XAI
eXplainable Artificial Intelligence
IA explicable

Cours 3 - mardi 3 octobre 2023

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

LIP6, Sorbonne Université

Au programme du jour

- 1. Un petit rab sur les opérateurs d'agrégation
- 2. Diversité des exemples contre-factuels
- 3. Le fameux LIME

Rappel : 4 familles principales

- Opérateurs conjonctifs
  - valeur élevée ssi u **et** v ont des valeurs élevées
  - exemples : $\min(u, v)$, $u \cdot v$, $\max(u + v - 1, 0)$
- Opérateurs disjonctifs
  - valeur élevée ssi u **ou** v ont des valeurs élevées
  - exemples : $\max(u, v)$, $u + v - u \cdot v$, $\min(u + v, 1)$
- Operateurs de compromis : famille très riche et expressive
  - autorisent la **compensation**
  - exemples : moyenne arithmétique (pondérée), OWA
- Opérateurs à attitude variable
  - **conjonctifs, disjonctifs ou de compromis** suivant les valeurs à agréger
  - exemple : intégrales de Gödel

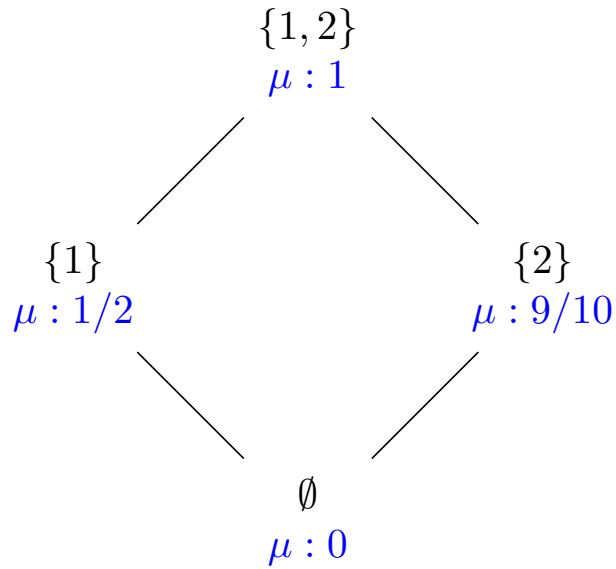
Opérateurs de compromis

- Moyenne arithmétique pondérée : $Agg(x_1, \cdots, x_n) = \sum_{i=1}^{n} w_i x_i$
- Approche ordinale : Ordered Weighted Average, OWA
  - **les poids dépendent de l'ordre, non des attributs**
  - σ permutation telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$
  - $OWA_w(x_1, \cdots x_n) = \sum_{i=1}^{n} w_i x_{\sigma(i)}$
- Encore plus expressifs : **intégrales de Choquet**
  - les poids dépendent de l'ordre **et** des attributs

Encore plus expressifs

- **Intégrales de Choquet**
  - les poids dépendent de l'ordre **et** des attributs
  - σ permutation telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$
  - $A_{\sigma(i)} = \{\sigma(i), \ldots, \sigma(n)\}$

- Capacité : $\mu : 2^{\{1,\ldots,n\}} \rightarrow [0,1]$
  - croissante + conditions aux limites
  - poids de tout sous-ensemble de critères

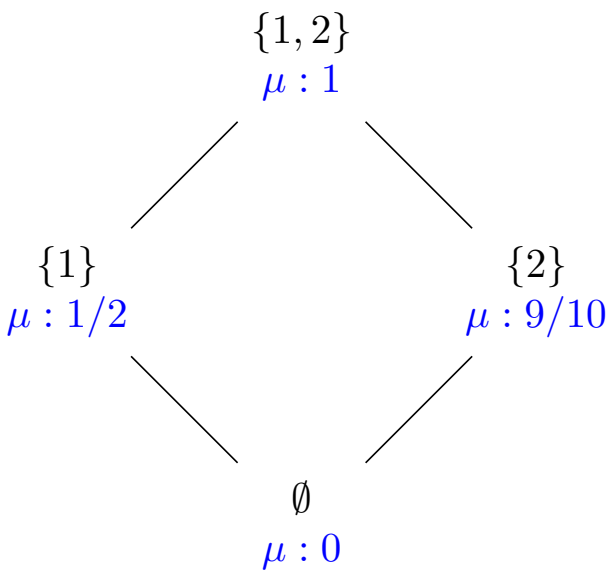


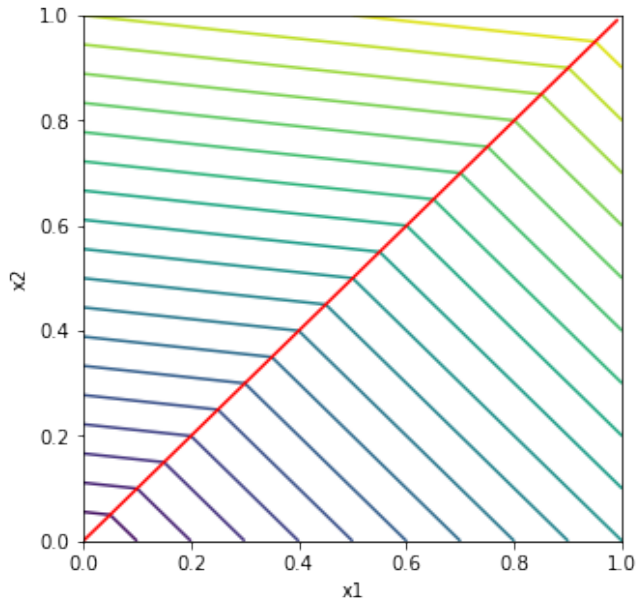
- Définition générale

$$C_\mu(x) = \sum_{i=1}^{n} \left(\mu(A_{\sigma(i)}) - \mu(A_{\sigma(i+1)})\right) x_{\sigma(i)}$$

