Sorbonne Université
Master DAC (2e année)
XAI

Année 2022-2023

G. Bourgne, M.-J. Lesot
C. Marsala, J.-N. Vittaut

Examen Session 1

Durée 2h - tout document autorisé
tout support électronique interdit

Le barème n'est donné qu'à titre indicatif.

Exercice 1 : Exemples contrefactuels *(5 points)*

On note

- $f : \mathcal{X} \to \{0, 1\}$ un classifieur binaire,
- x une donnée dont on veut expliquer la prédiction par f,
- $\delta \in \mathbb{R}^+$ et $\lambda \in [0, 1]$ deux hyperparamètres,
- $\hat{p}_y$ une estimation de la densité des données de la classe $y \in \{0, 1\}$, par exemple apprise par KDE,
- $\mathcal{E}(x, \hat{p}, \delta) = \{e \in \mathcal{X} \mid f(e) \neq f(x) \wedge \hat{p}_{f(e)}(e) > \delta\}$.

Il a été proposé d'expliquer la prédiction $f(x)$ par la donnée e^* définie comme l'optimum de la fonction de coût suivante

$$e = \arg \min_{e \in \mathcal{E}(x,\hat{p},\delta)} \|x - e\|_2 + \lambda \|x - e\|_0$$

1. Commenter les différentes composantes de cette fonction de coût en indiquant leur sens et leur influence sur le choix de l'explication. Indiquer en particulier en quoi on peut considérer que le résultat de son optimisation constitue une explication contre-factuelle.
2. Indiquer les hypothèses d'agnosticité par rapport aux données et par rapport au modèle sur lesquelles repose cette approche.
3. Représenter graphiquement (schématiquement) des données en 2D, une frontière de décision et une donnée x dont on veut expliquer la prédiction et indiquer, en justifiant, dans quelle zone se situe l'explication fournie par cette méthode. L'exemple choisi doit illustrer l'intérêt de cette fonction de coût par rapport à la définition classique des explications contre-factuelles, qu'il est demandé de commenter.

Exercice 2 : Vecteur d'importance de poids *(4 points)*

1. L'objectif de cette question est d'illustrer l'influence du paramètre de définition du voisinage pour la méthode LIME.

   Représenter graphiquement (schématiquement) des données en 2D, une frontière de décision et une donnée x dont on veut expliquer la prédiction. Représenter de plus les résultats obtenus par LIME pour deux valeurs différentes du paramètre de voisinage, en justifiant ces derniers et commenter.

   En particulier, une des deux valeurs de paramètre devra illustrer les limites de la méthode.
2. En argumentant votre réponse, indiquer quelle est, parmi les méthodes de calcul de vecteur d'importance de poids, votre méthode favorite et son apport à l'interprétabilité d'un classifieur.

Exercice 3 : Modèles de substitution *(6 points)*

Remarque : les questions de cet exercice sont indépendantes les unes des autres.

1. Dans l'arbre de décision représenté dans la figure 1, quels sont les attributs catégoriels ? Proposer un nouvel arbre de décision issu de cet arbre et obtenu en transformant les attributs catégoriels par un encodage "one hot".

