XAI
eXplainable Artificial Intelligence
IA explicable

Cours 1 - mardi 19 septembre 2023

Marie-Jeanne Lesot
Christophe Marsala
Jean-Noël Vittaut
Gauvain Bourgne

LIP6, Sorbonne Université

Contexte

- Omniprésence de l'IA
  - qualité des résultats obtenus dans multiples domaines
  - quelques exemples : santé, marketing, cybersécurité, transport, océanographie, assistants personnels, études, ...
- Complexité accrue des modèles de Machine Learning
  - réseaux profonds, XGBoost, forêts aléatoires
  - modèles **boîtes noires**
- **Dangers : risque de biais, d'opacité**, de discrimination, manque de confiance, incompréhension, ...
  - scandale COMPAS 2016 : prédiction du risque de récidive, outil utilisé par plusieurs juridictions aux Etats-Unis, mais fortement biaisé

Contexte

- Problèmes de biais et d'opacité

Chouldechova 2017

"Milla Jovovich"

Sharif et al. 2016

"Does your car have any idea why my car pulled it over?"

The New Yorker

RGPD (UE, 2016)

Règlement général sur la protection des données

- Soit aussi GDPR, General Data Protection Regulation
- Quelques unes des principales dispositions (source : Wikipedia)
  - accord explicite et positif pour les cookies sur les sites internet
  - droit à l'effacement, à la portabilité
  - *privacy by design*, notifications en cas de fuite, obligation en cas de cyberattaque, délégué à la protection des données
  - profilage : ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire
- Mention d'un **droit à l'explication**

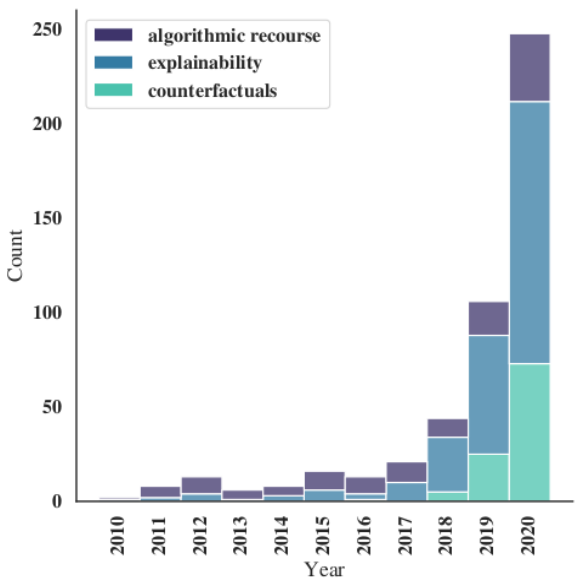
Le RGPD et avant

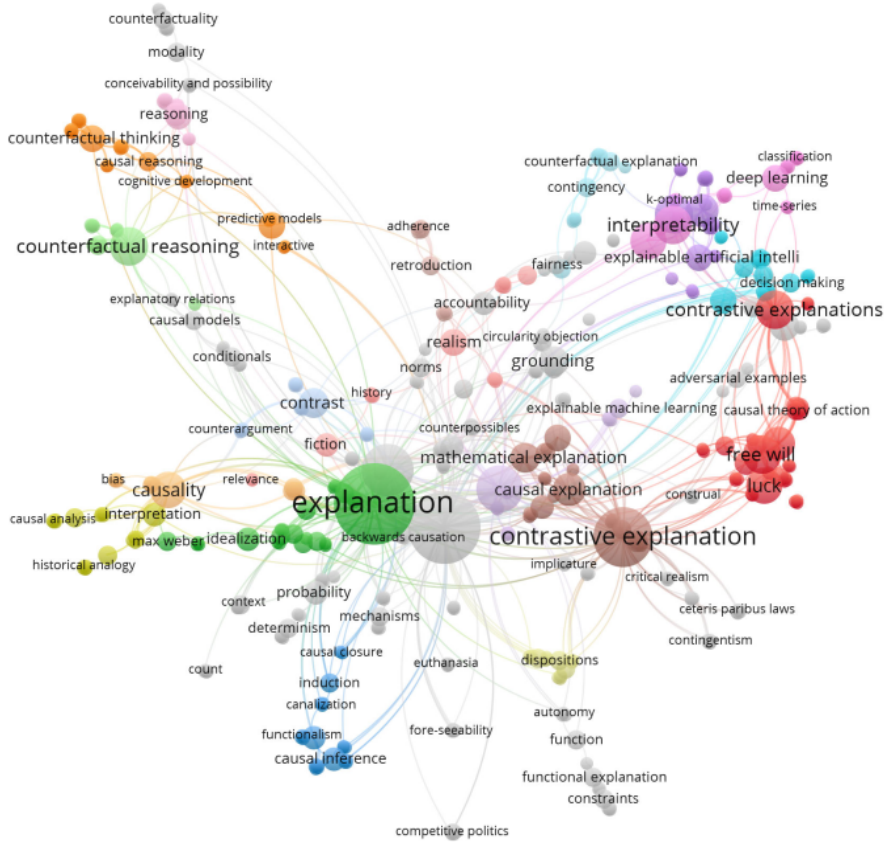
- Mention d'un **droit à l'explication**
  - si les données sont traitées par un algorithme,
  - une explication de la décision doit être fournie
- En France : loi pour une république numérique, 2016
  - si une administration publique prend une décision basée sur un traitement algorithmique,
  - les règles qui définissent le traitement et ses principales caractéristiques doivent être communiquées à l'utilisateur à sa demande
- Exemples
  - calcul des impôts, affectation des élèves dans les établissements scolaires/universitaires
  - c-à-d le débat Parcoursup

