Explicabilité et *loyauté* des règles de décision *boites-noires* en Intelligence Artificielle

Laurent Risser
Ingénieur de Recherche
CNRS - IMT - 3IA ANITI

Ronan Pons
Doctorant
Université Toulouse Capitole - 3IA ANITI

lrisser@math.univ-toulouse.fr

Décisions ou prédictions automatiques en utilisant de **règles apprises sur un jeu d'apprentissage**

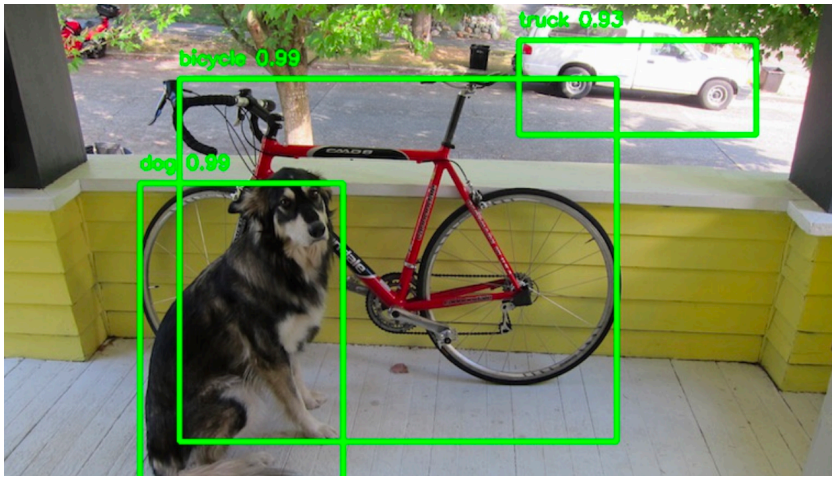
Modèles utilisant **des règles simples**
- Systèmes experts
- Modèles linéaires
- Arbres de décisions
- Perceptron

Modèles aux règles de décisions **peu à très peu interprétables**
- SVM à noyau
- Forets aléatoires

Modèles aux règles **encore moins interprétables**
- Réseaux de Neurones profonds

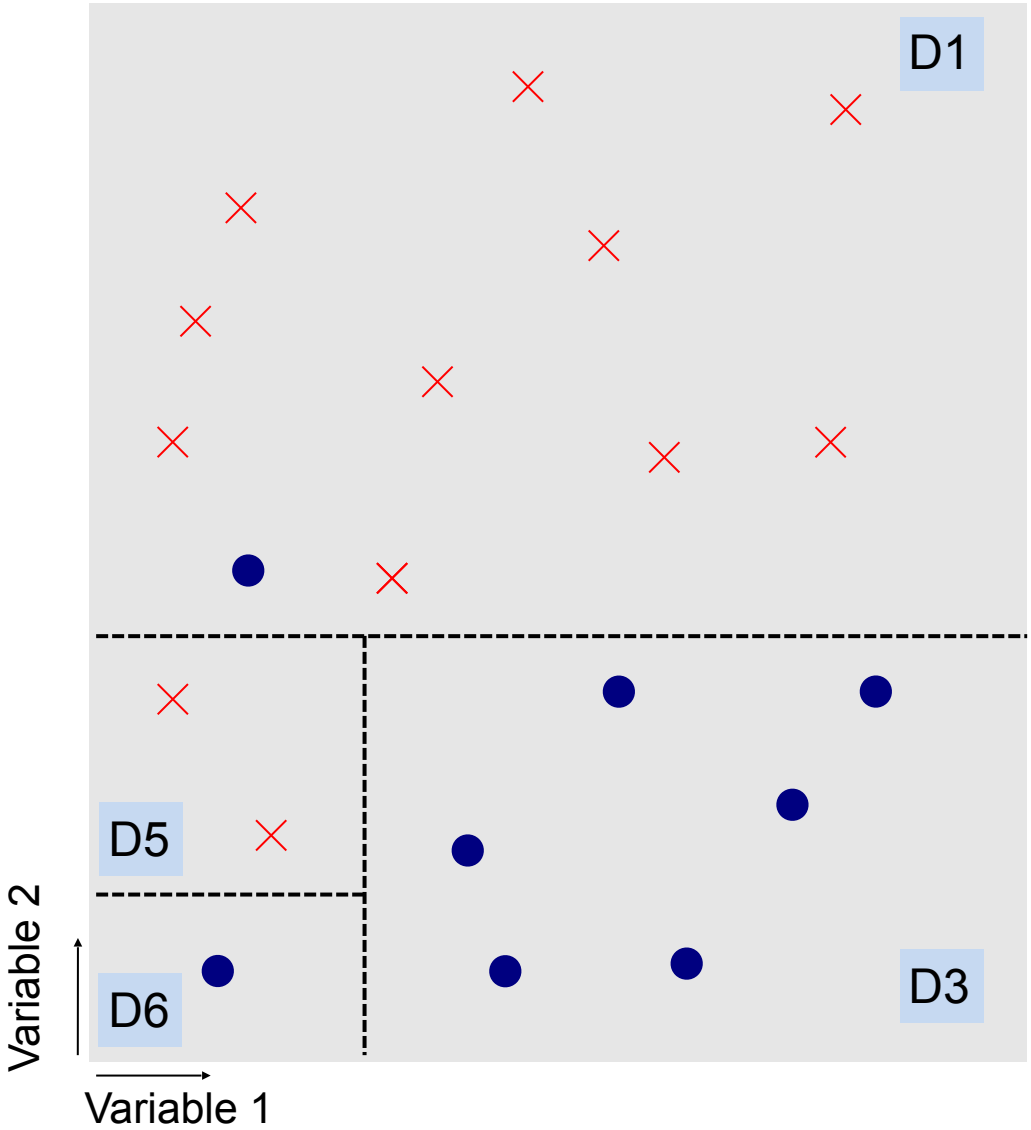
Années

Début des années 1980 : Aide au pilotage à l'aide de systèmes experts



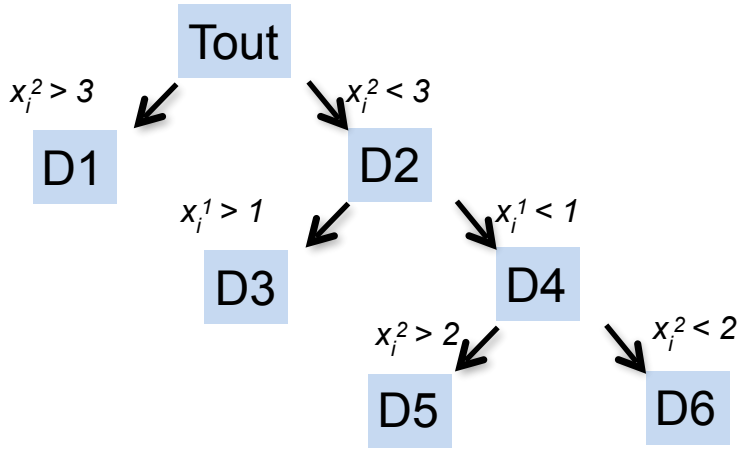
Fin des années 2010 : Détection temps réel de plus de 1000 objets par CNN dans des vidéos à 24 fps (Yolo v3).

