Loyauté des Algorithmiques d'Apprentissage Machine / Statistique

Philippe Besse

Université de Toulouse – INSA
Institut de Mathématiques – UMR CNRS 5219
CIMI – Projet AOC

Questions non traitées

- *La technologie elle elle neutre ?* : qui y a l'accès ?
- Pour quel usage ?
- Intermédiation technologique (Uber, Blablacar, Airbnb...)
- Entraves à la concurrence, *algorithmic pricing*, comparateurs *Virtual Competition* (Ezrachi et Stucke, 2016)
- Déontologie scientifique, épistémologie et reproductibilité
- *Open data*, anonymisation, fin du consentement libre et éclairé
  - Projet `care.data` NHS et Royaume Uni : (*Social License*)
  - Projet *Data Science Initiative* : $\mathcal{X}$ et CNAM, base Sniiram

Loyauté

- *Trustworthiness* : Mériter la confiance : fiabilité, crédibilité, non discriminatoire
- *Accountability* : responsabilité, capacité à rendre compte

Décision algorithmique

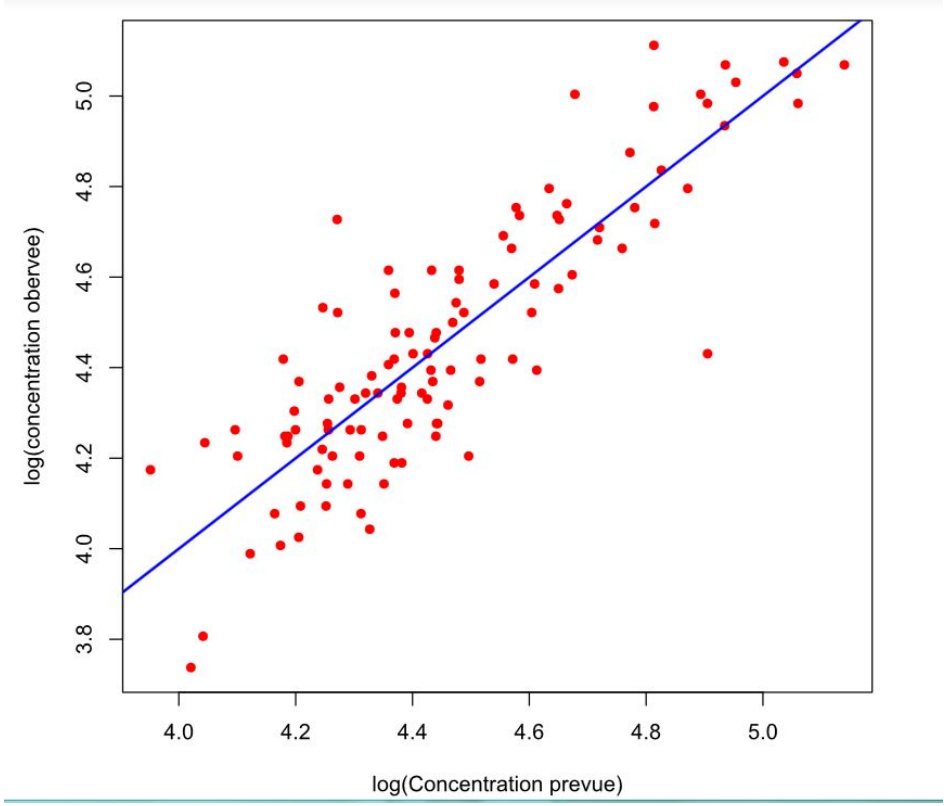
- Décision issue d'un traitement automatisée
- Algorithme procédural type APB (admission post-bac)
- Algorithme par Apprentissage Machine ou Statistique
  - Choix de :
    Traitement, action commerciale, maintenance préventive, accord de crédit, mise sous surveillance, d'un produit...
  - Prévision d'une probabilité ou risque de :
    Déclencher une maladie, départ d'un client, défaillance d'un système, défaut de paiement, radicalisation, d'appétence...
- Décision découle d'un Modèle ou Algorithme :
  - Estimé ou appris sur un *échantillon d'apprentissage*
  - Optimisé (compromis biais-variance) par *validation croisée*
  - Évalué sur un *échantillon test* indépendant

