XAI
eXplainable Artificial Intelligence
IA explicable

Cours 2 - mardi 26 septembre 2023

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

LIP6, Sorbonne Université

Au programme du jour

- 1. Résumé de l'épisode précédent
- 2. Un point sur les opérateurs d'agrégation
- 3. Agrégation et exemples contre-factuels
- 4. Exemples contre-factuels divers
- 5. Zoom sur DiCE

Définition des exemples contre-factuels

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

$$e^* = \arg\min_{e \in \mathcal{E}} c_x(e) \qquad \mathcal{E} \subseteq \{e / f(e) \neq f(x)\}$$

- 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 !

Fonction de coût : composantes

- **Minimalité du changement**: proximité entre e et x
  - distance l_2, éventuellement pondérée
- **Parcimonie du changement** : faible nombre d'attributs modifiés
  - distance l_0
- **Contextualisation / autres données** : réalisme
  - maximiser $p(e)$ ou $p_y(e)$ ou tout le chemin
  - existence d'ue justification par des données d'apprentissage
  - minimiser coût de reconstruction de e par auto-encodeur
- **Contextualiser / utilisateur** : personnalisation
  - maximiser l'actionnabilité / ensemble d'attributs modifiables
  - utiliser des attributs compréhensibles
  - vérifier les contraintes causales

$\Longrightarrow$ **Comment combiner ces critères ?**

Au programme du jour

- 1. Résumé de l'épisode précédent
- **2. Un point sur les opérateurs d'agrégation**
  - avec exercices TD
- 3. Agrégation et exemples contre-factuels
- 4. Exemples contre-factuels divers
- 5. Zoom sur DiCE

Définition générale

- Définition générale : $Agg : \bigcup_{n \in \mathbb{N}} [0,1]^n \to [0,1]$
  - identité si unique : $Agg(u) = u$
  - conditions aux limites : $Agg(0, \ldots, 0) = 0$ et $Agg(1, \ldots, 1) = 1$
  - monotonie croissante
    $Agg(u_1, \ldots, u, \ldots, u_n) \leq Agg(u_1, \ldots, v, \ldots, u_n)$ si $u \leq v$

- Abondante littérature : par exemple
  - Calvo, Mayor, Mesiar. *Aggregation operators*. Springer. 2002
  - Detyniecki. *Fundamentals on aggregation operators*. 2001
  - Grabisch, Marichal, Mesiar, Pap. *Aggregation Functions*. Number 127 in Encyclopedia of Mathematics and its Applications. 2009

- Quatre grandes familles

Opérateurs conjonctifs

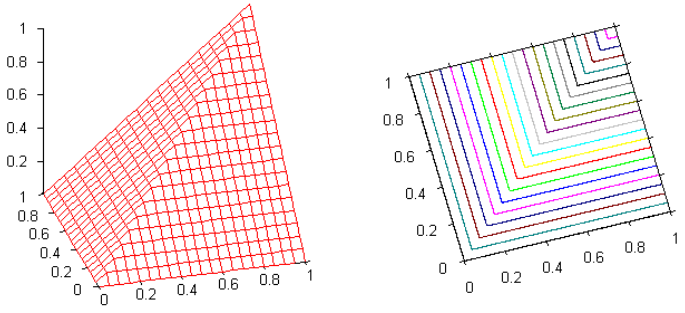
- Attitude sévère
  - valeur élevée ssi u **et** v ont des valeurs élevées

- Exemples :
  - $Agg(u, v) = \min(u, v)$
  - $Agg(u, v) = u \cdot v$
  - $Agg(u, v) = \max(u + v - 1, 0)$

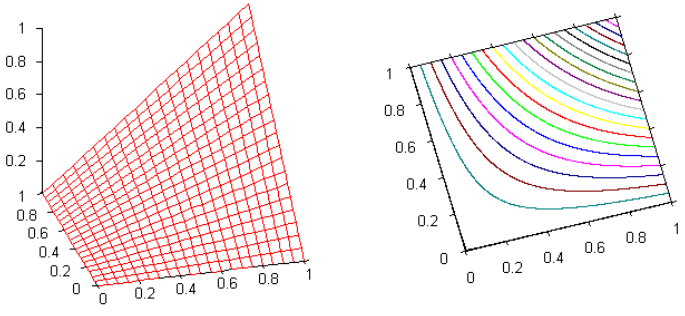
- Exercice : tracer les lignes de niveau de ces opérateurs
  - $L_t = \{(u, v)/Agg(u, v) = t\}$