Interprétabilité — Arbres de décisions

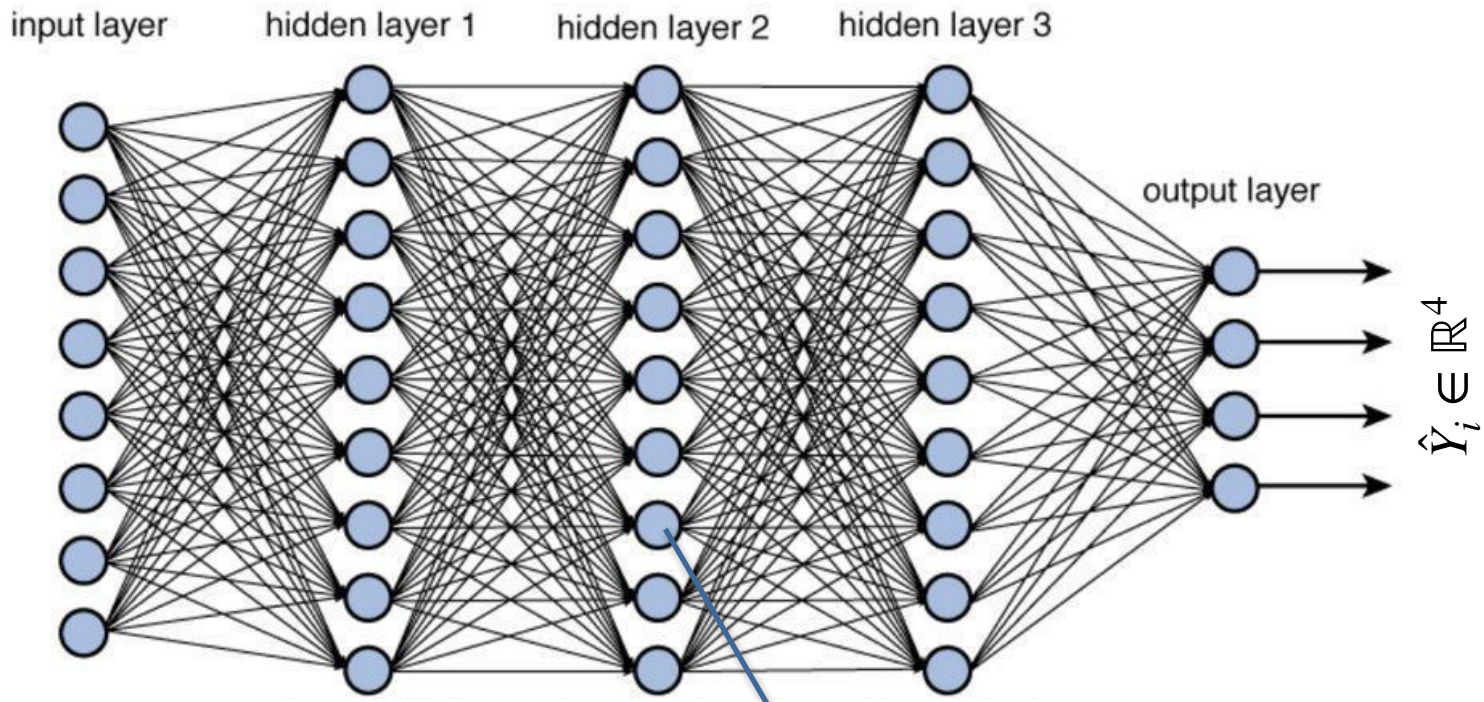


$\boldsymbol{x_i} = (x_i^1, x_i^2)$ une observation d'entrée

y_i un labels. *(ici : × → 1 et ● → -1)*

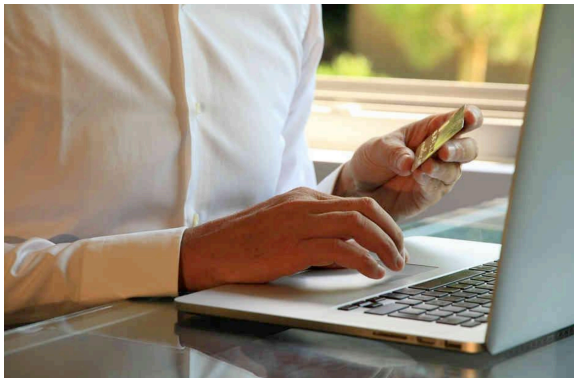


Interprétabilité — Réseaux de neurones



Réseau « *pas très profond* » prenant des décisions à partir de 270 paramètres $\theta = \{\theta_{pk}^l\}_{p,k,l}$

Vision actuelle de l'I.A. $\approx$ Applications de *l'apprentissage supervisé* avec
- des **règles de décisions complexes**
- des **bases d'apprentissage massives**



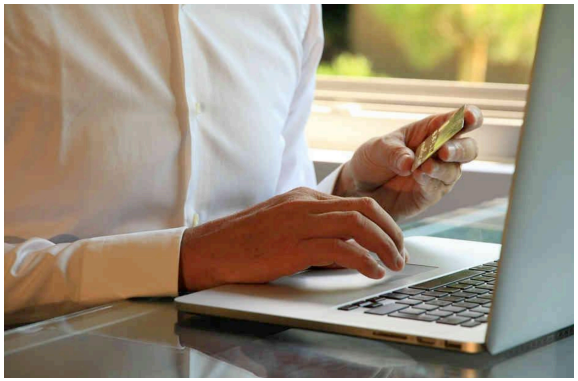
Publicité en ligne

Aide au diagnostic

Véhicules autonomes

...

Vision actuelle de l'I.A. $\approx$ Applications de *l'apprentissage supervisé* avec
- des **règles de décisions complexes**
- des **bases d'apprentissage massives**



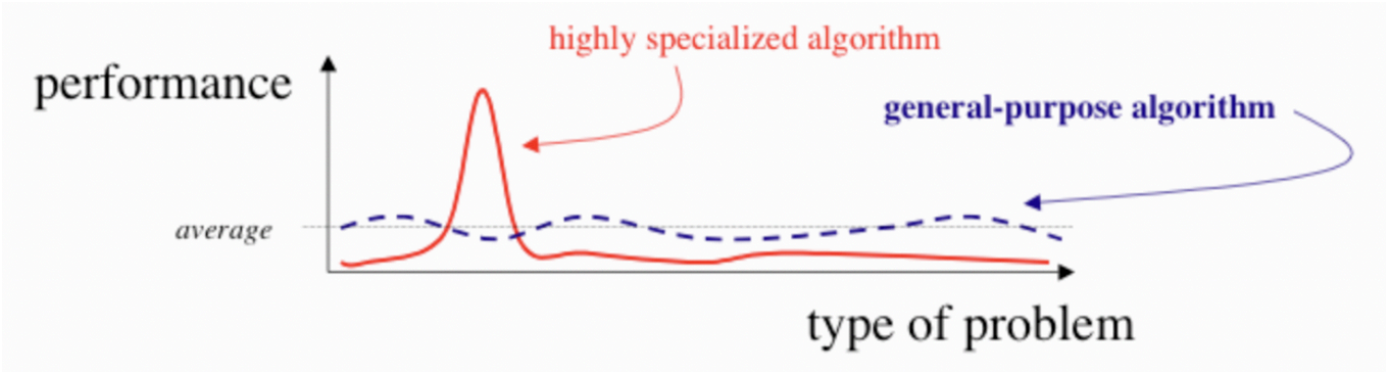
Publicité en ligne

Aide au diagnostic

Véhicules autonomes

...

Mais… problème (classique) du biais !

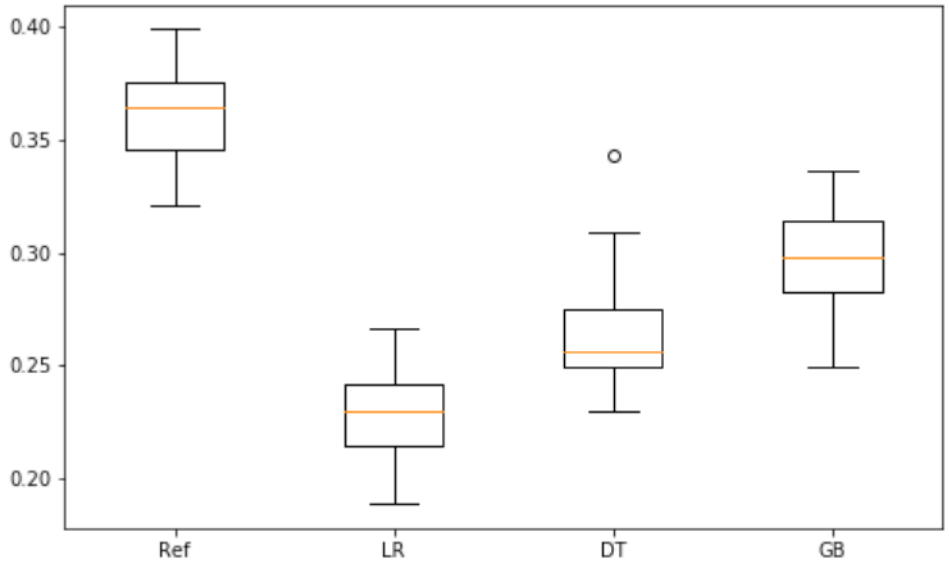


Depiction of the No Free Lunch theorem, where better performance at a certain type of problem comes with the loss of generality (Fedden, 2017)

Pire… le biais peut être déjà contenu dans les données d'apprentissage !

Exemple : Accord ou refus d'un prêt en fonction du dossier client (https://archive.ics.uci.edu/ml/datasets/adult)

Déséquilibres dans la base d'apprentissage



Disparate Impact (DI) de prédictions
(**ici** $acc \in [0.85,0.87]$)

$$DI = \frac{\mathbb{P}(\hat{Y} = \textbf{granted} \,|\, S = \textbf{Female})}{\mathbb{P}(\hat{Y} = \textbf{granted} \,|\, S = \textbf{Male})}$$

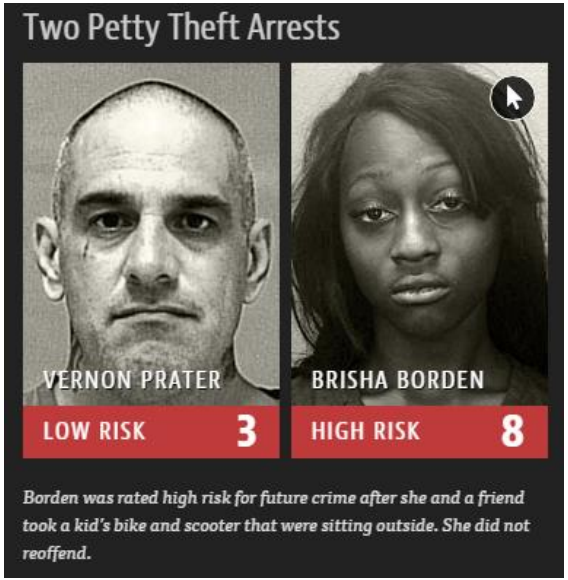
P. Besse, E. Del Barrio, P. Gordaliza, J.M. Loubes, L. Risser: *A story of bias in Machine Learning* (submitted)

Besoins forts et urgents de :

- **Cadre légal** pour profiter des avancés de l'I.A. tout en limitant ses dérives.
- Outils qui **détectent** un problème potentiel.
- Outils qui **interprètent** des règles de décisions complexes afin de les **corriger**.

