Une introduction aux méthodes d'IA explicables

Pierre-François Gimenez

"Papers please", 31 mars 2022

CentraleSupélec

Contexte

De l'IA de plus en plus utilisée par les décideurs

- Aide à la décision d'acceptation de crédits bancaires
- Aide à la décision de libération sous caution
- Aide au diagnostic médical

Des enjeux sociétaux importants

- Les modèles peuvent être biaisés, par exemple si les données d'apprentissage sont elles-mêmes biaisées
- Les modèles sont souvent des boîtes noires difficiles à comprendre
- L'explication est une manière de s'assurer qu'un système est digne de confiance
- RGPD: droit à l'explication (loi similaire aux US)

Nous allons voir différentes stratégies d'explication

CentraleSupélec

Les familles de méthodes d'explication

Quatre grandes familles

- Modèles intrinsèquement interprétables
- Explication applicable à tout modèle
  - Étude de l'effet des attributs sur la prédiction
  - Modèle de substitution (*surrogate model*) global
  - Modèle de substitution local
- Explication par exemples
  - Exemple contrefactuel
  - Instances influentes
  - k-NN
- Carte de saillance (spécifique aux réseaux de neurones)

Modèles intrinsèquement interprétables

Des modèles interprétables

Le plus simple est d'utiliser directement un modèle interprétable, comme :

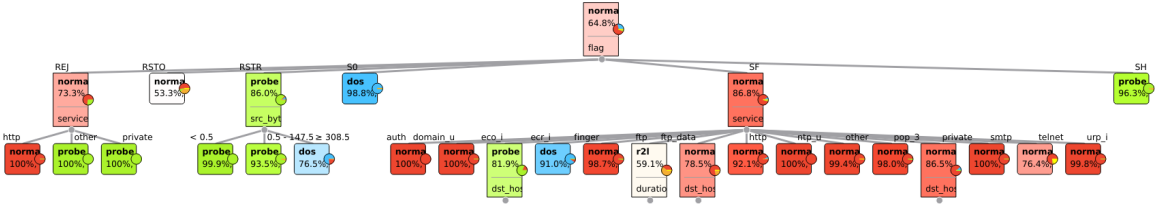
- un arbre de décision
- des règles de décision
- une régression linéaire
- un classifieur bayésien naïf
- les k plus proches voisins

Inconvénient

Généralement, ces modèles n'ont pas les meilleures performances

Modèles intrinsèquement interprétables

Exemple : un arbre de décision pour IDS réseau



Explication applicable à tout modèle

Étude de l'effet des attributs sur la prédiction

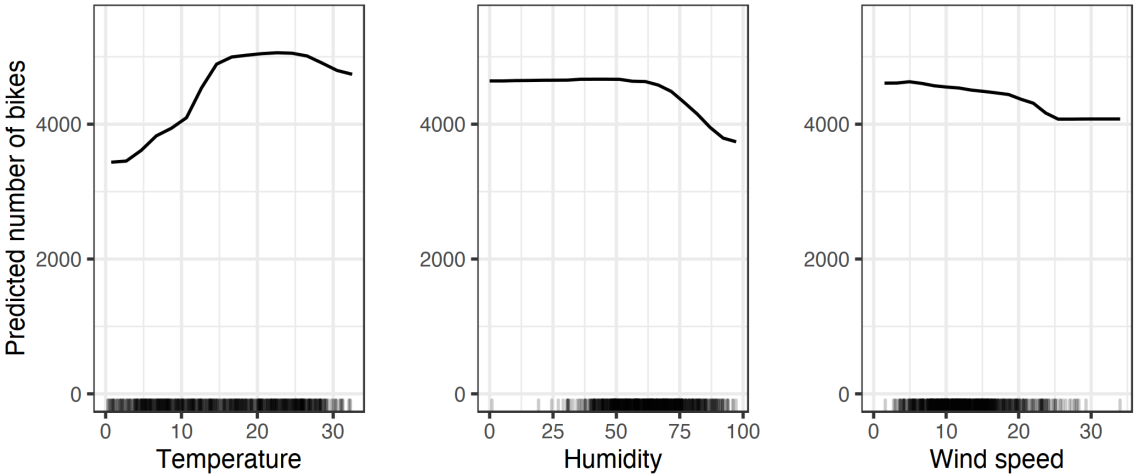
- Plusieurs méthodes : PDP, M-plots, ALE plots
- Permet de connaître l'influence d'un seul attribut sur la prédiction
- Pour cela, on modifie les valeurs de l'attribut concerné dans l'ensemble d'apprentissage et on vérifie la prédiction

Inconvénient

Ne permet d'étudier qu'un attribut à la fois

Explication applicable à tout modèle

Exemple : PDP



CentraleSupélec

Explication applicable à tout modèle

Modèles de substitution global

- Idée : on approche un modèle complexe avec un modèle interprétable
- On mesure la fidélité du nouveau modèle par leur différence d'accuracy
- Ce modèle est utilisé pour générer des explications

Inconvénient

Est-ce vraiment une explication du modèle, ou juste une illusion d'explication ?