Loyauté des Algorithmes

- *Accountability* et *Trustworthiness*
- Se traduisent et s'évaluent par leur :
  - Explicabilité et transparence
  - Qualité de prévision et justesse de décision
  - Biais et discrimination
- Contraintes juridiques vs. caractéristiques techniques
  - Explicabilité et décret du 14/03 (Loi RN)
  - Qualité ?
  - Biais et discrimination individuelle ou collective
- Zone de non droit ou *disruption*
- Quelle éthique ?
- Enjeu : Acceptabilité d'une nouvelle technologie

Explicabilité : modèle linéaire du ”siècle dernier”

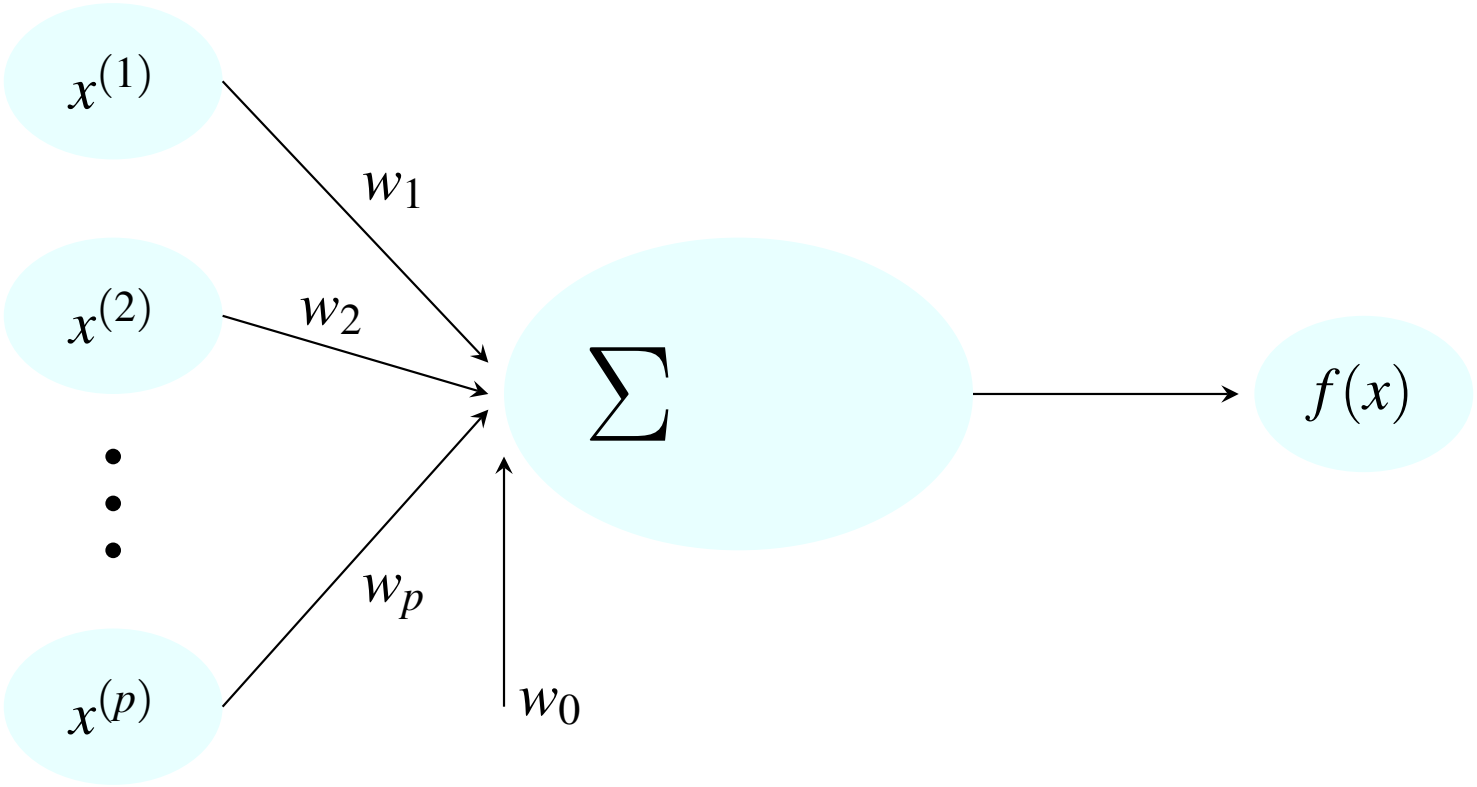
Prévoir la Concentration en Ozone



$$\begin{aligned}\log(\text{ConcODemain}) &= 2,4 + 0,35 \times \log(\text{ConcOJour}) + 0,05 \times \text{Sec} + \\ &+ 0,03 \times \text{T12} - 0,03 \times \text{Ne9} + 0.1 \times \text{Vx9}\end{aligned}$$