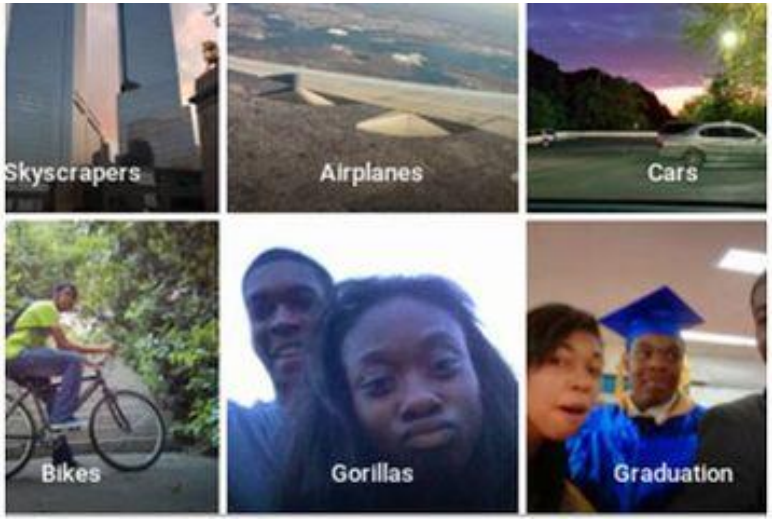
Risque juridique du *machine learning* : la discrimination algorithmique

COMPAS



Source : *Propublica*

GOOGLE PHOTOS



Source : *The Verge*

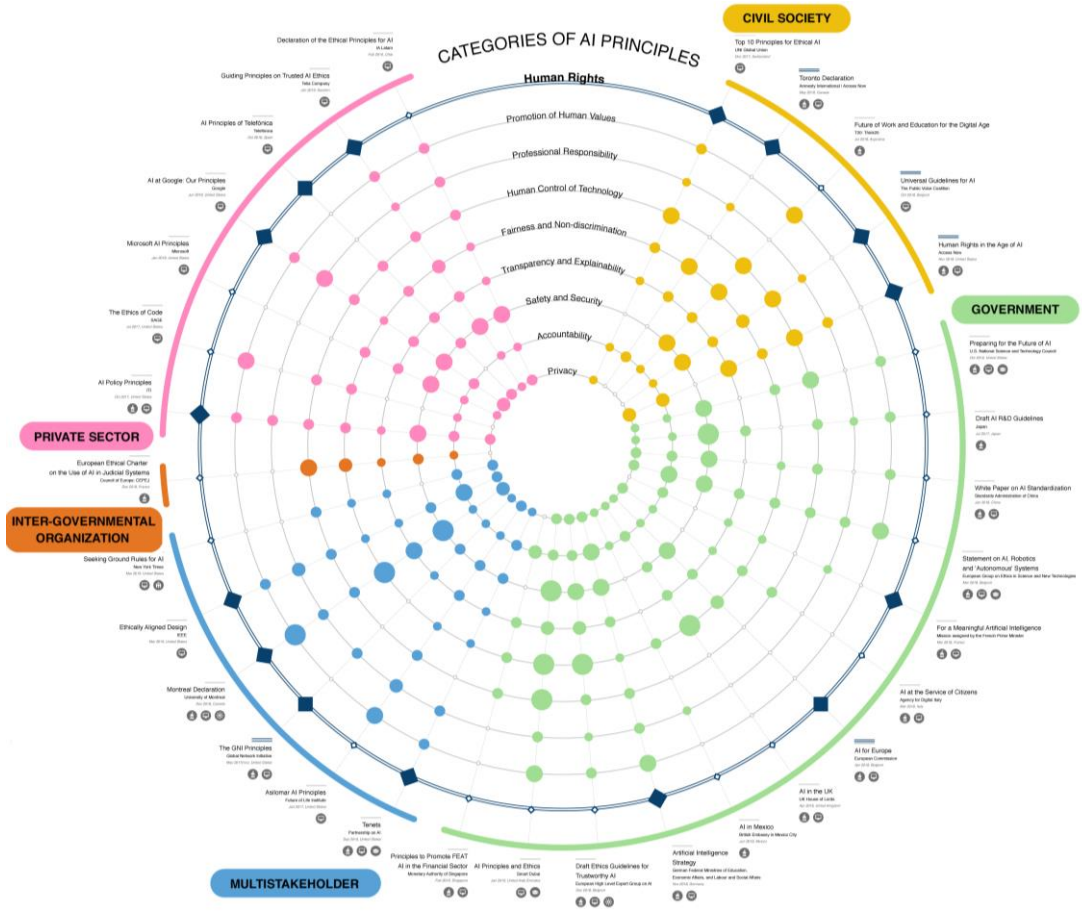
STREET BUMP

Source : *Boston.gov*

Les principes éthiques de l'IA

<u>9 principes éthiques :</u>

- Respect des droits de l'Homme ;
- Protection des valeurs humaines ;
- Responsabilité professionnelle ;
- Contrôle humain ;
- Equité et non-discrimination ;
- Transparence et Explicabilité ;
- Sureté et sécurité ;
- Reddition de comptes ;
- Vie privée ;



PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches July 4, 2019

Authors: Jessica Fjeld, Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Joshua Feldman, Sally Kagay
Design: Arushi Singh (arushisingh.net)

La discrimination directe

« Constitue une ***discrimination*** *toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, [...] de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle ,de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leurs capacités à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée.*

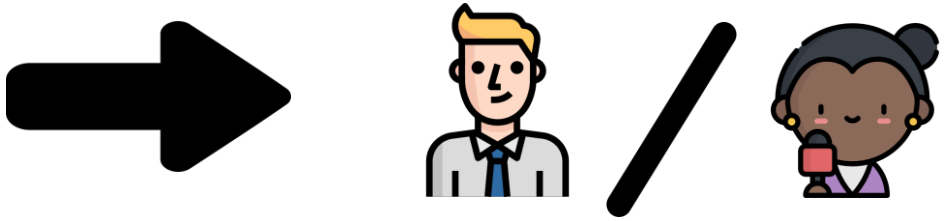
Article 225-1 du code pénal

La discrimination indirecte

« Constitue une ***discrimination indirecte*** *une disposition, un* ***critère ou une pratique neutre en apparence****, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un* ***désavantage particulier pour des personnes par rapport à d'autres personnes****, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifiée par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés. »*

Article 1 de la loi n°2008-496 du 27 mai 2008

Critère « neutre »



Conséquences discriminantes

Loi Informatique et Libertés (LIL) : l'interdiction des décisions automatisées

*« Aucune décision produisant des **effets juridiques** à l'égard d'une personne ou l'affectant de **manière significative** ne peut être prise **sur le seul fondement** d'un **traitement automatisé** de **données à caractère personnel**, y compris le profilage »*

Article 47 LIL

Secteurs concernés

Public

Privé

Gravité de la décision pour l'individu

Absence d'intervention humaine

Traitement de données personnelles

Règlement Général à la Protection des Données personnelles (RGPD) : Obligation d'information et Droit d'accès

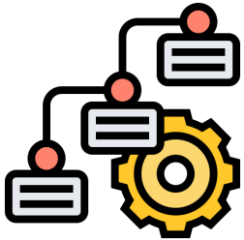
« [...] ***l'existence*** *d'une prise de décision automatisée, [...] et, des* ***informations utiles*** *concernant la* ***logique sous-jacente,*** *ainsi que* ***l'importance et les conséquences prévues de ce traitement pour la personne concernée.*** *»*

Articles 13.2.f), 14.2.g) et 15.1.h)

Moment de la communication

Acteurs concernés

Contenu de l'information



Logique de l'algorithme

Conséquences pour l'individu

Loi pour une République Numérique (LRN) : le cas des administrations publiques

Obligation de publication :

> *« [...] les administrations [...]* ***publient*** *en ligne les* ***règles définissant les principaux traitements algorithmiques*** *utilisés dans l'accomplissement de leurs missions* ***lorsqu'ils fondent des décisions individuelles.*** *»*
>
> Article L. 312-1-3 CRPA

Obligation de communication à l'intéressé :

> *« [...] une décision prise sur le fondement d'un traitement d'un traitement algorithmique comporte une* ***mention explicite*** *en informant l'intéressé.*
>
> *Les* ***règles*** *définissant ce traitement ainsi que les* ***principales caractéristiques*** *de sa mise en œuvre sont* ***communiquées à l'intéressé s'il en fait la demande.*** *»*
>
> Article L. 311-3-1 CRPA

Secteurs

Administrations

Décisions

Toutes décisions automatisées

LRN (2/2) : les détails des informations à fournir

Communiquées *« sous une forme **intelligible** » :*

1. *Le **degré et le mode de contribution** du traitement algorithmique à la prise de décision ;*
2. *Les **données traitées** et leurs **sources ;***
3. *Les **paramètres de traitement** et, le cas échéant, leur **pondération, appliqués à la situation de l'intéressé ;***
4. *Les **opérations effectuées** par le traitement »*

Article R. 311-3-1-2 CRPA

→ Relation humain / machine au cours du processus de décision

→ Données entrées (*inputs*) et leurs origines

→ Variables en jeu et leur poids respectifs

→ Étapes du processus

Les projets à venir : l'exemple de la loi bioéthique

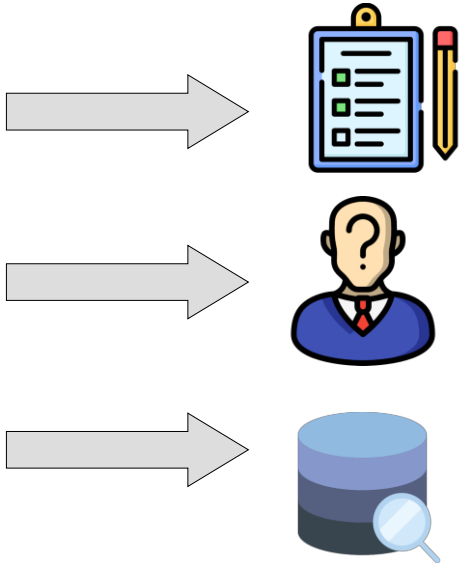
Article 11 du projet de loi:

Préserver une garantie humaine dans les résultats :

1) Bonne information du patient lorsqu'un traitement par IA est utilisé à l'occasion d'un acte de soins

2) Garantie d'une intervention humaine

3) Traçabilité des actions et des données utilisées pour le traitement



- Quelles informations ?
- Quel format ?

