Quelle transparence pour les algorithmes d’apprentissage machine ?

Issam Ibnouhsein

17 octobre 2019

Discours sur les algorithmes

Importance croissante du discours sur le rôle social des algorithmes : surveillance, discrimination, prise de décision, automatisation du travail...

Saisir la singularité des algorithmes

Ne pas rejouer la critique de la bureaucratie : certains algorithmes sont juste une automatisation de procédures préexistantes.

Intérêt pour l'AM : des algorithmes qui ne décident pas comme les humains.

Plan

1. Les sens de la transparence
2. L'intelligibilité des algorithmes en AM : remarques transverses
3. Enjeux techniques autour de l'intelligibilité des sorties des algorithmes d'AM
4. Conclusions

1 Les sens de la transparence

Deux familles de propriétés

Distinction de deux familles :

- Famille de propriétés **normatives** extrinsèques :
  - **Loyauté :** Un algorithme est loyal si la fonctionnalité affichée auprès de l'utilisateur est identique à la fonctionnalité connue du fournisseur.
  - **Équité :** un algorithme est équitable si son fonctionnement ne provoque pas d'effets discriminants à l'égard d'une partie de la population.

- Famille de propriétés **épistémiques** intrinsèques :
  - **Intelligibilité :** un algorithme est intelligible s'il est possible de comprendre son comportement dans l'état de l'art scientifique.
  - **Explicabilité :** Un algorithme est explicable s'il est possible de faire comprendre son fonctionnement à un utilisateur (sans expertise scientifique).

Relations entre les concepts : l'épistémique comme fondement du normatif

L'explicabilité dépend de l'intelligibilité : il est nécessaire de comprendre pour expliquer.

Intelligibilité = explicabilité fondamentale.

L'intelligibilité est fondamentale pour vérifier qu'un algorithme est loyal et équitable :

- Il faut expliquer pour être loyal.
- Le manque d'intelligibilité peut créer des effets discriminants inattendus.

2 L’intelligibilité des algorithmes d’AM : remarques transverses

Une distinction stratégique face à la prolifération terminologique

Intelligibilité des sorties du modèle (numériques, graphiques, prédiction, décision, action sur son environnement...)

VS

Intelligibilité de la procédure (fonctionnement de l'algorithme et du programme)

On peut bien comprendre une procédure, d'un point mathématique ou intuitif, sans bien comprendre une sortie donnée.

Intelligibilité des sorties dans les procédures bureaucratiques

Les procédures bureaucratiques sont **compositionnelles** (suite de décisions élémentaires simples).

Elles permettent une **explicabilité par extraits** : sélection de quelques éléments simples, compréhensibles et pertinents dans l'arbre de décision.

Essentiel pour permettre la **croissance en taille des procédures sans compromettre l'explicabilité** pour le public.

Un enjeu stratégique pour l'emploi de l'apprentissage machine

Des algorithmes "conventionnels" sont souvent d'une grande sophistication mathématique : l'explicabilité de l'ensemble de la procédure est compromise.

La simple taille des procédures bureaucratiques les plus courantes rend leur intelligibilité difficile.

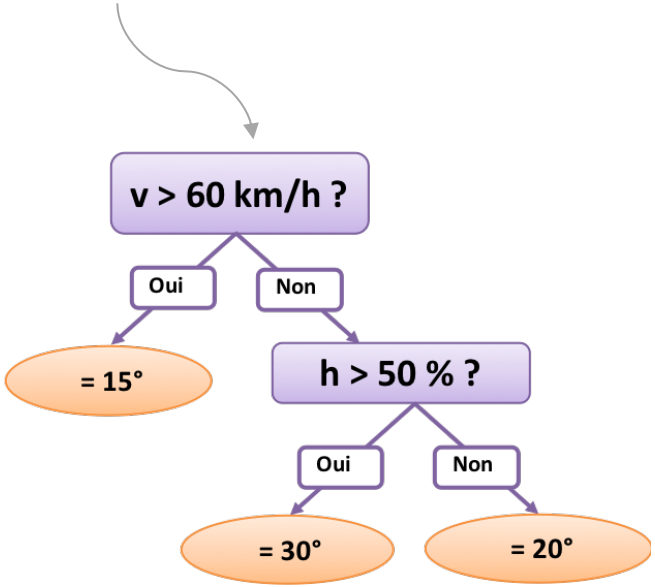
Enjeu pratique de l'explicabilité de l'AM : non pas tant faire de la vulgarisation d'une classe de techniques originale comme les réseaux de neurones profonds, mais extraire une explication brève et compréhensible des sorties.

Est-ce possible ?