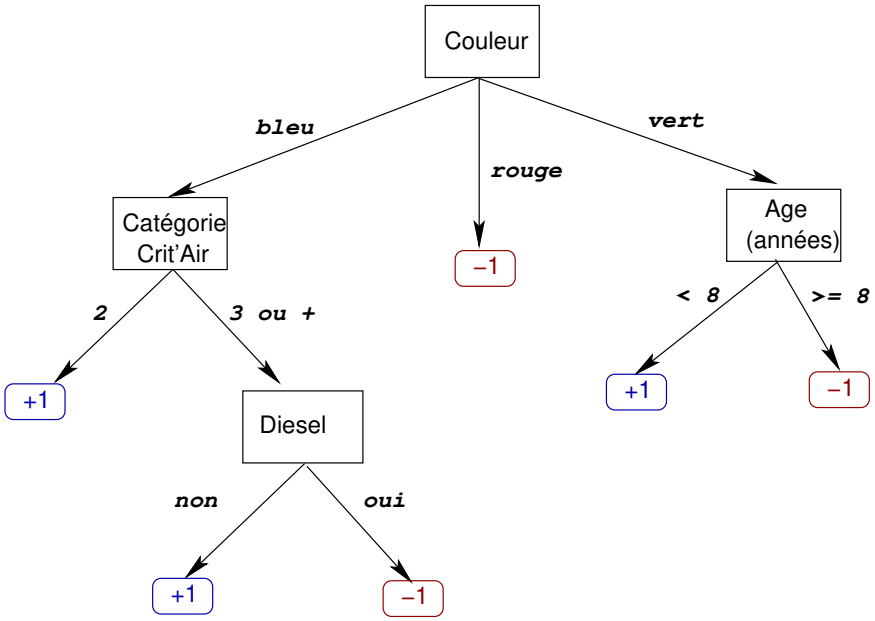

FIGURE 1 – Arbre de décision

2. Rappeler les étapes de la construction d'une explication pour la décision $f(\mathbf{x})$ prise par un modèle boite noire f pour la donnée $\mathbf{x}$ grâce à l'approche LORE.
3. On considère l'exemple $\mathbf{x_0}$ qui a moins de 5 ans, est de couleur bleue, de crit'air 3 et qui n'est pas diesel. Un modèle f, boîte noire, classifie $\mathbf{x_0}$ dans la classe +1. Afin d'expliquer la décision prise par f, on a obtenu l'arbre donné dans la figure 1.
   (a) Quels sont les exemples contrefactuels permettant d'expliquer la décision prise par f ?
   (b) Quelle explication est produite par LORE à partir des exemples contrefactuels obtenus ?

Exercice 4 : Sous-ensembles flous et interprétabilité *(5 points)*

On souhaite faire un système flexible de réservation de billets d'avion, permettant aux utilisateurs de trouver facilement les vols les plus appropriés par rapport à leurs souhaits. On considère qu'on a accès aux 5 possibilités de vols vers la destination désirée, décrits par leur aéroport de départ, leur durée de vol et leur prix :

id	aéroport	durée (min)	prix (€)
v_1	Orly	125	195
v_2	Orly	90	199
v_3	Roissy	90	200
v_4	Orly	150	200
v_5	Orly	140	250

1. L'utilisateur u_1 souhaite un vol partant d'Orly, d'une durée strictement inférieure à 2 heures et d'un prix inférieur ou égal à 200 euros. Proposer une formalisation de cette requête et indiquer le résultat obtenu pour les vols indiqués dans le tableau ci-dessus, en justifiant votre réponse.
2. L'utilisateur u_2 souhaite un vol partant d'Orly, d'environ 2 heures et autour de 200 euros. Proposer une formalisation complète de cette requête en détaillant ses composantes et indiquer le résultat obtenu pour les vols indiqués ci-dessus, en justifiant votre réponse.
3. (exemple totalement fictif et fantaisiste) Les émissions carbone du modèle d'avion ABX4242 sont optimisées pour le transport de 150 passagers. Pour un vol d'une durée de 2 heures, elles sont

calculées, en fonction du nombre de passagers p, par

$$e(p) = \frac{1}{2}(p - 150)^2 + 1000$$

Donner l'estimation des émissions de carbone si l'avion transporte environ 160 personnes, valeur approchée représentée par le sous-ensemble flou triangulaire de support $[130, 170]$ et de noyau 160.

Une réponse analytique et une représentation graphique (schématique et s'appuyant sur des points particuliers d'intérêt) sont demandées.

4. (idem : fictif et fantaisiste) On considère la règle suivante

   > Si l'avion transporte entre 120 et 150 personnes environ, la quantité de jus de tomate servi est d'approximativement 8 litres.

   et le cas d'un vol sur lequel environ 160 passagers ont embarqué.

   En quoi la logique floue peut-elle être vue comme influençant (dans quel sens) l'interprétabilité du résultat obtenu par rapport à la logique classique ?

   Remarque : aucun calcul n'est demandé dans cette question.