(N)!} \left(v(S \cup i) - v(S)\right).$$

The Shapley value is the only indicator that respects the four following properties:

- Additivity: $\Phi_i(N, v + w) = \Phi_i(N, v) + \Phi_i(N, w)$ for all i;
- Null player: if $v(S \cup i) = v(S)$ for all $S \subset N \setminus i$ then $\Phi_i(N, v) = 0$;
- Symmetry: $\Phi_{\pi i}(\pi N, \pi v) = \Phi_i(N, v)$ for every permutation π on N;
- Efficiency: $\sum_{i \in N} \Phi_i(N, v) = v(N)$.

In Machine Learning, Shapley Value can be expressed according the Equation (2).

$$\Phi_i = \sum_{S \subset F \setminus i} \frac{(card(F) - card(S) - 1)!card(S)!}{card(F)!} \left(f_{S \cup i}(\mathbf{x_{S \cup i}}) - f_S(\mathbf{x_S})\right), \tag{2}$$

where $f_S(x_S) = \int \hat{f}(x_1, \ldots, x_p) dP_{x \notin S} - E_X\left(\hat{f}(X)\right)$, F the features set, S a subset of features such as $S \subset F$, $v(S) = f_S(x_S) = E(\hat{f}(\mathbf{x}|\mathbf{x_S}) = \mathbf{x_S^*}$, with $\mathbf{x_S^*}$ real value of the features associate at the instance explained on features subset S.

3.2 Approximation

The exact computation of the Shapley Value involves $O(2^p)$ calculations, which is very quickly untractable to do. Some authors rewrite the Shapley Value problem as a weighted least square optimization problem (see e.g. Lundberg and Lee, 2017; Aas *et al.*, 2019).

Some authors use Monte Carlo techniques to approximate the Shapley Value (see e.g. Strumbelj and Kononenko, 2014; Maleki *et al.*, 2014). Maleki *et al.*, 2014 proves the good performance in terms of errors of Monte Carlo estimation of Shapley Value when the variance or the range of the players' marginal contributions is known.

3.3 Adaptation to a reference population

In classical application of Shapley Value, the reference population is the average prediction and we measure the contribution of each feature to the difference between this prediction and the prediction made for the instance. Merrick and Taly, 2019 propose a generalization of this definition. In the game proposed, the amount of wealth produced by S when they cooperate is written by:

$$v_{\mathbf{x},D^{ref}}(S) = E_{\mathbf{R}\sim D^{ref}}\left(f(\mathbf{z}(\mathbf{x},\mathbf{R},S))\right) - E_{\mathbf{R}\sim D^{ref}}\left(f(\mathbf{R})\right),$$

where $\mathbf{z} : (\mathbf{x},\mathbf{R},S) \mapsto (z_1,\ldots,z_p)$ with $z_i = x_i \times \mathbb{1}_{i\in\mathbb{S}} + r_i \times \mathbb{1}_{i\notin\mathbb{S}}$ for all $i \in (1,\ldots,p)$, D^{ref} a sampling distribution (e.g. uniform on the range of the features or a sampling from features marginal distribution).

4 Application

During this talk, after introducing the fair adversarial network and the Shapley Value, we will propose an illustration of the use of the Fair Adversarial Network on one use case which could be on COMPAS dataset, a dataset of justice predictive where it exists some known bias. We will compare the performance both in term of accuracy and in term of Fairness. For Fairness evaluation, in addition to the p-rule and the Differential Fairness, we will propose base rate metric based on the difference between two distributions which that will solve the threshold problem of the p-rule. Assume we consider a case of binary classification with $\hat{Y} = 1$ or 0 where a protected attribute A takes two modality: a_1, the discriminated modality, and a_2. We can compute kernel estimator of the densities of $P(\hat{Y} = 1|A = a_1)$ and $P(\hat{Y} = 1|A = a_2)$, and then compute the Kullback Leibler (KL) divergence between the two estimated distributions. We compute the max between of the two KL divergence when $A = a_1$ and $A = a_2$ are respectively the reference population. Another metric consists to compute the Kolmogorov-Smirnov statistic between the two empirical distributions. Last metric we propose is to compute the Dynamic Time Warping distance between $(p_{(1)},\ldots,p_{(n_1)})$ and $(q_{(1)},\ldots,q_{(n_2)})$, where for all $i \in \{1,\ldots,n_1\}$, $p_{(i)}$ (resp. for all $i \in \{1,\ldots,n_2\}$, $q_{(i)}$) is the statistic of order i of the probability predicted

by the classifier for the instance such as $A = a_1$ (resp. the probability predicted by the classifier for the instance such as $A = a_2$), with n_1 (resp. n_2) the cardinal of the instances which have a_1 modality (resp. a_2 modality). Thanks to the Shapley Value, we will explain the difference between a classical classifier, which has no Fairness constraint, and the Fair classifier. Moreover we will illustrate the contribution in terms of data protection made possible by the Fair Adversarial Network.

5 Acknowledgement

This work is funded under the SPARTA project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892.

Reference

Aas, Jullum, and Loland (2019), Explaining individual predictions when features are dependent: more accurate approximations to shapley values
Baratt (2017), InterpNET: Neural Introspection for Interpretable Deep Learning
Foulds and Pan (2018), An Intersectional Definition of Fairness
Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, Roth (2018), A comparative study of fairness-enhancing interventions in machine learning, Proceedings of the Conference on Fairness, Accountability, and Transparency
Kamiran, Mansha, Karim and Zhang (2018), Exploiting reject option in classification for social discrimination control, Information Sciences
Louizos, Swersky, Li, Welling, Zemel (2015), Variational Fair Autoencoder, ICLR 2016
Louppe, Kagan, and Cranmer (2017), Learning to Pivot with Adversarial Networks, Advances in Neural Information Processing Systems 30
Lundberg and Lee (2017), A Unified Approach to Interpreting Model, NIPS 2017
Maleki, Tran-Thanh, Hines, Rahwan, and Rogers (2014). Bounding the Estimation Error of Sampling-based Shapley Value Approximation
Merrick and Taly (2019), The explanation Game: Explaining Machine Learning Models with cooperative Game Theory
Raff, Sylvester, Mills (2017), Fair Forests: Regularized Tree Induction to Minimize Model Bias, AAAI / ACM conference on AIES 2018
Ribeiro, Singh, Guestrin (2016), Local Interpretable Model-Agnostic Explanations
Shapley (1953), A Value for n-person Games. Contributions to the Theory of Games.
Strumbelj and Kononenko (2014), Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems
Zafar, Valera, Gomez Rodriguez, Gummadi (2015), Fairness Constraints: Mechanisms for Fair Classification, Proceedings of the 20th IAISTATS