Modèle / Neurone Linéaire

Modéliser / prévoir une variable quantitative

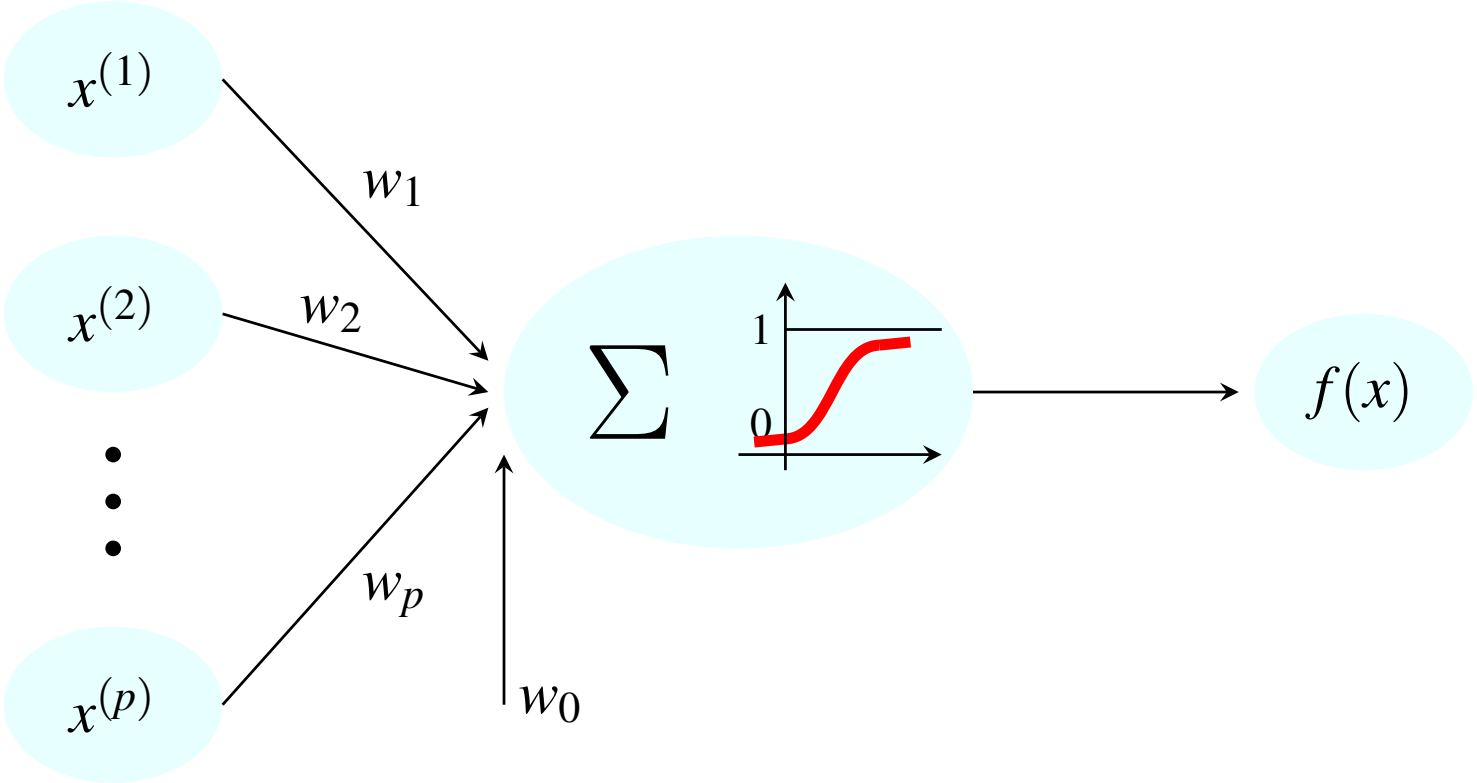


$$f(x) = w_0 + w_1 \times x^{(1)} + w_2 \times x^{(2)} + \cdots + w_p \times x^{(p)}$$

Compléments sur `wikistat.fr`

Modèle / **Neurone** *logistique*