Le RGPD et après

- Appel à projets de la DARPA 2017 : terme **XAI**
  - eXplainable AI
  - comme une implémentation du droit à l'explication
  - objectif général : fournir des réponses à la question **Pourquoi ?**
- Des tas de néologismes et concepts liés, par exemple
  - interpretability, explainability, accountability
  - transparency, fairness
- Des tas de sessions spéciales, workshops, conférences dédiées
  - WHI/HiLL@ICML, XAI@IJCAI, FATML@KDD, NeurIPS, IUI, ...
  - AIES, FAccT, CHI, ...
  - GdR IA GT Explicabilité, atelier madics FENDER, ...

Côté publications scientifiques





(Figures tirées de Pawelczyk et al. 21, Stepin et al. 21)

⟹ **XAI : une thématique cruciale et en pleine expansion**

Ampleur du domaine

$\Longrightarrow$ **Enorme diversité des méthodes !**

- Absence de consensus

  (Doshi-Velez et Kim 17, Lipton 17, Mueller et al. 19, Weller 19)

  - sur une définition formelle
  - sur des propriétés souhaitées : qu'est-ce qu'une bonne explication ?
    comment mesurer la qualité d'une explication ?
    qu'est-ce qu'une explication ?
    quelles questions se pose l'utilisateur ?

    (Wachter et al. 18, Miller 18, Liao et al. 20)

- Multiplicités d'approches, de catégorisations, d'axes de discussion

  (Guidotti et al. 18, Biran et Cotton 19, Artelt et Hammer 19, Carvalho et al. 19, Molnar 19, Verma et al. 20, Rudin et al. 21, Guidotti et al. 22)

- Problématique transdisciplinaire (Miller 19, Wachter et al. 18)
  - informatique, sciences cognitives, pédagogie, philosophie, droit, ...

Ampleur du domaine

- **Diversité des termes**
  - interprétabilité, explicabilité, accountability, transparence, équité, ...
- **Diversité des tâches**
  - classification, régression, clustering, détection d'exceptions, ...
  - recommandation, planification, agents autonomes, allocation de ressources, interactions,
- **Diversité des buts** : expliquer pour
  - comprendre, augmenter la confiance, optimiser la décision, ...
  - corriger le modèle, augmenter l'équité, ...
- **Diversité des destinataires** de besoins et expertises variables
  - expliquer à des utilisateurs, des experts, des concepteurs du modèle
- **Diversité des hypothèses** sur les informations disponibles
  - accès au modèle, aux données d'apprentissage, à des données non étiquetées, à des connaissances expertes, ...

Multiplicité des formes d'explications

- Représentation graphique
  - *partial dependence plots*, effet marginal d'un attribut sur la prédiction
  - *saliency maps* et les approches pour les images
- Classifieur lui-même, s'il est suffisamment simple
  - arbre de décision peu profond, régression parcimonieuse
  - *self-explaining classifier* (Alvarez Melis and Jaakoloa, 18)
  - classifieur de substitution ou *surrogate models*
    (Craven et Shavlik 96, Hara et Hayashi 16, Ribeiro et al. 16, Guidotti et al., 18)