Opérateurs conjonctifs

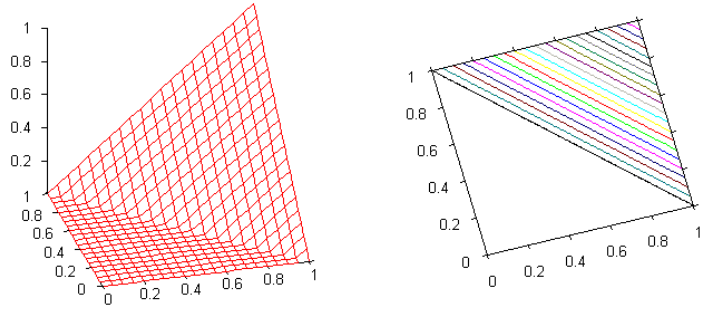
- $Agg(u, v) = \min(u, v)$



- $Agg(u, v) = u \cdot v$



- $Agg(u, v) = \max(u+v-1, 0)$



Opérateurs disjonctifs

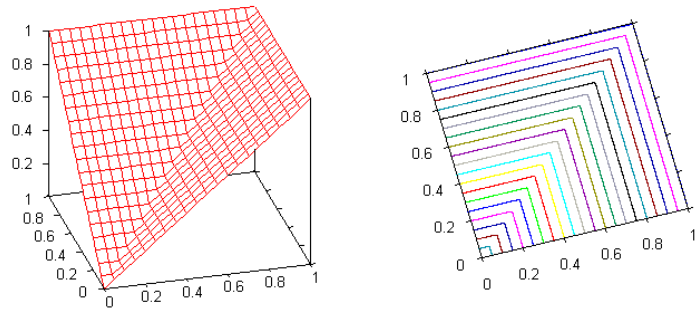
- Attitude tolérante
  - valeur élevée ssi u **ou** v ont des valeurs élevées

- Exemples :
  - $Agg(u, v) = \max(u, v)$
  - $Agg(u, v) = u + v - u \cdot v$
  - $Agg(u, v) = \min(u + v, 1)$

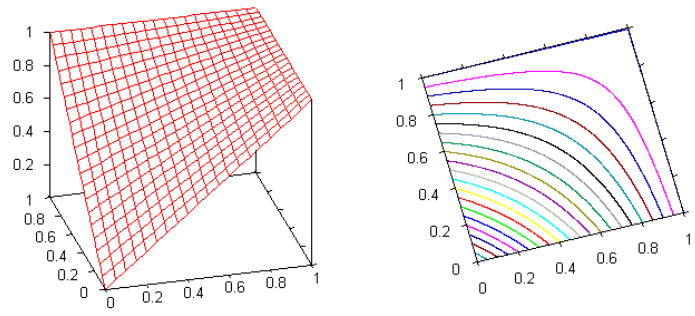
- Exercice : lignes de niveau ?

Opérateurs disjonctifs

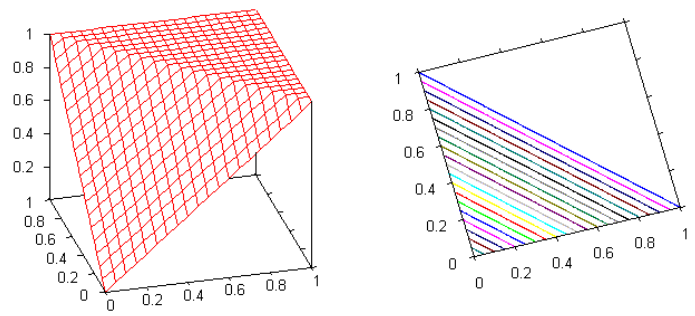
- $Agg(u, v) = \max(u, v)$



- $Agg(u, v) = u + v - u \cdot v$



- $Agg(u, v) = \min(u + v, 1)$



Dualité opérateurs conjonctifs-disjonctifs

$$AggDisj = 1 - AggConj(1-u, 1-v)$$

- Exercice : à montrer pour les couples

- $Agg(u,v) = \min(u,v)$
- $Agg(u,v) = u \cdot v$
- $Agg(u,v) = \max(u+v-1, 0)$
- $Agg(u,v) = \max(u,v)$
- $Agg(u,v) = u + v - u \cdot v$
- $Agg(u,v) = \min(u+v, 1)$

Opérateurs de compromis

- Autorise la **compensation**
- Exemples :
  - **moyenne arithmétique** : $Agg(x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - variantes pondérées : $Agg(x_1, \cdots, x_n) = \sum_{i=1}^{n} w_i x_i$
- Lignes de niveaux ?