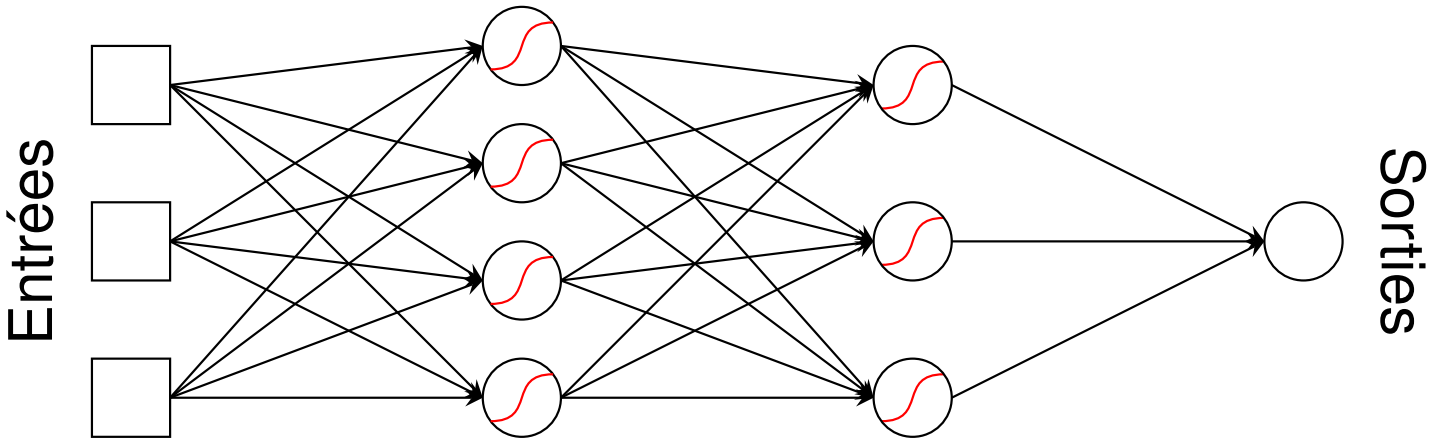
Variable binaire : Maladie, Panne, Départ, Faillite...



Exemple en épidémiologie : évaluer les facteurs de risque

Compléments sur `wikistat.fr`

Explicabilité : réseau de neurones : (Perceptron)