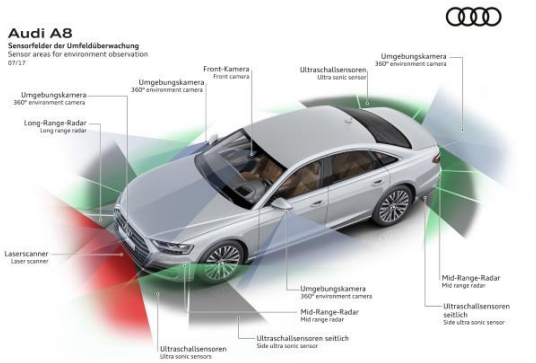
- Qui est-ce ?
- Quel rôle joue-t-il ?
- Que garantie-t-il ?

- Comment ?
- Quel objectif ?
- Quelle protection ?

Et à l'international ?

En Allemagne :

German Road Traffic Act Amendment Regulating the Use for ***« Motors Vehicles with Highly or Fully Automated Driving Function »*** *- July 17, 2017*



Source : *Audi*

Aux États-Unis :

Projet de loi fédérale « *Algorithm Accountability Act* » dont l'objectif est de lutter contre la discrimination algorithmique

Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms

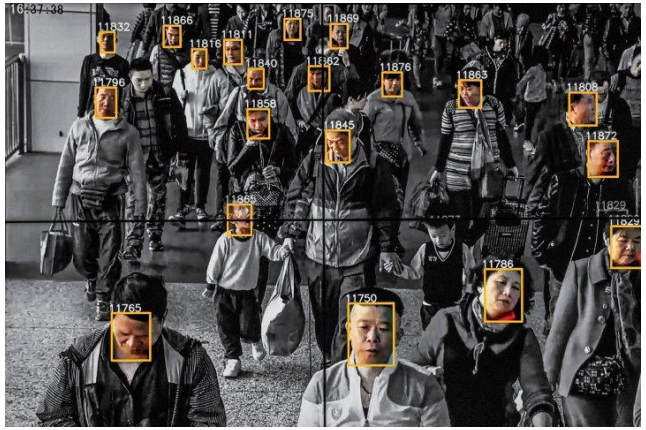
April 10, 2019

Share Tweet Print

WASHINGTON, D.C. -- U.S. Senators Cory Booker (D-NJ) and Ron Wyden (D-OR), along with Rep. Yvette D. Clarke (D-NY) today introduced the Algorithmic Accountability Act, which requires companies to study and fix flawed computer algorithms that result in inaccurate, unfair, biased or discriminatory decisions impacting Americans.

Source : *booker.senate.gov*

En Chine :



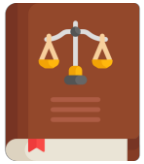
How China Is Using "Social Credit Scores" to Reward and Punish Its Citizens

By Charlie Campbell / Chengdu

Yi Tingyue has a lot going for her. She won a scholarship to China's prestigious Sichuan University, where she graduated with a master's in graphic design. She drives an Audi A4 and owns a penthouse

Source : *Time magazine*

Du juridique à la technique



Etat de l’art du droit

Recommandations au législateur d’un cadre juridique applicable

Développement d’outils techniques conformes aux futures dispositions légales

ET MAINTENANT ???

On arrête d'utiliser des boites noires en I.A. ?

Tests pour détecter des biais de discrimination.

Méthodes d'interprétabilité de règles de décision *boites noires*.

Méthodes de corrections des règles biaisées.

Recherche industrielle - Interprétabilité et loyauté

Google AI

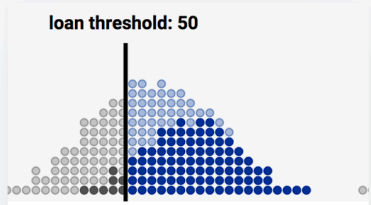
About Stories Research Education Tools Responsibilities Blog

Philosophy Publications People Teams & Focus Areas Join Us Outreach

RESEARCH › TEAMS & FOCUS AREAS › BRAIN TEAM ›

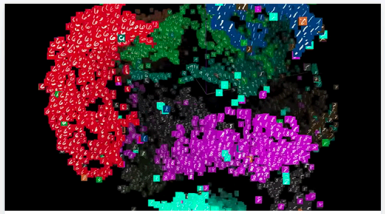
People + AI Research (PAIR)

Human-centered research and design to make AI partnerships productive, enjoyable, and fair.



Explaining fairness

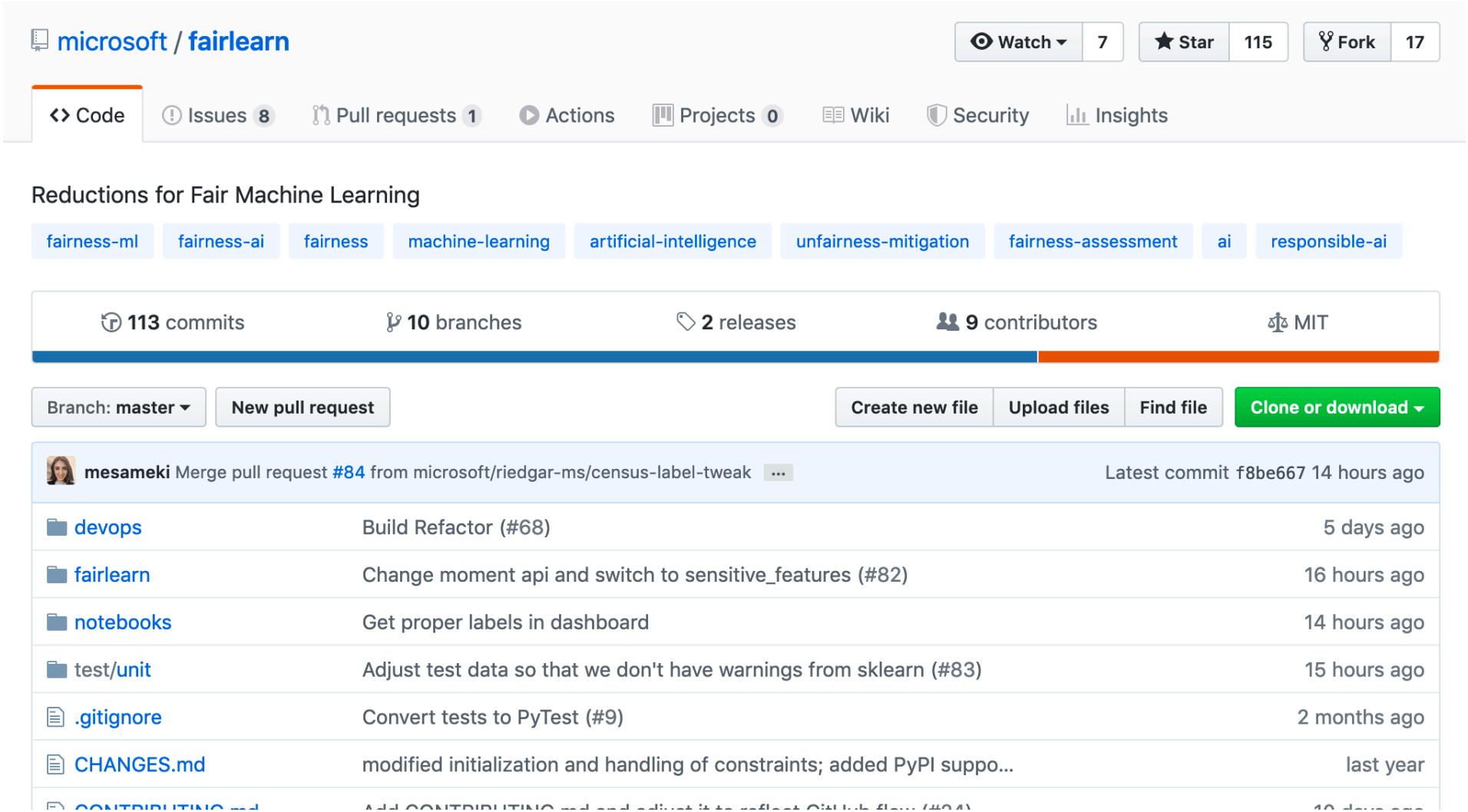
Policy tradeoffs in machine learning can be complex, but visualizations and interactive explanations can help people understand these critical issues.



Model interpretibliity

Can machines explain the "why" behind their decisions? We're investigating ways for people to understand more about ML models, starting with visualizations that look under the hood of complex systems. See...