Opérateurs de compromis ordinaux

- Principe des **Ordered Weighted Average, OWA**
  - **les poids dépendent de l'ordre, non des attributs**
  - σ permutation telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$
  - $OWA_w(x_1, \cdots x_n) = \sum_{i=1}^{n} w_i x_{\sigma(i)}$
- Exemple : $w_1 = \frac{9}{10}$, $w_2 = \frac{1}{10}$

$$OWA(u,v) = \begin{cases} \frac{9}{10}u + \frac{1}{10}v & \text{si } u \leq v \\ \frac{9}{10}v + \frac{1}{10}u & \text{sinon} \end{cases}$$

- Exercices
  - lignes de niveaux
  - $\max = OWA_w$ pour quel w ?
  - $\min = OWA_w$ pour quel w ?

Encore plus expressifs

- **Intégrales de Choquet**
  - pour la prochaine séance

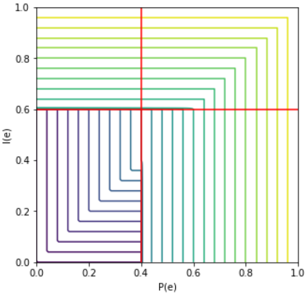
Bilan des opérateurs d'agrégation par compromis

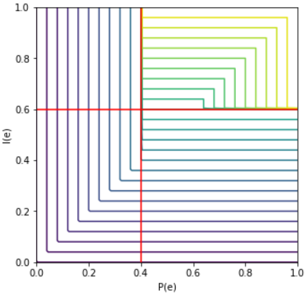
- Famille très riche et expressive !

- Une dernière famille : opérateurs à attitude variable
  - conjonctifs, disjonctifs ou de compromis
    suivant les valeurs à agréger

Opérateurs à attitude variable

- Exemple : intégrales de Gödel





$G_\mu^{\otimes}(u,v)$ $\qquad\qquad$ $G_\mu^{\rightarrow}$

$$= \begin{cases} \min(u,v) & \text{si } u \leq 1-\alpha_u \text{ et } v \leq 1-\beta_v \\ \max(u,v) & \text{si } u > 1-\alpha_u \text{ et } v > 1-\beta_v \\ u & \text{si } u > 1-\alpha_u \text{ et } v \leq 1-\beta_v \qquad // \qquad v \\ v & \text{si } u \leq 1-\alpha_u \text{ et } v > 1-\beta_v \qquad // \qquad u \end{cases}$$

Caractérisations théoriques

- Propriétés éventuelles d'intérêt
  - associativité :
    $Agg(u, v, w) = Agg(Agg(u, v), w) = Agg(u, Agg(v, w))$
  - symétrie : $Agg(u, v) = Agg(v, u)$
  - élément absorbant : $Agg(u, k) = k$
  - élément neutre : $Agg(u, v, k) = Agg(u, v)$
  - idempotence : $Agg(u, u) = u$
  - compensation : $\min(u, v) \leq Agg(u, v) \leq \max(u, v)$
  - contre-effet : $\forall t, \forall u, \forall v \exists w \; Agg(u, v, w) = t$
  - renforcement : $(u \geq k \wedge v \geq k) \Rightarrow Agg(u, v) \geq \max(u, v)$
  - ...

Au programme du jour

- 1. Résumé de l'épisode précédent
- 2. Un point sur les opérateurs d'agrégation
- **3. Agrégation et exemples contre-factuels**
- 4. Exemples contre-factuels divers
- 5. Zoom sur DiCE