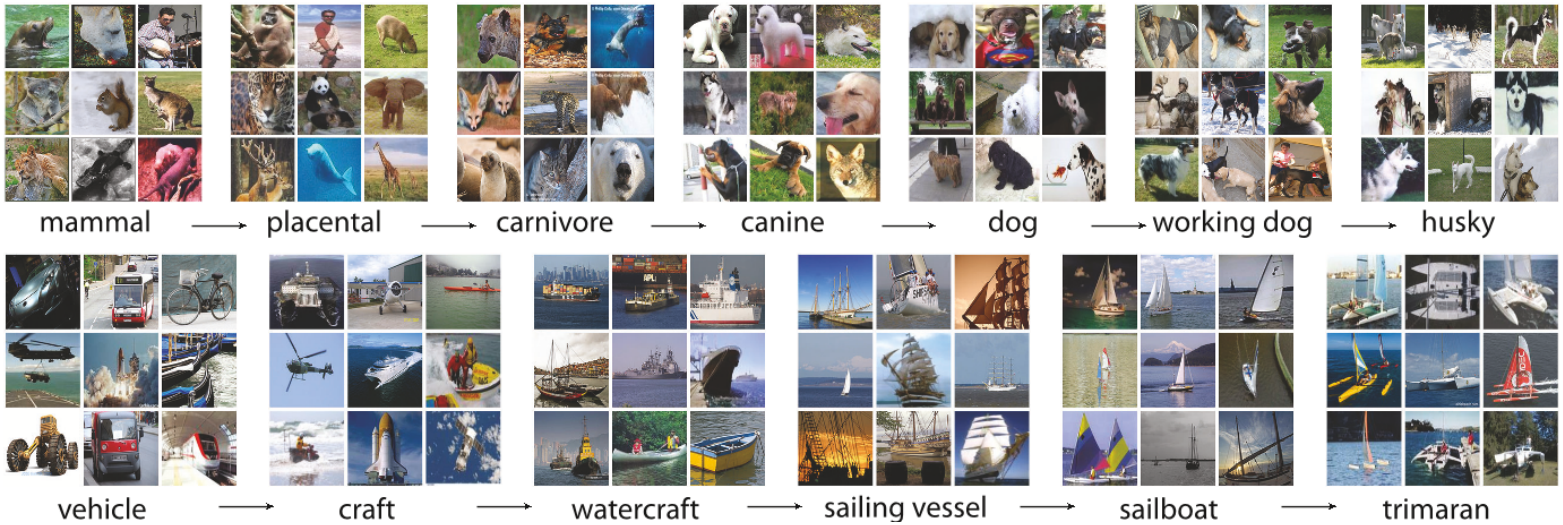


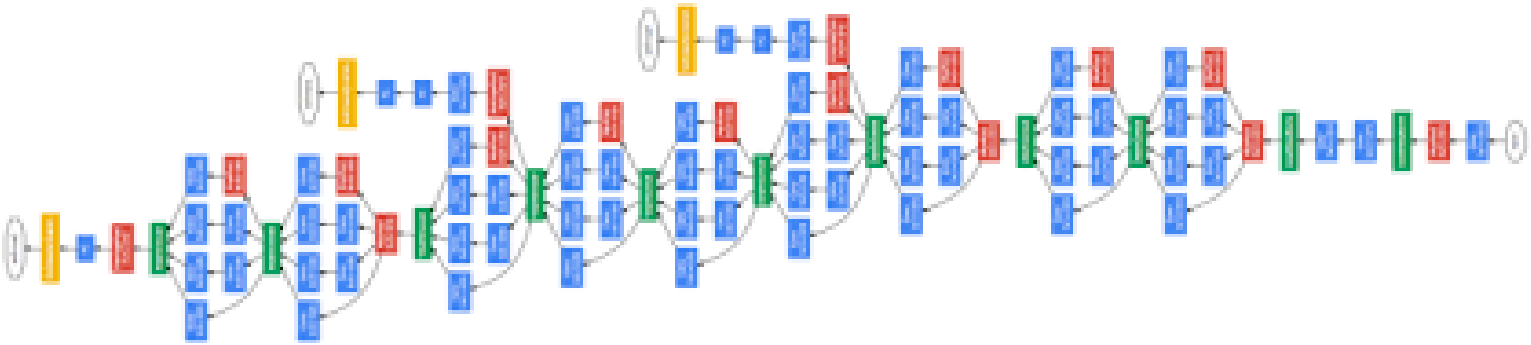
$x = (x^{(1)}, \ldots, x^{(p)})$ *Couche 1* *Couche 2* $y = F(x)$

- Explication impossible : *Boîte Noire*
- Aides à l'interprétation
- Idem pour k-p.p.v., SVM, *boosting, random forest*...