Recherche industrielle - Loyauté



microsoft / fairlearn

Watch 7 | Star 115 | Fork 17

Code | Issues 8 | Pull requests 1 | Actions | Projects 0 | Wiki | Security | Insights

Reductions for Fair Machine Learning

fairness-ml fairness-ai fairness machine-learning artificial-intelligence unfairness-mitigation fairness-assessment ai responsible-ai

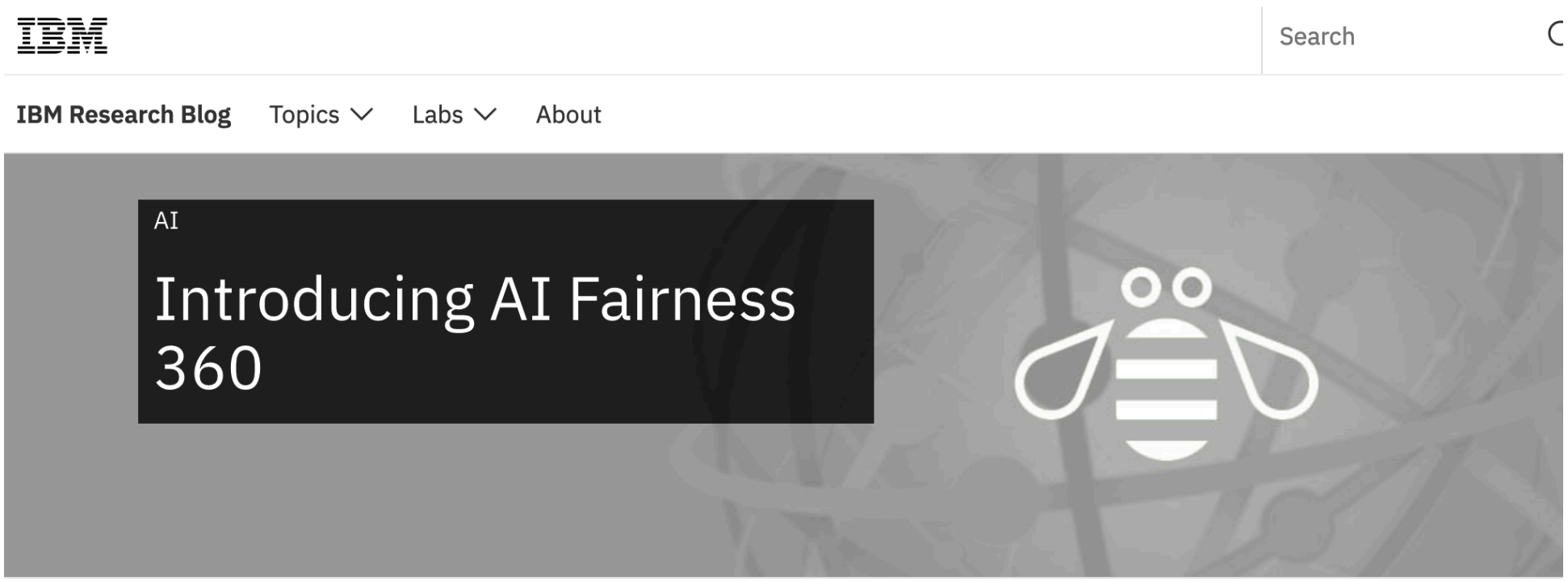
113 commits | 10 branches | 2 releases | 9 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

mesameki Merge pull request #84 from microsoft/riedgar-ms/census-label-tweak ... Latest commit f8be667 14 hours ago

devops	Build Refactor (#68)	5 days ago
fairlearn	Change moment api and switch to sensitive_features (#82)	16 hours ago
notebooks	Get proper labels in dashboard	14 hours ago
test/unit	Adjust test data so that we don't have warnings from sklearn (#83)	15 hours ago
.gitignore	Convert tests to PyTest (#9)	2 months ago
CHANGES.md	modified initialization and handling of constraints; added PyPI suppo...	last year

Recherche industrielle - Loyauté



September 19, 2018 | Written by: Kush R. Varshney

Categorized: AI

Share this post:

Recherche académique - Interprétabilité

marcotcr / lime

Watch 276 | Star 7.2k | Fork 1.2k

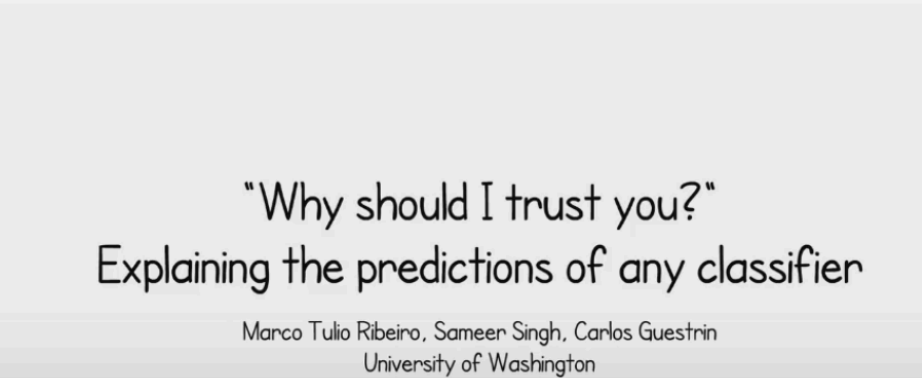
Code | Issues 30 | Pull requests 10 | Actions | Projects 0 | Security | Insights

Lime: Explaining the predictions of any machine learning classifier

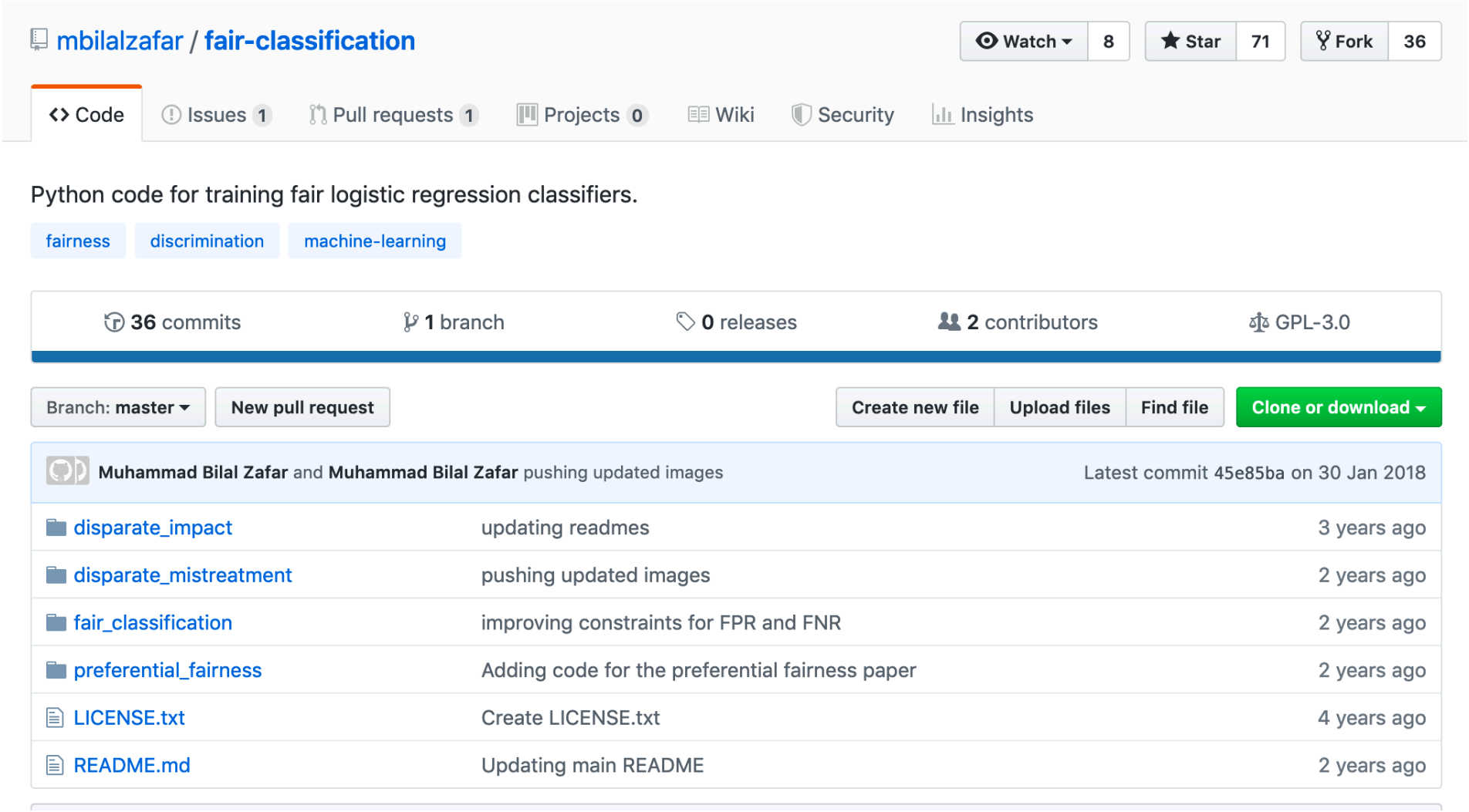
446 commits | 5 branches | 0 packages | 17 releases | 37 contributors | BSD-2-Clause

Branch: master | New pull request | Find file | Clone or download

This project is about explaining what machine learning classifiers (or models) are doing. At the moment, we support explaining individual predictions for text classifiers or classifiers that act on tables (numpy arrays of numerical or categorical data) or images, with a package called lime (short for local interpretable model-agnostic explanations). Lime is based on the work presented in this paper (bibtex here for citation). Here is a link to the promo video:



Recherche académique - Loyauté



mbilalzafar / **fair-classification**

Watch 8 | Star 71 | Fork 36

Code | Issues 1 | Pull requests 1 | Projects 0 | Wiki | Security | Insights

Python code for training fair logistic regression classifiers.