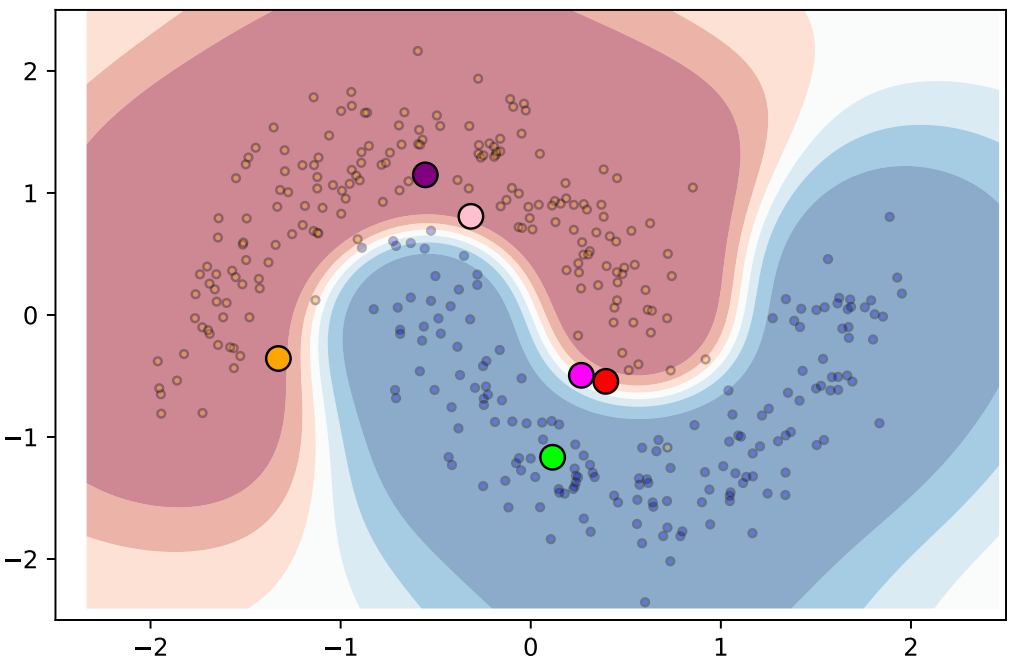
Choix classiques

- Critères de qualité : tous souhaitables mais souvent incompatibles
  - exemple : maximiser la parcimonie et la proximité

  $\Longrightarrow$ pas d'agrégation conjonctive
- **Compromis explicite** : moyenne pondérée (Mahajan et al., 19)
  - problème 1 : choix des poids, évidemment
  - problème 2 : gestion de la commensurabilité des critères
- **Compromis implicite** : priorité entre les critères
  - optimisation sous contrainte :
    définition de $\mathcal{E} = \{e / f(e) \neq f(x) \wedge p(e) > \eta\}$
    (Artelt et Hammer, 20 ; FACE, Poyiadzi et al., 20 ; Ustun et al., 19)
  - heuristique :
    Growing Spheres : "optimise" la parcimonie après la proximité

Importance du problème d'agrégation

- **Influence sur les résultats obtenus, évidemment**
  - illustration avec les `half-moons`
  - x : point vert, classifieur f : SVM
  - $c_x(e) = agg(d_2(e,x), -\log p_y(e))$, p_y Gaussian KDE



- rouge : proximité seule
- violet : densité seule
- rose : moyenne pondérée
- magenta : proximité prioritaire
- orange : densité prioritaire

Importance du problème d'agrégation

- Influence sur les résultats obtenus, évidemment
- **Conséquences indésirables**
  - compromis non souhaitable : aucun des critères n'est satisfait
  - effet compris par l'utilisateur ?
    - ni le respect de la causalité, ni la proximité ne sont satisfaits
    - l'un vient aux dépens de l'autre

$\Longrightarrow$ **Problème aussi crucial que le choix des critères eux-mêmes**

Au programme du jour

- 1. Résumé de l'épisode précédent
- 2. Un point sur les opérateurs d'agrégation
- 3. Agrégation et exemples contre-factuels
- **4. Exemples contre-factuels divers**
- 5. Zoom sur DiCE

Exemples contre-factuels multiples

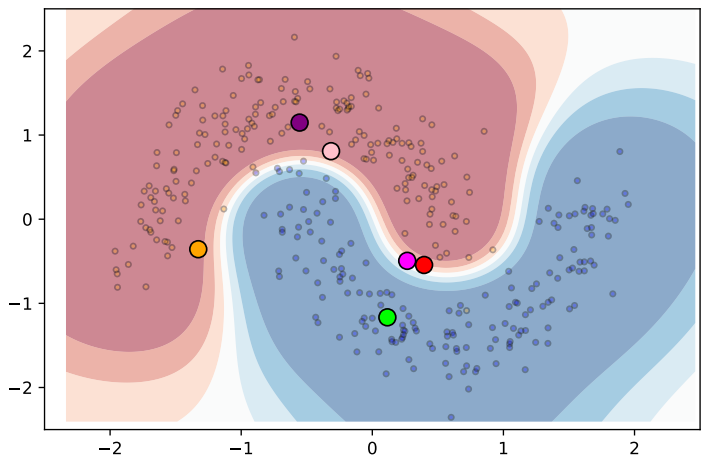
- Principe
  - ne pas choisir un des exemples
  - mais **en générer plusieurs**