Explicabilité : Deep Learning 1

Exemple : base de données ImageNet :
15 millions d'images, 22000 catégories
2016 : 152 couches et mieux que l'expert humain





Grosses données et qualité de prévision

- Plus de données entraine-t-il une meilleure prévision ?
- *L'efficacité prédictive sera d'autant plus grande qu'elle sera le fruit de l'agrégation de données massives*
  in *La Gouvernementalité Algorithmique* (Rouvroy et Berns, 2013)
- **Vrai** et **Faux**
- Ne pas confondre estimation / prévision d'une moyenne
  (*loi des grands nombres*)
  et celle d'un comportement individuel
- Taux d'erreur élevés en marketing (15 à 30%)

Fiabilité des algorithmes

Exemples : *Google Flu Trend* (2008-2015), Médiamétrie

FINAL FINAL POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Entre échantillonnage et big data, les nouveaux enjeux de la mesure d'audience

28 août 2013

Qualité : Deep Learning 2

M PIXELS | CHRONIQUES DES (R)ÉVOLUTIONS NUMÉRIQUES

Derrière les dérapages racistes de l'intelligence artificielle de Microsoft, une opération organisée

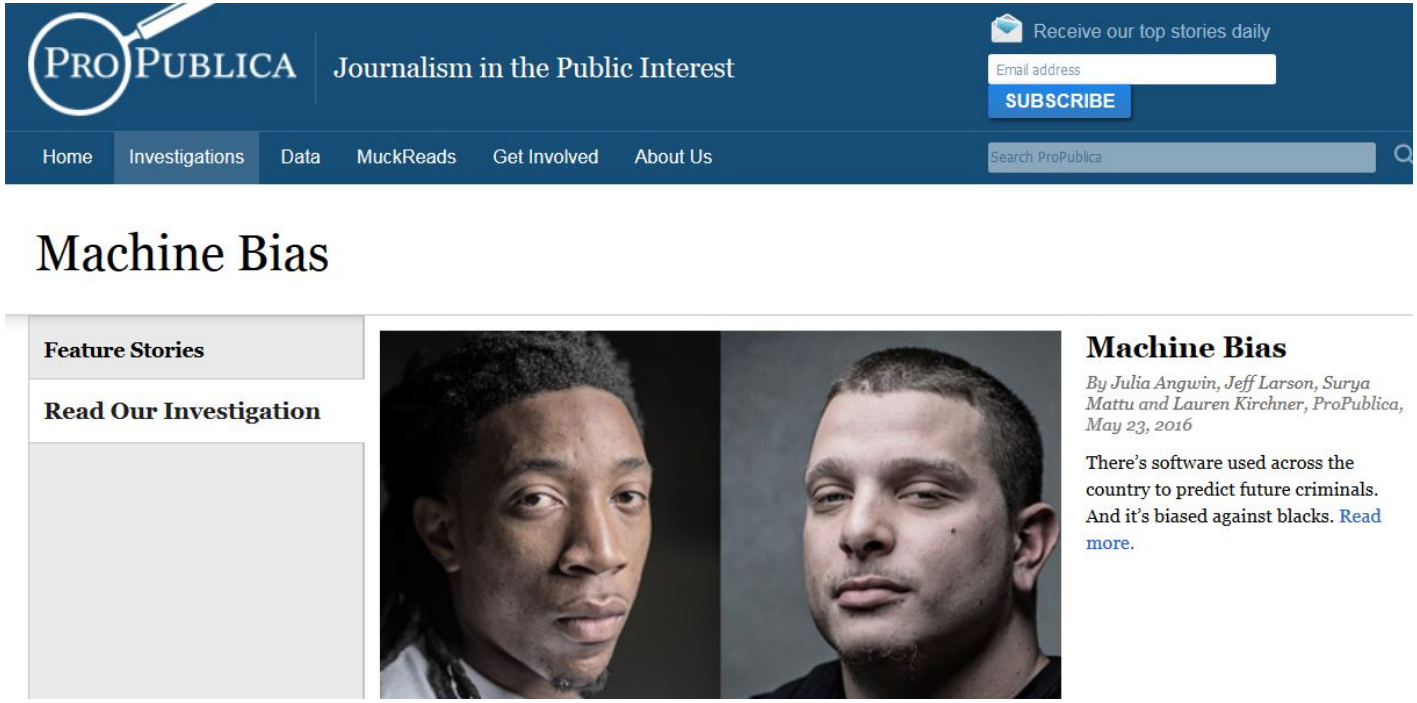
Partisans de Donald Trump, soutiens du GamerGate, ou simples internautes adeptes du chaos se sont associés pour transformer Tay, un programme censé imiter la conversation d'une adolescente, en nazie.