CentraleSupélec

Explication applicable à tout modèle

Modèles de substitution local

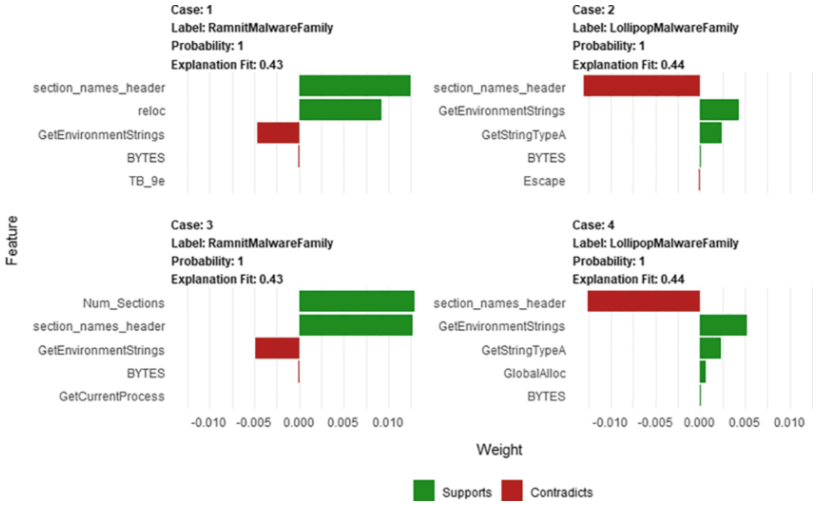
- Idée : on approche un modèle complexe localement (autour d'une prédiction)
- LIME : approche le modèle avec un modèle explicable (régression linéaire / arbre de décision) en s'appuyant sur les prédictions d'instances voisines
- SHAP : s'appuie sur les "Shapley values". Similaire à LIME.

Inconvénient

Peu d'inconvénients, mais pas évident de bien définir la notion de voisinage

Explication applicable à tout modèle

Exemple : explication calculée avec LIME



Explication par exemples

Exemple contrefactuel

- Idée : expliquer quelle modification de l'entrée permettrait de changer de classe prédite
- Permet d'avoir une explication du type "Si vous gagnez 50€ de plus par mois, alors le crédit serait accordé"
- Pas toujours fiable (cf. exemple adversarial)

Inconvénient

- De nombreuses explications différentes possibles : quelle est la plus pertinente ?
- Par exemple : "Si vous étiez un an plus jeune, le crédit serait accordé". Oui, sauf qu'on ne peut pas rajeunir...

Explication par exemples

Instances influentes

- Une instance influente est une instance du jeu d'apprentissage qui a un effet important dans l'apprentissage
- On estime l'influence d'une instance par la modification du modèle appris si cette instance est absente
- Permet de mieux comprendre le jeu de données
- Une autre technique (utilisant des prototypes et des critiques) fonctionne d'une manière similaire

Inconvénient

Complexité : demande de nombreux réapprentissages...

Carte de saillance

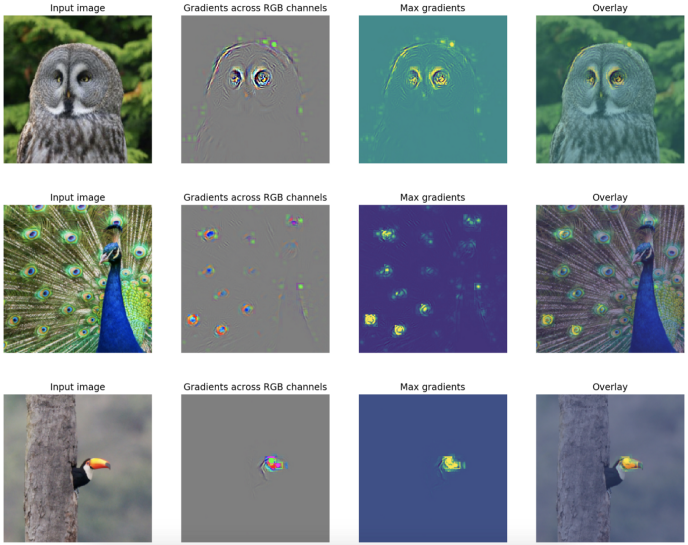
Carte de saillance (*saliency map*)

- Pour chaque attribut, on calcule son score de pertinence dans la prédiction
- Limité aux réseaux de neurones. On retire la couche soft-max final s'il y en a une.
- Il y plusieurs méthodes (Gradient, x $\odot$ Grad, Integrated Gradient, Layer-wise Relevance Propagation) qui s'appuient toute sur le gradient
- Principalement utilisé sur les images pour savoir sur quelle partie s'appuie la prédiction

Inconvénient

Explications fragiles, principalement qualitatives

Carte de saillance



CentraleSupélec

Caractéristiques des méthodes

Méthodes	Véritable ?	Agnostique ?	Global / local
Modèle intrinsèquement interprétable	✓	×	Global
PDP / ALE	×	✓	Global
Substitution globale	×	✓	Global
Substitution locale (LIME)	×	✓	Local
Exemple contrefactuel	×	✓	Local
Instances influentes	×	✓	Local
Carte de saillance	×	×	Local

CentraleSupélec

Et la sécurité de tout ça ?

Encore et toujours des exemples adversariaux

- Pas mal d'articles sur comment modifier les explications pour faire passer un modèle injuste pour un modèle juste
- Cela concerne quasiment toutes les méthodes
- Même des méthodes intrinsèquement explicables peuvent être manipulées en utilisant des attributs corrélés (utiliser le code postal au lieu de l'ethnie sur les données des USA...)
- Encore du boulot...

Conclusion

L'utilisé des explications

Les explications permettent :

- de s'assurer que le modèle apprend la bonne chose (reconnaître les loups, pas la neige)
- de vérifier qu'il s'appuie sur des attributs pertinents
- de comprendre les mauvaises prédictions pour améliorer le modèle (ou le dataset...)

Beaucoup de méthodes différentes aux caractéristiques variées

- Méthodes approximative ou non
- Explication spécifique à un modèle ou agnostique/générique
- Explication locale ou globale

Dans tous les cas, garder à l'esprit qu'actuellement les explications restent manipulables...

Plus d'infos sur `https://christophm.github.io/interpretable-ml-book/`