Multiplicité des formes d'explications

- Conditions suffisantes : règles de décision
  - MES : sélection de la meilleure explication candidate par score d'information mutuelle par rapport au modèle à interpréter (Turner, 15)
  - LORE : extraction d'un arbre de décision appris sur le voisinage (Guidotti et al., 18)
  - explication par abduction (Ignatiev et al., 19)
- Vecteur d'importance d'attribut : *feature importance vector*
  - décroissance de précision quand on permute les valeurs d'un attribut (Breiman 01, Fisher et al. 19)
  - LIME : modèle linéaire local (Ribeiro et al. 16)
  - gradient du classifieur (Baehrens et al. 10;,Selvaraju et al. 16)
  - SHAP et ses variantes (Stumberlj et al. 09, Lundberg et Lee 17)

Multiplicité des formes d'explications

- Données particulières
  - prototype (Kim et al. 14)
  - données d'apprentissage les plus influentes, identifiées par réapprentissage (Kabra et al. 15, Sharchilev et al. 18)
  - explication contrefactuelle
    (Martens et Provost 14, Lash et al. 17, Wachter et al. 18, Guidotti, 22)

Trois axes de discussion classiques

(Doshi-Velez et Kim 17, Lipton 17, Guidotti et al. 18 Carvalho et al. 19)

- Auto-explication ou explication *post hoc*
- Hypothèses d'agnosticité
- Explications locales ou globales

Axe de discussion I

- Auto-explication : classifieur générant ses propres explications
  - p. ex. modèle linéaire, (petit) arbre de décision
  - souvent limité par taux de bonne classification
- Explication *post hoc*
  - distinction des étapes de prédiction et d'explication
  - p. ex. approximation de la frontière de décision par modèle simple

Axe de discussion II

- Hypothèses d'agnosticité : quelles connaissances disponibles ?
  - le classifieur, le type du classifieur
  - les données d'apprentissage, d'autres données, la distribution des données, des graphes causaux sur les attributs
- Rien : *model-agnostic* et *data-agnostic*
  - p. ex. LORE, Growing Spheres
  - approches flexibles qui respectent la confidentialité des données
  - risques d'explications moins pertinentes

Axe de discussion III

- Explications globales
  - fournir des connaissances sur le classifieur en général, son comportement global
  - p. ex. raisonnement à partir de cas et production de prototypes (Kim et al. 14)
- Explications locales (Guidotti et al. 18)
  - expliquer une prédiction en particulier
  - p. ex. LIME : approximation locale du classifieur au voisinage de la donnée considérée

Objectifs du cours XAI

- Ne pourra pas traiter toutes ces thématiques

- Cas principaux considérés
  - explication de classifieurs de données tabulaires :
    - expliquer la prédiction : pourquoi le modèle prédit C ?
  - explication de données tabulaires : résumé interprétable
  - approches numériques et approches logiques

- Organisation pratique
  - cours et TME
  - évaluation : certains TME, présentation d'articles, examen

Cours XAI 1 : Explications par exemples contre-factuels

ou *counterfactual explanations*

- 1. Principes
- 2. Méthodes fortement agnostiques
- 3. Limites et risques
- 4. Méthodes moins agnostiques
- 5. Propriétés souhaitables
- 6. Questions additionnelles

Principes

- Raisonnement contre-factuel : (Wikipedia)
  modifier en imagination l'issue d'un événement en modifiant l'une de ses causes
  - ex : si James Dean avait pris le train le jour de son accident de voiture, il ne serait pas mort
  - étudié du point de vue de la philosophie, la psychologie, les sciences cognitives
- Cas des explications (Bottou et al 13, Wachter et al. 18, Artelt et Hammer 19)
  - analyser les prédictions en envisageant des modifications susceptibles de changer les conclusions
  - considérer des variantes de la donnée x à expliquer
  - répondre à la question
    que faudrait-il changer pour avoir une prédiction différente ?

Principes

- Exemple typique :
  - décision = demande de crédit rejetée
  - **pourquoi ?**
  - explication : si le demandeur gagnait 500 euros de plus par mois et avait eu un accident de moins au cours des 3 dernières années, sa demande aurait été acceptée
- Modification de la question
  Pourquoi C ? $\longrightarrow$ **Pourquoi C et non C' ?**
- Notion d'explication **contrastive** (Miller 19)
  - en effet, explication différente selon C' :
    pourquoi l'appartement est-il cher et non abordable ?
    $\neq$ pourquoi l'appartement est-il cher et non exorbitant ?