Le Monde.fr | 25.03.2016 à 15h50 • Mis à jour le 25.03.2016 à 17h04 |

Par William Audureau

Abonnez vous à partir de 1 € Réagir Ajouter

Justice prédictives : ProPublica vs. Equivant (NorthPointe Inc.)



Angwin et al. (2016)

ProPublica vs. Equivant (NorthPointe Inc.)

- Absence de discrimination selon NorthPointe Inc.
  - Distributions des scores (m_1 et m_2) similaires
  - Taux d'erreur ($FN + FP/n$) similaires
- Discrimination selon ProPublica

Matrice de confusion

Observation Récidive	Score Faible	Score Élevé	
Oui	FN	VP	q_1
Non	VN	FP	q_2
	m_1	m_2	n

  - Taux de faux positifs$= FP/q_2$
    afro-américain (45%) vs. caucasiens (25%)
- Taux de récidive afro-américain plus élevé (Chouldechova, 2016)
- Taux d'erreur très élevé (40%)

Loi *vs.* Algorithme

- Explicabilité et décret de la loi République Numérique
  - Algorithmes procéduraux et explicables (linéaire, arbre)
  - Algorithmes opaques ?
- Justesse de décision et qualité de prévision : rien !
- Discrimination et biais
  - Quelle Définition (légale ?) du biais ?
    Pedreschi et al. (2012), Zliobaité (2015) ...
  - Biais collectif (statistique) ou discrimination individuelle
  - Qui à charge de preuve ? pratique du *testing* ?

Actions *a priori*

- Objectifs : respect de la loi, éthique et acceptabilité
- Meilleur compromis interprétation / qualité (Zeng et al. 2016)
- Qualité de prévision : ”obligation” d’informer (cf.sondages)
- Algorithme le moins biaisé ou moins discriminatoire
  - Débiaiser l’échantillon d’apprentissage avec variable sensible connue
    Feldman et al. (2015) ...
  - Variable sensible anonymisée (confidentialité différentielle)
    Ruggiery (2014), Hajian et al. (2014)
  - Débiaiser l’algorithme
    Zemel et al. (2013) , Zafar et al. (2017)...

Éthique "industrielle"

Amazon, Google, Facebook, IBM, Microsoft, Apple...

1. We believe that artificial intelligence technologies hold great promise for raising the quality of people's lives and can be leveraged to help humanity address important global challenges such as climate change, food, inequality, health, and education.

...

7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.

Actions *a posteriori*

- Face aux Disruptions technologiques
- Faire évoluer le cadre juridique
- Auditer, contrôler un algorithme
- Comment ?
  - Vérifier la précision, l'explicabilité ou l'interprétabilité
  - Détecter un biais, collectif ou individuel (Ruggieri et al. 2010)
  - Par Testing ?
- Par qui ?
  - CNIL, DGCCRF (répression des fraudes)
  - Plateformes collaboratives : *Data transparency lab*, TransAlgo (INRIA)
  - Médias : *ProPublica*
  - Associations : Quadrature du Net, *Bayes Impact*
  - ... ?