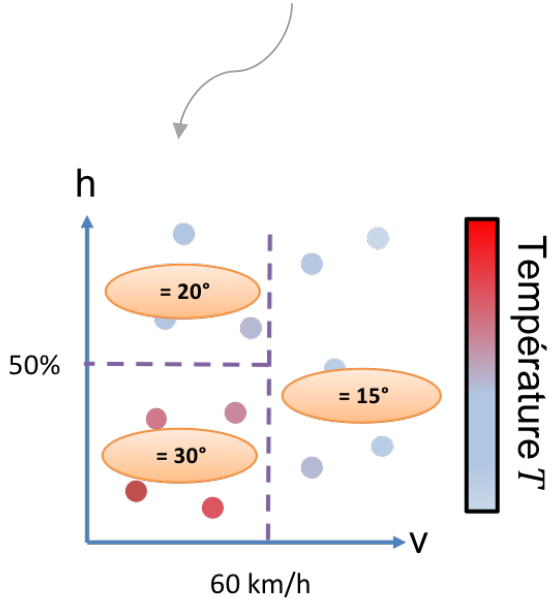
3 Enjeux techniques autour de l’intelligibilité des sorties

Explicitation des segmentations établies par un arbre de décision dans un espace de données très simple à deux dimensions (v,h)

Un arbre de décision est une représentation graphique de segmentations dans l'espace des données : elles sont intelligibles car les variables sont explicites et les frontières bien définies

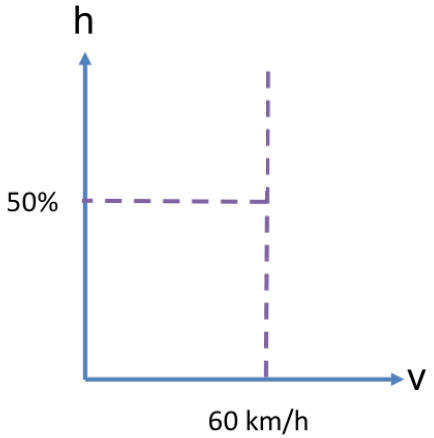


Segmentation établie par un arbre de décision dans l'espace de l'humidité et de la vitesse du vent pour prédire la température



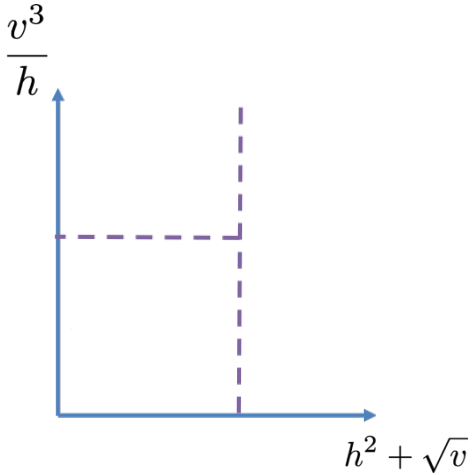
***Note :** la profondeur de l'arbre est fixée à l'avance et correspond à un **hyper-paramètre** du modèle, à optimiser indépendamment de ses paramètres que sont le choix des variables et seuils à chaque noeud

Mais les dimensions selon lesquelles est établie une segmentation ne sont pas toujours facilement interprétables…



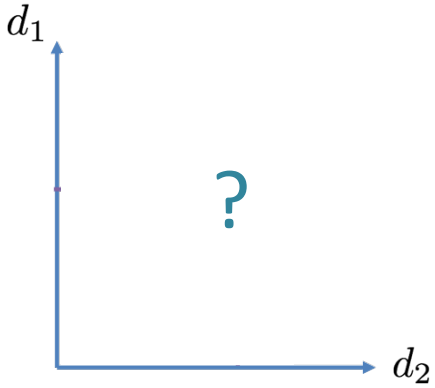
Ex : arbre de décision

Segmentation claire en fonction des variables d'entrée



Ex : modèles paramétriques

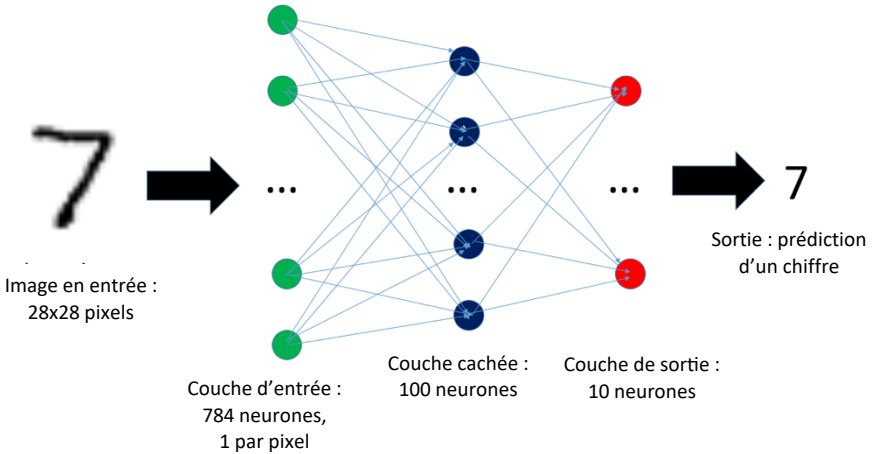
Transformation intelligible des variables
Segmentation formalisable mais dont le sens peut être opaque, ex : crédit part.