Motivations pour les explications contre-factuelles

(Miller, 19)

- Motivations cognitives
  - mode de raisonnement naturel
  - aide à l'apprentissage : expliquer par l'exemple, en comparant
    - exemple : si l'animal avait des oreilles plus longues, ce serait un lièvre et non un lapin
- Motivations pratiques
  - donne des indications pratiques à l'utilisateur
  - actions à réaliser : particulièrement pertinent dans le cadre RGPD

Formalisation

- Etant donné un classifieur f et une donnée x
  - construire e tel que $f(e) \neq f(x)$ **en miminisant l'effort**
  - explication $= e - x$, changement **minimal** à apporter
- soit

$$e^* = \arg\min_{e \in \mathcal{X}} c_x(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

- A définir :
  - la fonction de coût c_x
  - l'espace de recherche pour e
  - la méthode d'optimisation, ou l'heuristique d'identification
- Multiplicité des approches
  - Guidotti 22 en cite une soixantaine !

Discussion : quelques exemples

$$e^* = \arg\min_{e \in \mathcal{X}} c(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

- Fonction de coût c
  - p. ex. distances l_2, l_1, l_0

    (Lash et al. 17, Wachter et al. 18, Guidotti et al. 18)
  - coût non uniforme sur tous les attributs
- Espace de recherche : $\mathcal{X}$ ou un sous-ensemble
  - $\mathcal{X} = \mathcal{X}_d \cup \mathcal{X}_i \cup \mathcal{X}_u$ (Lash et al. 17)
  - corrélation entre attributs, voire causalité
- Méthode d'optimisation suivant les hypothèses d'agnosticité
  - p. ex. méthodes efficaces pour classifieur linéaire

    (Ustun et al. 19, Russell et al. 19)
  - échantillonnage aléatoire (Laugel et al. 18)

Cours XAI 1 : Explications par exemples contre-factuels

- 1. Principes
- **2. Méthodes fortement agnostiques**
- 3. Limites et risques
- 4. Méthodes moins agnostiques
- 5. Propriétés souhaitables
- 6. Questions additionnelles

Méthode de référence : Wachter et al, 17

- Fonction de coût : composée de deux termes

$$c_x(e) = \lambda(f(e) - C')^2 + d(x, e)$$

  - $(f(e) - C')^2$: le résultat associé à e par le classifieur f doit être proche de la classe souhaitée C'
  - $d(x, e)$: le contre-factuel e doit être proche de x
  - λ : pondération relative des deux termes
    - maximiser λ en minimisant $c_x(e)$, sous la contrainte $|f(e) - C'| < \epsilon$

- Choix crucial de la distance d

Interlude : rappel sur les distances

- Vecteurs $\equiv$ points dans un espace vectoriel
  - $x = (x_i)_{i=1..p} \in \mathbb{R}^p$
- Distances
  - euclidienne l_2 $\quad d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
  - Manhattan l_1 $\quad d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$
  - Tchebychev $\quad d(x,y) = \max_{i=1..p} |x_i - y_i|$
  - Minkowski $\quad d_\gamma(x,y) = \left(\sum_{i=1}^{p} |x_i - y_i|^\gamma\right)^{\frac{1}{\gamma}}$

Interlude : rappel sur les distances

- Vecteurs $\equiv$ points dans un espace vectoriel
  - $x = (x_i)_{i=1..p} \in \mathbb{R}^p$
- Distances
  - Minkowski $\quad d_\gamma(x, y) = \left(\sum_{i=1}^{p} |x_i - y_i|^\gamma \right)^{\frac{1}{\gamma}}$

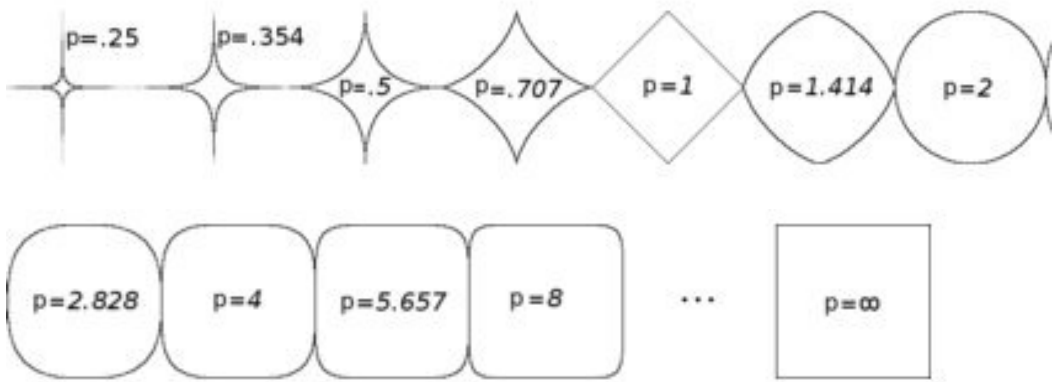


figure tirée de http://en.wikipedia.org/wiki/Minkowski_distance

Interlude : rappel sur les distances

- Vecteurs $\equiv$ points dans un espace vectoriel
  - $x = (x_i)_{i=1..p} \in \mathbb{R}^p$
- **Distances pondérées**