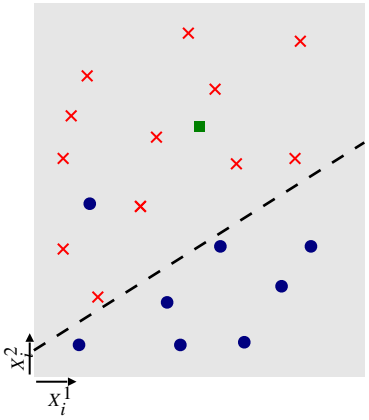
fairness | discrimination | machine-learning

36 commits | **1** branch | **0** releases | **2** contributors | GPL-3.0

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

Muhammad Bilal Zafar and **Muhammad Bilal Zafar** pushing updated images — Latest commit 45e85ba on 30 Jan 2018

disparate_impact	updating readmes	3 years ago
disparate_mistreatment	pushing updated images	2 years ago
fair_classification	improving constraints for FPR and FNR	2 years ago
preferential_fairness	Adding code for the preferential fairness paper	2 years ago
LICENSE.txt	Create LICENSE.txt	4 years ago
README.md	Updating main README	2 years ago

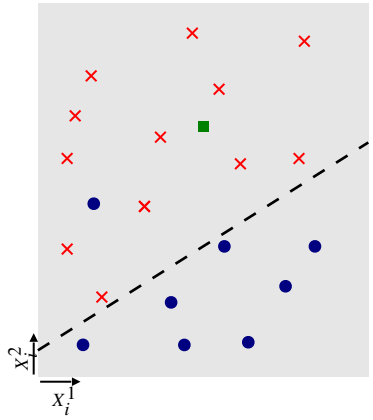


- Apprentissage effectué
- Base de test :
  - $X_i \in \mathbb{R}^p \rightarrow$ Donnée d'entrée test (ex : revenu, niveau d'épargne,…)
  - $S_i \in \{0,1\} \rightarrow$ Variable sensible (ex : age, sexe,…)
  - $\hat{Y}_i \in \{0,1\} \rightarrow$ Décision (ex : obtention d'un prêt)

Disparate impact

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

Test statistique dans : Besse, Del Barrio, Gordaliza, Loubes (2018) — https://arxiv.org/pdf/1807.06362.pdf



- Apprentissage effectué
- Base de test :
  - $X_i \in \mathbb{R}^p \rightarrow$ Donnée d'entrée test (ex : revenu, niveau d'épargne,…)
  - $S_i \in \{0,1\} \rightarrow$ Variable sensible (ex : age, sexe,…)
  - $\hat{Y}_i \in \{0,1\} \rightarrow$ Décision (ex : obtention d'un prêt)

Disparate impact

$$DI(X, \hat{Y}, S) = \frac{\mathbb{P}(\hat{Y} = 1 \,|\, S = 0)}{\mathbb{P}(\hat{Y} = 1 \,|\, S = 1)}$$

Test statistique dans : Besse, Del Barrio, Gordaliza, Loubes (2018) — https://arxiv.org/pdf/1807.06362.pdf

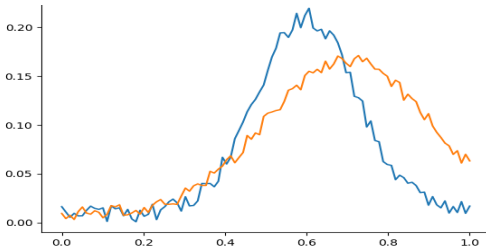
Equality of Opportunity Hardt, Price, Srebro (2016) — https://arxiv.org/pdf/1610.02413.pdf

$$OP_0(\hat{Y}, Y, S) = \mathbb{P}(\hat{Y} = 1 \,|\, S = 0, Y = 1) \quad \text{et} \quad OP_1(\hat{Y}, Y, S) = \mathbb{P}(\hat{Y} = 1 \,|\, S = 1, Y = 1)$$

avec $Y_i \in \{0,1\} \rightarrow$ Décision sans biais de discrimination

Distance de Wasserstein entre les distributions $\{\mu_{\theta,0}, \mu_{\theta,1}\}$ de Y pour $S = \{0,1\}$

$$W_2^2(\mu_{\theta,0}, \mu_{\theta,1}) = \int_0^1 \left(H_0^{-1}(\tau) - H_1^{-1}(\tau)\right)^2$$

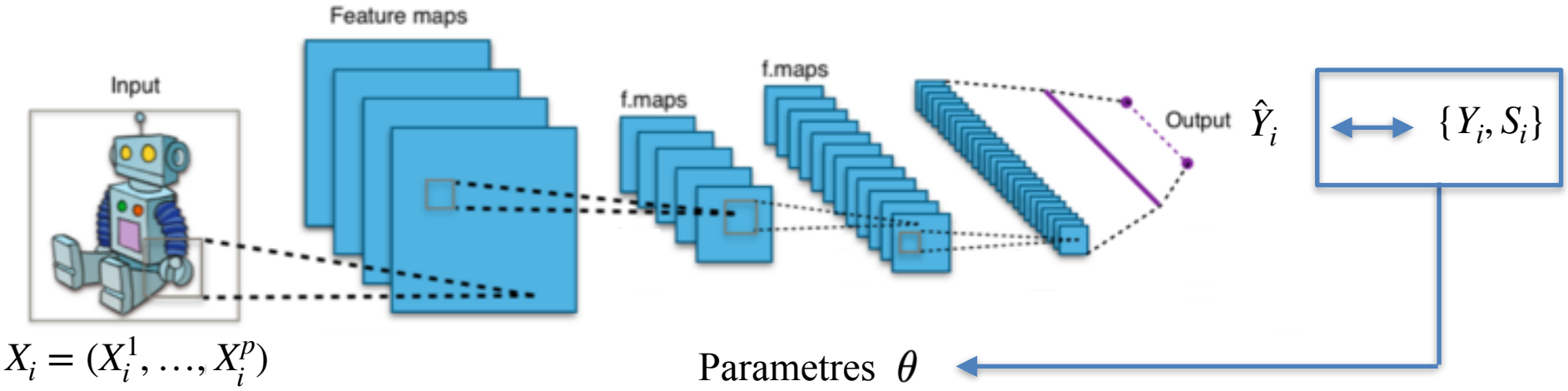


False negative rate, equality of odds …

Méthode la plus efficace pour corriger une règle de décision injuste quel que soit $\{X_i, S_i\}$

… mais qualité de prédiction mauvaise !

***Rétro-propagation* de contraintes de loyauté dans un réseau de Neurones :**

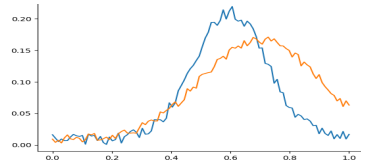


$$\hat{\theta} = \arg\min_{\theta} R(\theta) + \lambda W_2^2(\mu^n_{\theta,0}, \mu^n_{\theta,1}) \qquad \text{où} \qquad W_2^2(\mu^n_{\theta,0}, \mu^n_{\theta,1}) = \int_0^1 \left(H_0^{-1}(\tau) - H_1^{-1}(\tau)\right)^2 d\tau$$

Qualité de prédiction

Contrainte de loyauté

Distance entre les predictions pour lesquelles *S=0* et *S=1*



Risser, Vincenot, Couellan, Loubes (2019) — https://arxiv.org/pdf/1908.05783.pdf

Algorithm 1 Batch training procedure for neural-networks with Wasserstein-2 regularization

ht

Require: Weight λ and the training observations $(X_i, S_i, Y_i)_{i=1,\dots,n}$, where $X_i \in \mathbb{R}^p$, $S_i \in \{0,1\}$ and $y_i \in \{0,1\}$.
Require: Neural network f_θ with initialized parameters θ.

for e in Epochs **do**
  for b in Batches **do**
    Pre-compute H_0 and H_1.
    Draw the batch observations B.
    Compute the $f_\theta(X_i)$, $i \in B$
    Approximate $\mathbb{E}\left[\frac{\partial loss(f_\theta(X), y)}{\partial f_\theta(X)}\right]$ using Eq. (5).
    Approximate $\mathbb{E}\left[\frac{\partial W_2^2(\mu_{\theta,0}, \mu_{\theta,1})}{\partial f_\theta(X)}\right]$ using Eq. (13).
    Backpropagate the approximated derivative of $R(\theta) + \lambda W_2^2(\mu^n_{\theta,0}, \mu^n_{\theta,1})$.
    Update the parameters θ.
  end for
end for
return Trained neural network f_θ.