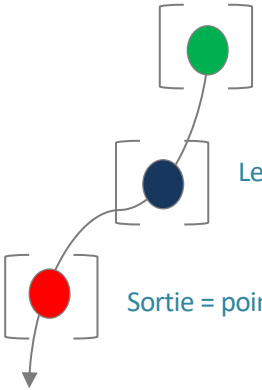


Ex : calcul intermédiaire d'un réseau de neurones

Transformation non-intelligible des variables
Des segmentations peuvent être établies à chaque étage du réseau, mais leur évolution et sens restent largement opaques

Examinons concrètement le cas d'un réseau de neurones : souvent, seules les dimensions au départ et à la fin possèdent un sens clair





Entrée = point d'un espace vectoriel de dimension 784

Les points intermédiaires n'appartiennent pas à des espaces à dimensions facilement intelligibles

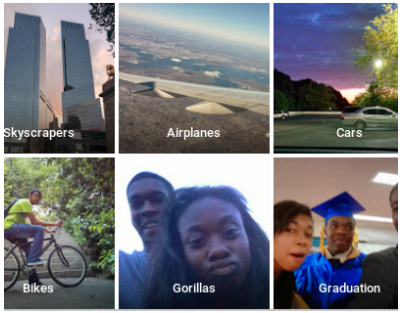
Sortie = point d'un espace vectoriel de dimension 10

***Note :** le nombre de neurones par couches est un exemple d'hyper-paramètre d'un réseau de neurones

- L'intelligibilité des sorties d'un algorithme d'AM n'est donc pas toujours garantie, car même si les critères de production de la sortie peuvent être approximés par des critères explicites, ces derniers ne sont pas nécessairement intelligibles, y compris pour un expert.
- De là naît le sentiment d'opacité entourant certaines applications de l'AM, en particulier l'analyse d'images par des réseaux de neurones profonds

Certaines procédures d'IA sont aujourd'hui limitées en taille par leur absence d'explicabilité par extrait

Deux exemples de polémiques récentes :



Google obligé de ré-entraîner son modèle sans gorilles dans les données : couplage fort de la procédure globale avec calcul d'une sortie ≠ processus bureaucratique, exemple : code sécurité sociale

Tesla pas responsable car le conducteur disposait de 7s pour réagir : cadeau juridique empoisonné car le but d'une voiture autonome à terme est d'être… autonome !

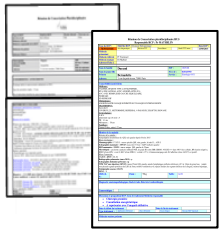
L’explicabilité est un facteur extrêmement important pour l’IA en médecine

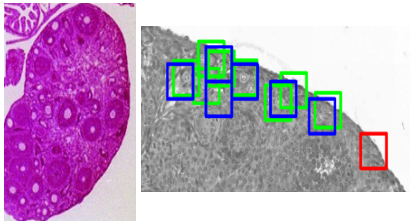
A novel machine learning-derived decision tree including uPA/PAI-1 for breast cancer care

December 2018 · Clinical Chemistry and Laboratory Medicine

DOI: 10.1515/cclm-2018-1065

Nathalie Reix · Massimo Lodi · Stéphane Jankowski · Show all 18 authors · Carole Mathelin





High-throughput ovarian follicle counting by an innovative deep learning approach

Charlotte Sonigo, Stéphane Jankowski, Olivier Yoo, Olivier Trassard, Nicolas Bousquet, Michael Grynberg, Isabelle Beau & Nadine Binart

- Constitution de cohortes à partir de données textuelles
- Modalités de détection/décompte d’objets sur des images
- Etc.

4 Conclusion

AM et intelligibilité

- Ne pas opposer systématiquement AM et et autres familles d'algorithmes
- Ne pas tomber dans une catégorisation simpliste des modèles en « modèles opaques » et « modèles transparents »
- Affiner les méthodes d'explicabilité par extrait des sorties de modèles pour pouvoir croître en taille dans les procédures
- Se souvenir qu'il s'agit d'un champ scientifique jeune et en pleine évolution…