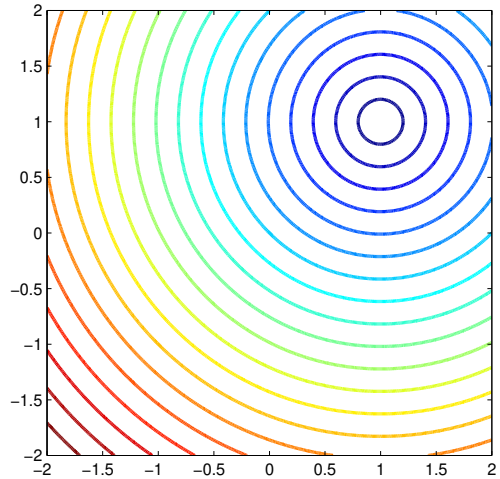
$$d(x, y) = \sqrt{(x - y)^t A (x - y)} = d(A^{1/2}x, A^{1/2}y)$$

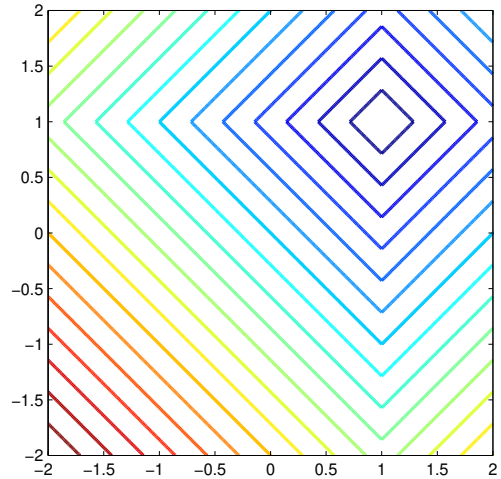
  - équivalente à une transformation des données
  - cas particulier : distance de Mahalanobis
    - Σ matrice de covariance des données

$$d(x, y) = \sqrt{(x - y)^t \Sigma^{-1} (x - y)}$$

Interlude : rappel sur les distances

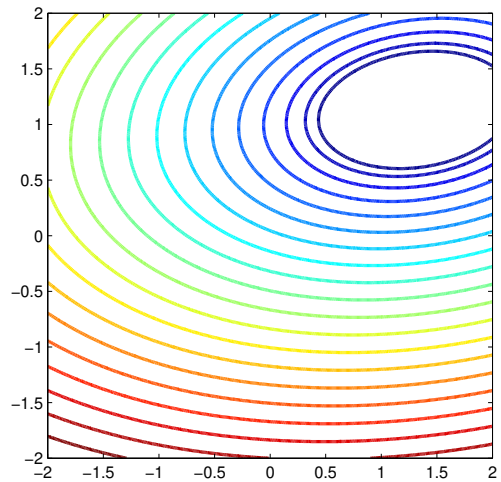
euclidienne

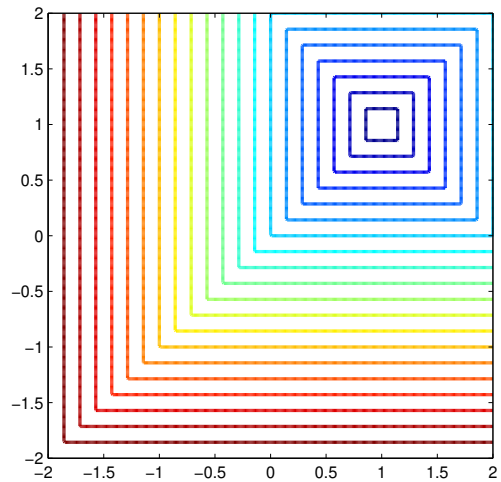




Manhattan

euclidienne pondérée





Tchebychev

Méthode de référence : Wachter et al, 17

- Fonction de coût : composée de deux termes

$$c_x(e) = \lambda(f(e) - C')^2 + d(x, e)$$

  - $(f(e) - C')^2$: le résultat associé à e par le classifieur f doit être proche de la classe souhaitée C'
  - $d(x, e)$: le contre-factuel e doit être proche de x
  - λ : pondération relative des deux termes
    - maximiser λ en minimisant $c_x(e)$, sous la contrainte $|f(e) - C'| < \epsilon$