Algorithme d'apprentissage

	Acc	DI
SLR	0.83	0.28
LRn	0.72	0.44
NNn	0.82	0.34
ZFA	0.79	0.68
ZFN	0.66	1.04
LRW2	0.65	0.97
NNW2	0.78	0.68

Résultats sur données UCI Adult (avec discrimination H/F)

Risser, Vincenot, Couellan, Loubes (2019) — https://arxiv.org/pdf/1908.05783.pdf

Base de test

$\{X_i, Y_i\}_{i=1,\ldots,n}$

avec

$X_i = \{X_i^1, \ldots, X_i^p\}$

Règles de décisions « boites noires »

X_i^1 →
X_i^2 →
X_i^3 →
… …
X_i^p →

→ $\hat{Y}_i := f(X_i)$

Question : Comment mesurer l'influence de chaque X^p sur les sorties $\hat{Y}$

Base de test

$\{X_i, Y_i\}_{i=1,\ldots,n}$

avec

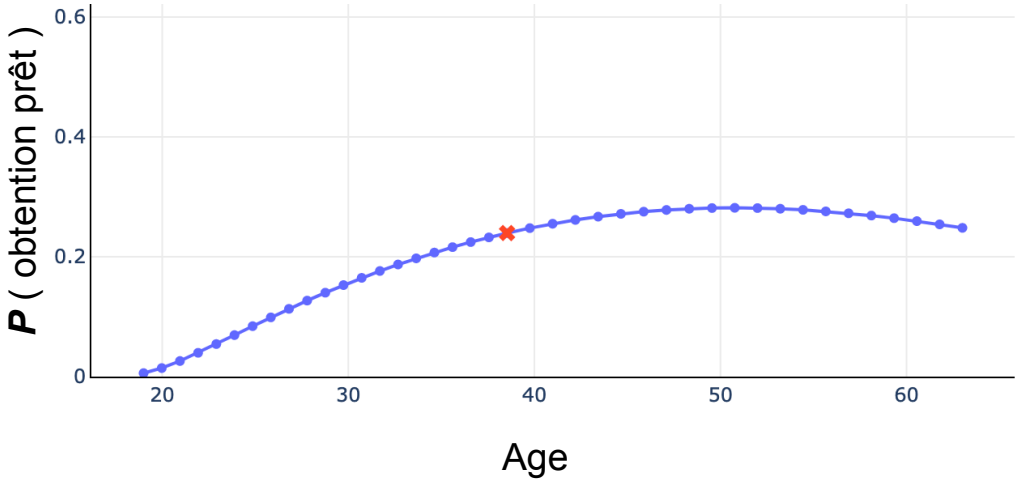
$X_i = \{X_i^1, \ldots, X_i^p\}$

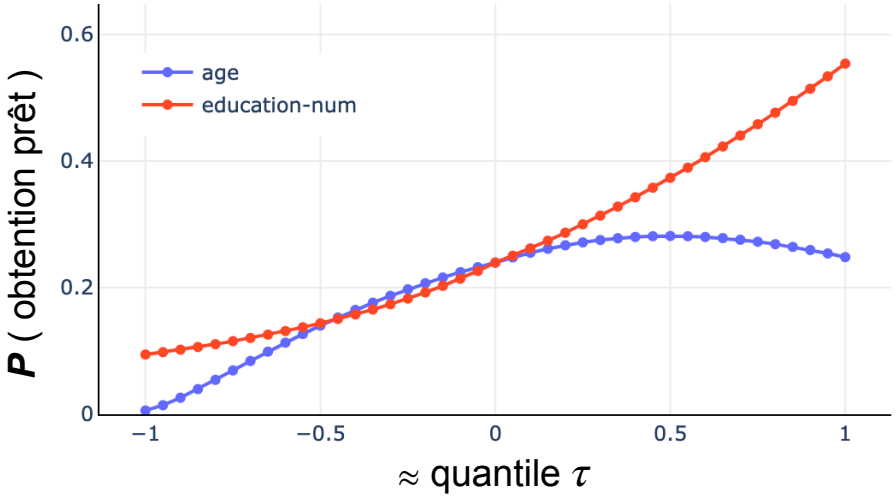
Règles de décisions « boites noires »

X_i^1 →
X_i^2 →
X_i^3 →
… …
X_i^p →

→ $\hat{Y}_i := f(X_i)$

Intuition : Modifier les propriétés moyennes de chaque X_i^p et observer les variations en sortie





Base de test

$\{X_i, Y_i\}_{i=1,\dots,n}$

avec

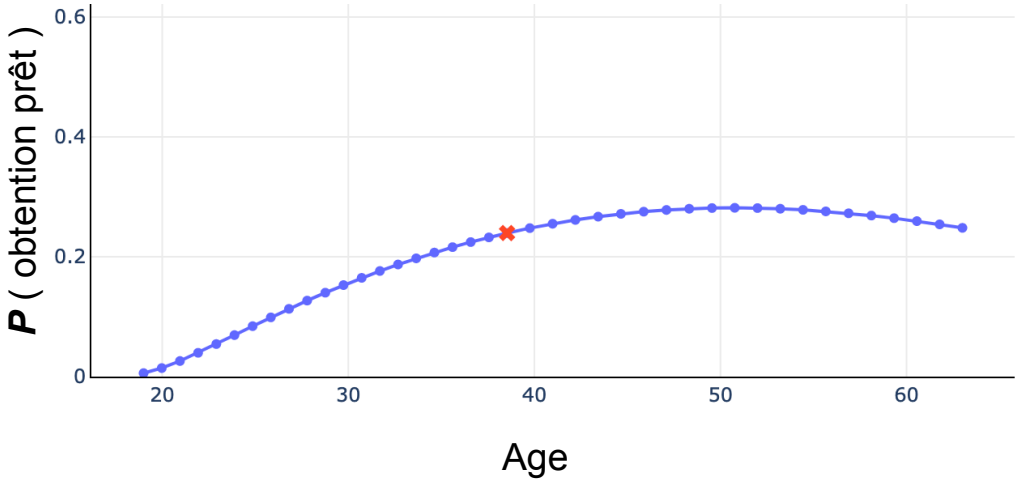
$X_i = \{X_i^1, \dots, X_i^p\}$

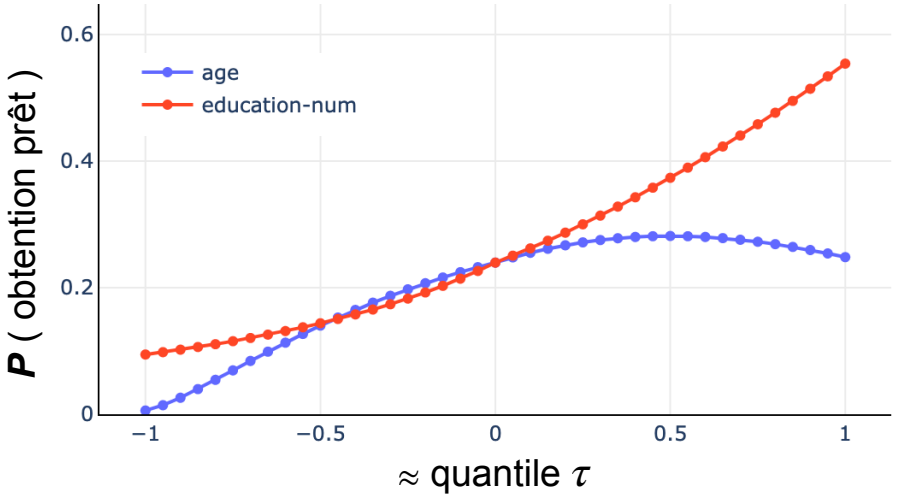
Règles de décisions « boites noires »

X_i^1 → X_i^2 → X_i^3 → … → X_i^p →

→ $\hat{Y}_i := f(X_i)$

Intuition : Modifier les propriétés moyennes de chaque X_i^p et observer les variations en sortie





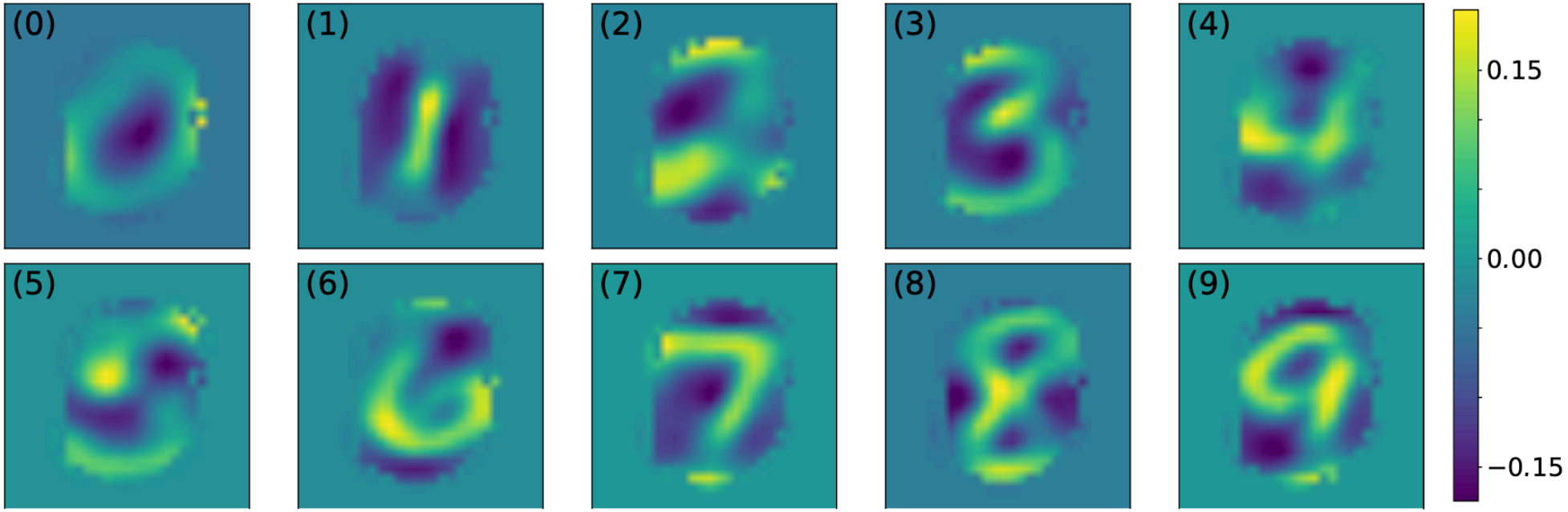
Difficultés :