Biais : Apprentissage Machine condamné

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

SPECIAL OFFER: JOIN NOW

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Oil at One-Year High on Falling Stockpiles

U.S. Stocks Rise on Oil Rally, Bank Earnings

Platinum Partners' Flagship Hedge Fund Files for Bankruptcy

MARKETS

U.S. Government Uses Race Test for $80 Million in Payments

Checks are ready for minority borrowers allegedly discriminated against on Ally Financial auto loans

By ANNAMARIA ANDRIOTIS and RACHEL LOUISE ENSIGN

Updated Oct. 29, 2015 9:32 p.m. ET

Recommended Videos

Conclusion : des questions

- Enjeu : Acceptabilité ou Rejet (cf. nanotech., OGM, care.data...)
- *Trustworthiness* et *Accountability*
  - Précision, explicabilité, biais des algorithmes
  - Cadre juridique restreint : proposer des modifications ?
  - Quels critères contrôler ?
  - Comment ? *Testing* déloyal ?
  - Par qui ?

Références

- Angwin J., Larson J., Mattu S., Kirchner L. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2017.
- Chouldechova A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv pre-print.
- Datta A., Sen S., Zick Y. (2016). Algorithmic Transparency via Quantitative Input Influence : Theory and Experiments with Learning Systems, in IEEE Symposium on Security and Privacy.
- Ezrachi A., Stucke M. (2016). *Virtual Competition The promise and perils of algorithmic-driven economy*, Harward University Press.
- Feldman M., Friedler S., Moeller J., Scheidegger C., Venkatasubramanian S. (2015). Certifying and removing disparate impact, arXiv-preprint.

- Feldman M., Friedler S., Moeller J., Scheidegger C., Venkatasubramanian S. (2015). Certifying and removing disparate impact, arXiv-preprint.

- Goodman B. (2016). A Step Towards Accountable Algorithms?:Algorithmic Discrimination and the European Union General Data Protection, in 29th Conference on Neural Information Processing Systems (NIPS 2016).
- Goodman B., Flaxman S. (2016). EU regulations on algorithmic decision-making and a ”right to explanation”, ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York.
- Hajian S., Domingo-Ferrer J., Farràs O. (2014). Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining, Data Mining and Knowledge Discovery 28 (5-6), 1158-1188.
- Kamiran F., Calders T. (2011). Data Pre-Processing Techniques for Classification without Discrimination, Knowledge and Information Systems 33(1).
- Kamiran F., Calders T, Pechenizkiy M. (2010). Discrimination Aware Decision Tree Learning in ICDM, 869-874.

Références – suite

- Pedreschi D., Ruggieri S., Turini F. (2008). Discrimination-Aware Data Mining. In *KDD*, pp. 560-568.
- Rouvroy A., Berns T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation, *Réseaux*, 177, 163-196.
- Ruggieri S. (2014). Using t-closeness anonymity to control for non-discrimination, Transaction on Data Privacy, 7, 99-129.
- Ruggieri S., Pedreschi D., Turini F. (2010). Data mining for discrimination discovery. In TKDD 4(2).
- Zafar M., Valera I., Rodriguez M., Gummadi K. (2017). Fairness Constraints : Mechanisms for Fair Classication in International Conference on Articial Intelligence and Statistics (AISTATS), vol. 5.
- Wachter S., Mittelstadt B., Floridi L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, à paraître.
- Zafar M., Valera I., Rodriguez M., Gummadi K. (2017). Fairness Constraints: Mechanisms for Fair Classication in International Conference on Articial Intelligence and Statistics (AISTATS), vol. 5.
- Zemel R., Wu Y., Swersky K., Pitassi T., Dwork C. (2013). Learning Fair Representations in JMLR W&CP 28(3), 325 ?333.
- Zeng J., Ustuny B., Rudin C. (2016). Interpretable Classification Models for Recidivism Prediction, arXiv pre-print.
- Zliobaité I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

Quelques sites consultés

- Le Monde: IBM, Google, Microsoft et le cancer
- ArsTechnika: Programme Skynet de la NSA
- Mesure d'audience
- Le Monde: révélation du code source d'APB
- Predpol: Police prédictive
- NorthPointe : prévision de la récidive
- ProPublica : article sur le biais des "machines"
- Le Monde: *Tay, ChatBot* raciste de Microsoft
- Village de la Justice: Justice prédictive
- Wall Street Journal: condamnation du gouvernement
- Partenariat sur l'IA