- Choix crucial de la distance d : par exemple distance L1 normalisée

$$d(x, e) = \sum_{i=1}^{m} \frac{|x_i - e_i|}{MAD_i}$$

- Optimisation : descente de gradient si f est différentiable

Approche heuristique : Growing Spheres

(Laugel et al., 18)

$$e^* = \arg\min_{e \in \mathcal{X}} c(e) \qquad \text{tel que} \quad f(e) \neq f(x)$$

- Fonction de coût : combine deux distances

$$c_x(e) = ||e - x||_2 + ||e - x||_0$$

  - proximité et **concision** : minimiser le nombre d'attributs modifiés

- Heuristique d'optimisation séquentielle
  - minimisation de l_2 par une méthode de Monte Carlo

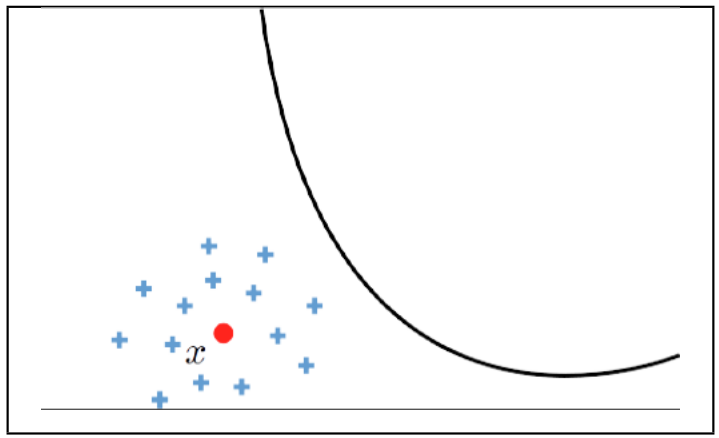
$$\tilde{e} \approx e^* = \arg\min_{z \in \mathcal{X}} \{||x - z||_2 \quad \text{tel que} \quad f(z) \neq f(x)\}$$

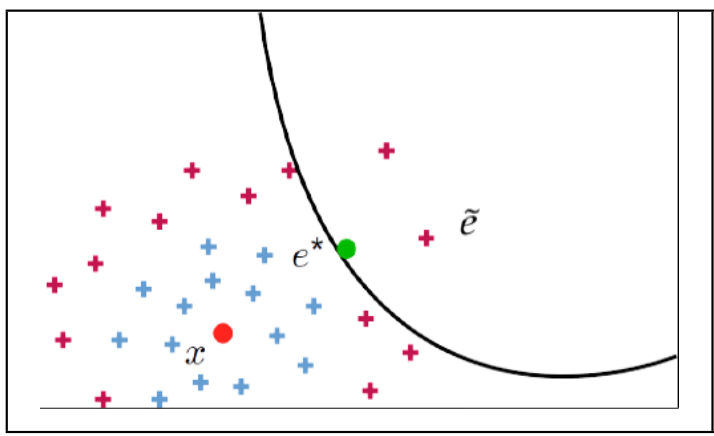
  - minimisation de l_0 : rendre le résultat intermédiaire $\tilde{e}$ parcimonieux

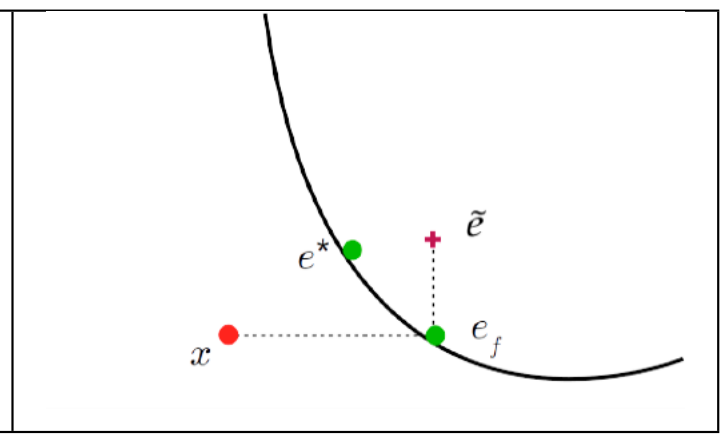
$$e_f = \arg\min_{e \in \mathcal{P}_{\tilde{e}}} ||e - x||_0 \qquad \text{tel que} \quad f(e) \neq f(x)$$