- Coût algorithmique en grande dimension
- Risque de tester des observations irréalistes

→ Re-pondération automatique des observations avec contrainte d'entropie

Bachoc, Gamboa, Halford, Loubes, Risser (2019) — https://arxiv.org/pdf/1810.07924.pdf

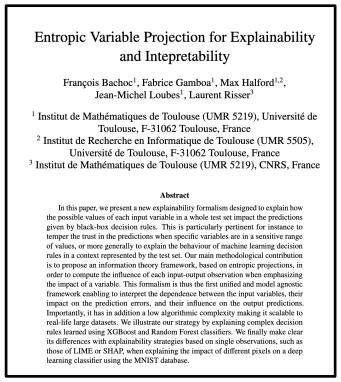
Données test : 10000 images 28x28 de la base MNIST / 10 classes



Détection des pixels les plus influents pour prédire chaque label (10 secondes de calculs en Python)

Bachoc, Gamboa, Halford, Loubes, Risser (2019) — https://arxiv.org/pdf/1810.07924.pdf

Article : https://arxiv.org/pdf/1810.07924.pdf



Entropic Variable Projection for Explainability and Intepretability

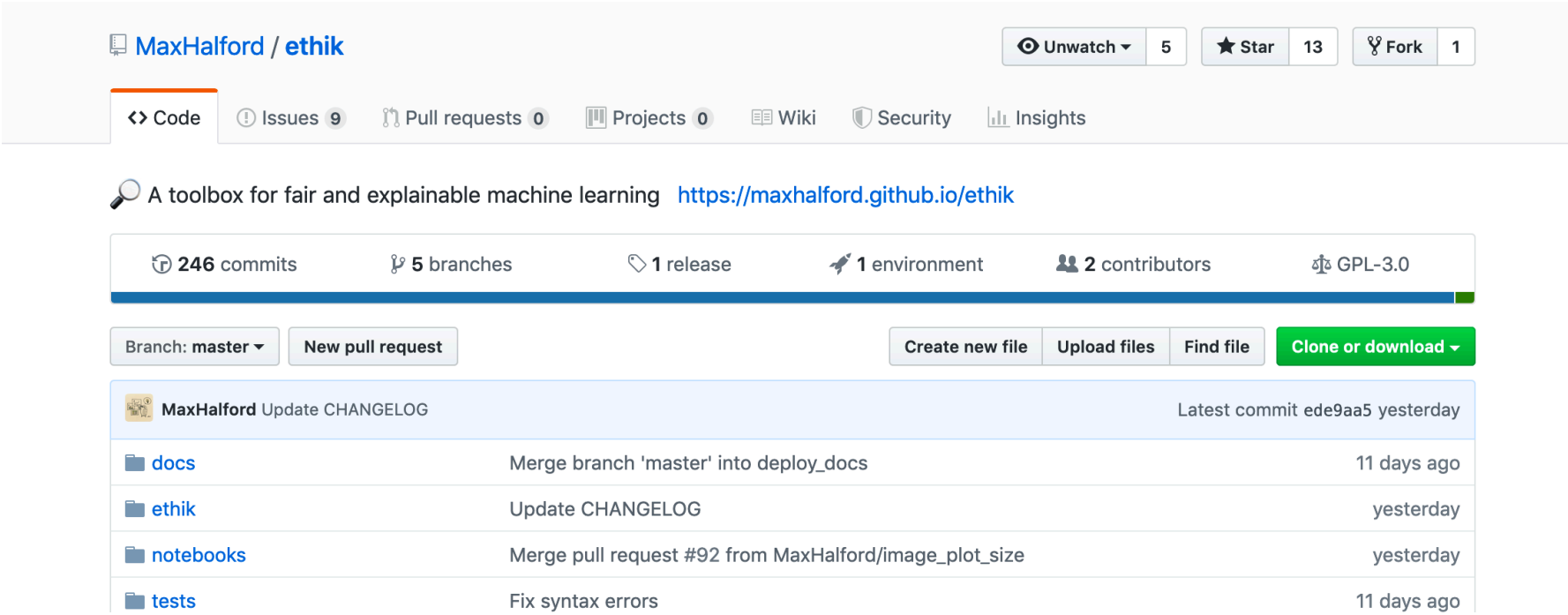
François Bachoc[1], Fabrice Gamboa[1], Max Halford[1,2], Jean-Michel Loubes[1], Laurent Risser[3]

[1] Institut de Mathématiques de Toulouse (UMR 5219), Université de Toulouse, F-31062 Toulouse, France
[2] Institut de Recherche en Informatique de Toulouse (UMR 5505), Université de Toulouse, F-31062 Toulouse, France
[3] Institut de Mathématiques de Toulouse (UMR 5219), CNRS, France

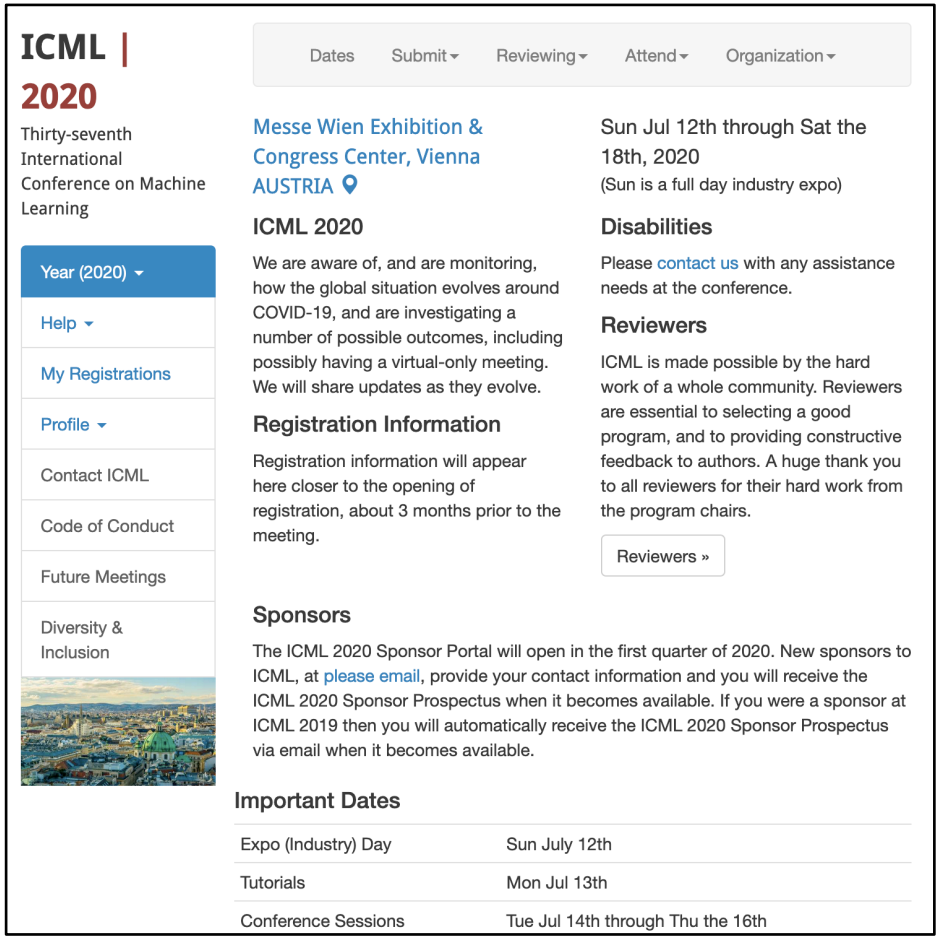
Abstract

In this paper, we present a new explainability formalism designed to explain how the possible values of each input variable in a whole test set impact the predictions given by black-box decision rules. This is particularly pertinent for instance to temper the trust in the predictions when specific variables are in a sensitive range of values, or more generally to explain the behaviour of machine learning decision rules in a context represented by the test set. Our main methodological contribution is to propose an information theory framework, based on entropic projections, in order to compute the influence of each input-output observation when emphasizing the impact of a variable. This formalism is thus the first unified and model agnostic framework enabling to interpret the dependence between the input variables, their impact on the prediction errors, and their influence on the output predictions. Importantly, it has in addition a low algorithmic complexity making it scalable to real-life large datasets. We illustrate our strategy by explaining complex decision rules learned using XGBoost and Random Forest classifiers. We finally make clear its differences with explainability strategies based on single observations, such as those of LIME or SHAP, when explaining the impact of different pixels on a deep learning classifier using the MNIST database.

Package python : https://maxhalford.github.io/ethik



Thématique 8 des JDEV 2020 : Programmez et déployez votre I.A.



ICML workshop : « Law and Machine Learning »

Merci !