Exemples contre-factuels multiples

- **Motivation informatique** : solution au problème d'agrégation
  - problème d'optimisation multi-critères
  - notion de front de Pareto : candidats non dominés
- **Motivations cognitives** : effet sur la compréhension
  - cadre médical : plusieurs explications aident à meilleur diagnostic (Wang et al. 2019)
  - cadre éducatif : unique analogie $\rightarrow$ risque d'idée fausse (Spiro et al., 1989)
  - ce que dit Miller : permettre de choisir une explication préférée parmi un ensemble d'explications plausibles (Miller, 2019)
- **Motivation pratique** : s'adapter à un besoin inconnu
  - offrir plus de flexibilité, et ainsi augmenter les chances de satisfaction de l'utilisateur
  - personnalisation de l'explication

Exemples contre-factuels divers

- De la multiplicité...
  - ne pas choisir un des exemples
  - mais en générer plusieurs
- ... à la **diversité**
  - les multiples exemples ne doivent pas être redondants
  - mais différer les uns des autres



- Plein de façons de définir la diversité !
  - pour la prochaine séance

Au programme du jour

- 1. Résumé de l'épisode précédent
- 2. Un point sur les opérateurs d'agrégation
- 3. Agrégation et exemples contre-factuels
- 4. Exemples contre-factuels divers
- **5. Zoom sur DiCE**

Diverse Counterfactual Explanations: DiCE

- Référence :

Ramaravind K. Mothilal, Amit Sharma and Chenhao Tan.
Explaining machine learning classifiers through diverse counterfactual explanations.
Proc. of the Int. Conf. on Fairness, Accountability, and Transparency, FAT 20*, pp. 607-617. 2020

Diverse Counterfactual Explanations: DiCE

- Paramètres
  - f : un classifieur
  - x : requête, donnée dont la prédiction est à expliquer
  - k : nombre de contre-factuels souhaités
  - λ_1, λ_2 : poids des termes
- Fonction de coût

$$\{e_1^*, \ldots, e_k^*\} = \arg\min_{e_1,\ldots,e_k} \frac{1}{k}\sum_{i=1}^{k} yloss(f(e_i), f(x)) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(e_i, x) - \lambda_2\, div(e_1, \ldots, e_k)$$

- Ajout de parcimonie

Détails des termes

$$\frac{1}{k}\sum_{i=1}^{k} yloss(f(e_i), f(x))$$

- **Validité** des exemples contre-factuels
  - prédits d'une classe autre que la requête x

- Choix de $yloss$
  - $|f(e) - f(x)|$ ou $(f(e) - f(x))^2$ trop contraignants
  - *hinge-loss* : cas où classe souhaitée $= 1$
    - 0 si $f(e) > t \geq 0.5$
    - proportionnelle à $f(e) - f(x)$ si $f(e) \in [0.5, t]$
    - pénalité élevée si $f(e) < 0.5$ (classe prédite $= 0$)

Détails des termes

$$\frac{\lambda_1}{k} \sum_{i=1}^{k} dist(e_i, x)$$

- **Proximité** des exemples contre-factuels à la requête
- Attributs continus

$$dist(e, x) = \sum_{l=1}^{d} \frac{|e_l - x_l|}{MAD_l}$$

  - MAD_l : median absolute deviation calculée sur les données d'apprentissage

- Attributs catégoriels : 1 si valeur différente

Détails des termes

$$div(e_1, \ldots, e_k) = det(K) \text{ avec } K_{ij} = \frac{1}{1 + dist(e_i, e_j)}$$

- **Diversité** des exemples contre-factuels
  - même distance que pour la proximité

- avec ajout de petites perturbations aux termes diagonaux pour éviter les déterminants mal définis

Optimisation et post-traitement

- Pour les modèles de sklearn
  - tirage aléatoire
  - algorithme génétique
  - recherche par kd-tree
- Pour les modèles de tensorflow et pytorch
  - descente de gradient

- **Parcimonie**
  - comme Growing Spheres
  - $e'_i = x_i$ itérativement tant que $f(e') \neq f(x)$

Contraintes additionnelles de faisabilité

- **Intégration de connaissances** des utilisateurs
- Attributs non actionnables
  - liste d'attributs dont les valeurs ne peuvent pas être modifiées
- Plages de variation des attributs
  - intervalles de valeurs associées à chaque attribut actionnable
- Extension : graphe de causalité des valeurs d'attributs $\mathcal{C}$
  - filtrer les candidats qui ne vérifient pas ces contraintes
  - intégration dans la fonction d'optimisation (Mahajan et al, NeurIPS19)

$$dist(e,x) = \sum_{i \in \mathcal{U}} d(e_i, x_i) + \sum_{j \in \mathcal{V}} d(e_j, \mathcal{C}(e_{jp_1}, \ldots, e_{jp_m}))$$