Approche heuristique : Growing Spheres

(Laugel et al., 18)







Cours XAI 1 : Explications par exemples contre-factuels

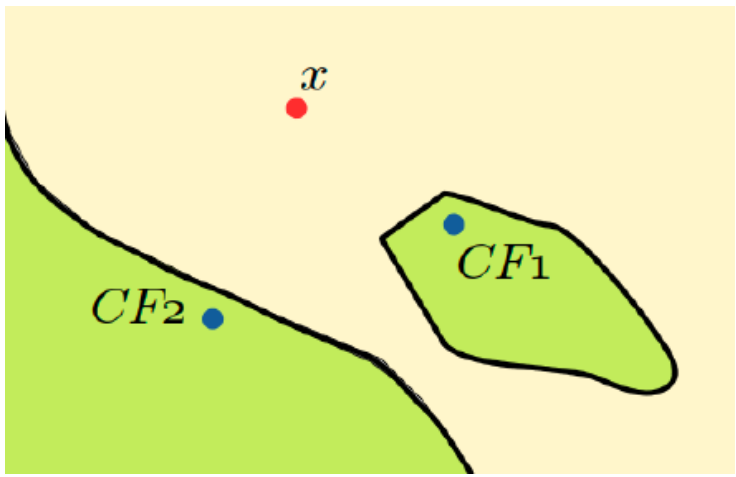
- 1. Principes
- 2. Méthodes fortement agnostiques
- **3. Limites et risques**
- 4. Méthodes moins agnostiques
- 5. Propriétés souhaitables
- 6. Questions additionnelles

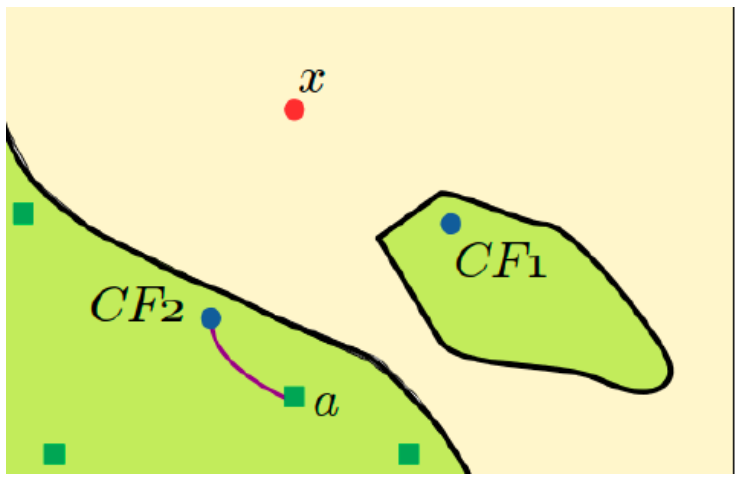
Risque d'explication non justifiée

(Laugel et al. 19)

- Explication contre-factuelle liée à des artefacts du classifieur
- Proposition d'une méthode de diagnostic
  - utilisant les données d'apprentissage X

$\Longrightarrow$ les méthodes existantes sont très sensibles à ce risque



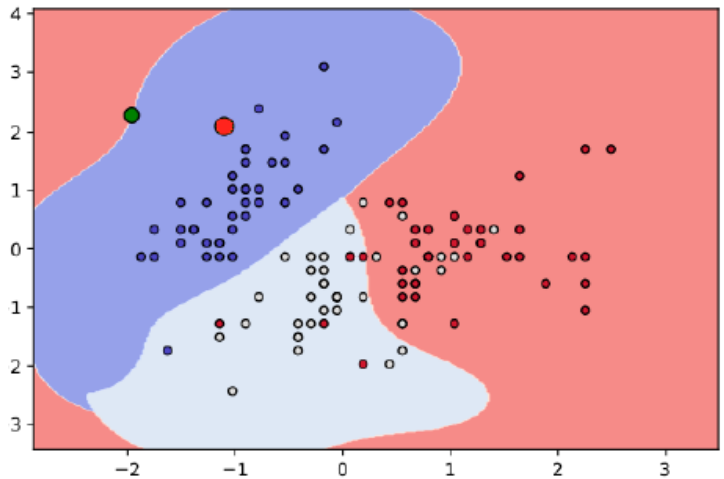


- Question : est-ce vraiment un comportement à éviter ?
  - vraie question sous-jacente, ardue : quelle explication pour un classifieur qui se trompe ?

Risque d'explication hors distribution

(Laugel et al. 19)

- Le classifieur peut "improviser" dans les zones où il ne dispose pas de données



- Risque d'explication contre-factuelle **non réaliste** : par exemple
  - valeur impossible : si le demandeur était âgé de 150 ans
  - combinaison de valeurs impossibles : si le demandeur avait (le même âge et) un niveau d'étude strictement supérieur

Au-delà : question d'actionnabilité

- Explication contre-factuelle **non actionnable**
  - changement non réalisable
  - exemple : si le demandeur avait 2 ans de moins
- Remarque : modification de la définition d'explication
  - ne vise plus à répondre à la question "Pourquoi P et pas P' ?"
  - mais "**Comment** avoir la prédiction P' ?"
  - l'objectif n'est plus de comprendre, mais de pouvoir agir : *algorithmic recourse* (Karimi et al, 20)
- Par défaut, risque de contre-factuel non actionnable
  - restriction de l'espace de recherche des candidats

Cours XAI 1 : Explications par exemples contre-factuels

- 1. Principes
- 2. Méthodes fortement agnostiques
- 3. Limites et risques
- **4. Méthodes moins agnostiques**
- 5. Propriétés souhaitables
- 6. Questions additionnelles

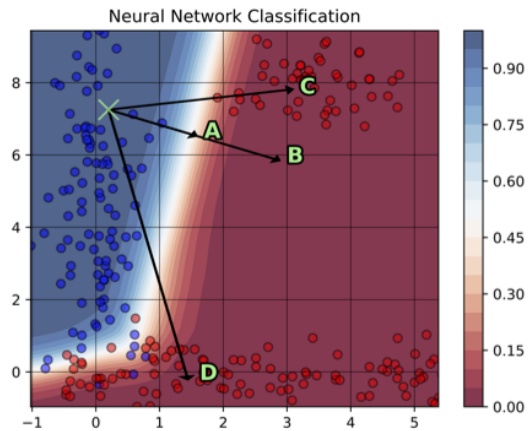
Cas où des données sont disponibles

$\Longrightarrow$ Apprendre des **informations sur la distribution** des données

- FACE : Feasible and Actionable Counterfactual Explanations

(Poyiadzi et al. 20)

  - trouver un chemin pondéré optimal entre x et e
  - passant par les données disponibles où la densité estimée est supérieure à un seuil
  - avec poids des arcs entre données $w_{ij} = p\left(\frac{x_i+x_j}{2}\right) d(x_i, x_j)$



Cas où des données sont disponibles

$\Longrightarrow$ Apprendre des **informations sur la distribution** des données

- Estimation de densité, intégrée à la fonction de coût

(Artelt et Hammer, 19)

$$e^* = \arg\min_{e \in \mathcal{X}} c_x(e) \qquad \text{tel que} \quad f(e) \neq f(x) \text{ et } \hat{p}_y(e) \geq \delta$$

- Apprentissage d'un graphe de causalité par VAE (Mahajan et al. 19)
  - équations structurelles causales : dépendances entre les valeurs d'attributs
  - **score causal** pénalisant les valeurs d'attributs non compatibles
  - définition d'une nouvelle mesure de distance

Cours XAI 1 : Explications par exemples contre-factuels

- 1. Principes
- 2. Méthodes fortement agnostiques
- 3. Limites et risques
- 4. Méthodes moins agnostiques
- **5. Propriétés souhaitables**
- 6. Questions additionnelles

Propriétés souhaitables

- Validité : $f(e^*) \neq f(x)$
- Proximité : minimalité en terme de distance $d(x, e^*)$
- Parcimonie : minimalité en terme de nombre d'attributs modifiés
- Plausibilité, faisabilité : avoir des valeurs réalistes
- Actionabilité : faisabilité des changements

- Stabilité : instances similaires associées à des explications similaires
- Temps de calcul

Et pour citer d'autres questions

- Questions de **présentation des résultats**
  - parfois trop techniques, tournés vers des utilisateurs experts en IA
  - domaine des **interfaces explicatives**
  - possibilité d'expression linguistique
    interprétabilité de "si le demandeur gagnait 317.62 euros de plus par mois" ?
- Questions d'**évaluation** : nécessité de mettre l'utilisateur au cœur
  - a-t-il compris ? quel est l'"objectif pédagogique" ?
  - sait-il quoi faire ? sans le pousser à la fraude
- De telles approches vont-elles rétablir la confiance des utilisateurs ou donner des outils pour les manipuler ?