Intégrales de Choquet

$$C_\mu(x) = \sum_{i=1}^{n} \left(\mu(A_{\sigma(i)}) - \mu(A_{\sigma(i+1)})\right) x_{\sigma(i)}$$

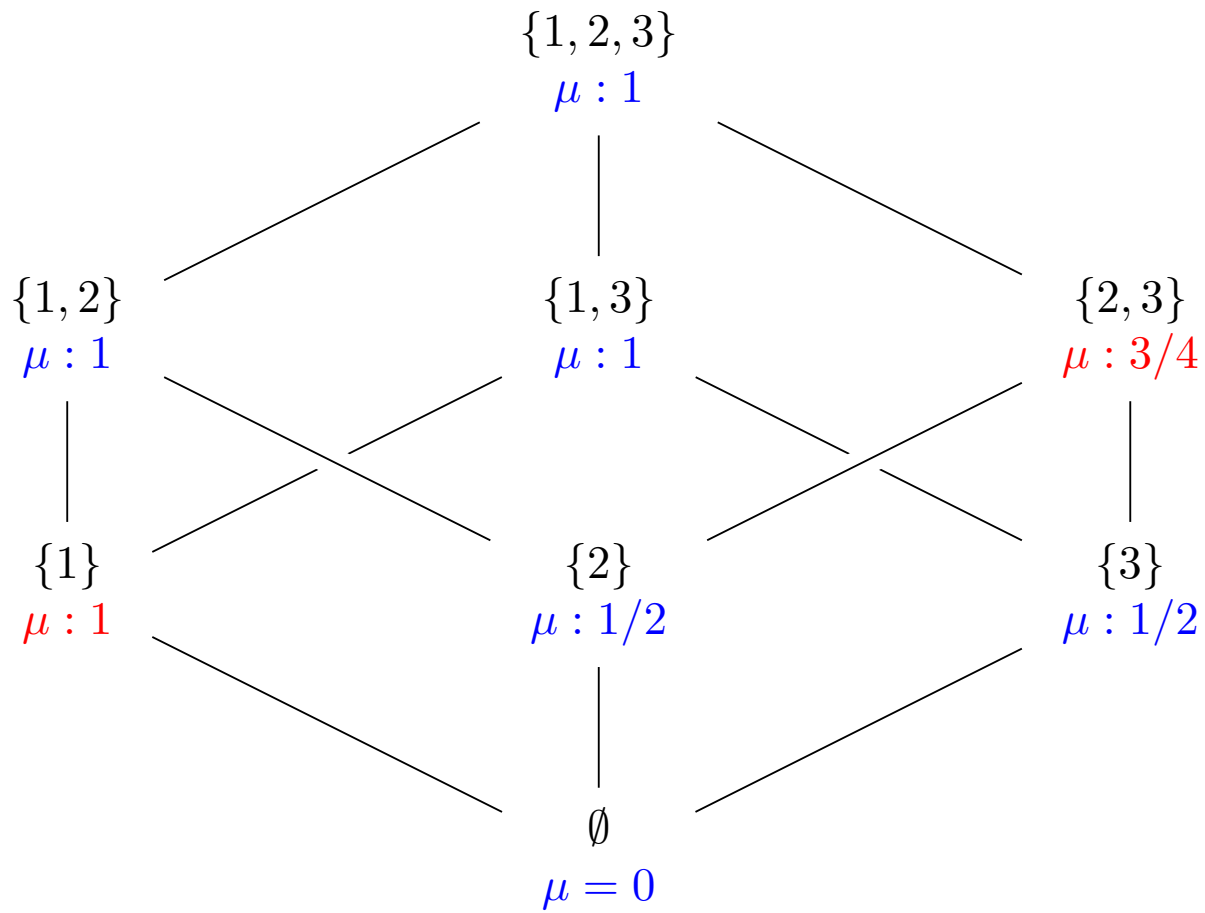




$$C_\mu(x) = \begin{cases} \text{if } x_1 \geq x_2 \\ \quad \frac{1}{2}(x_1 + x_2) \\ \text{else} \\ \quad \frac{1}{10}(9x_2 + x_1) \end{cases}$$

Intégrales de Choquet

- Cas à 3 attributs



OWA et intégrales de Choquet

- Notations
  - σ permutation telle que $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$
  - $A_{\sigma(i)} = \{\sigma(i), \ldots, \sigma(n)\}$

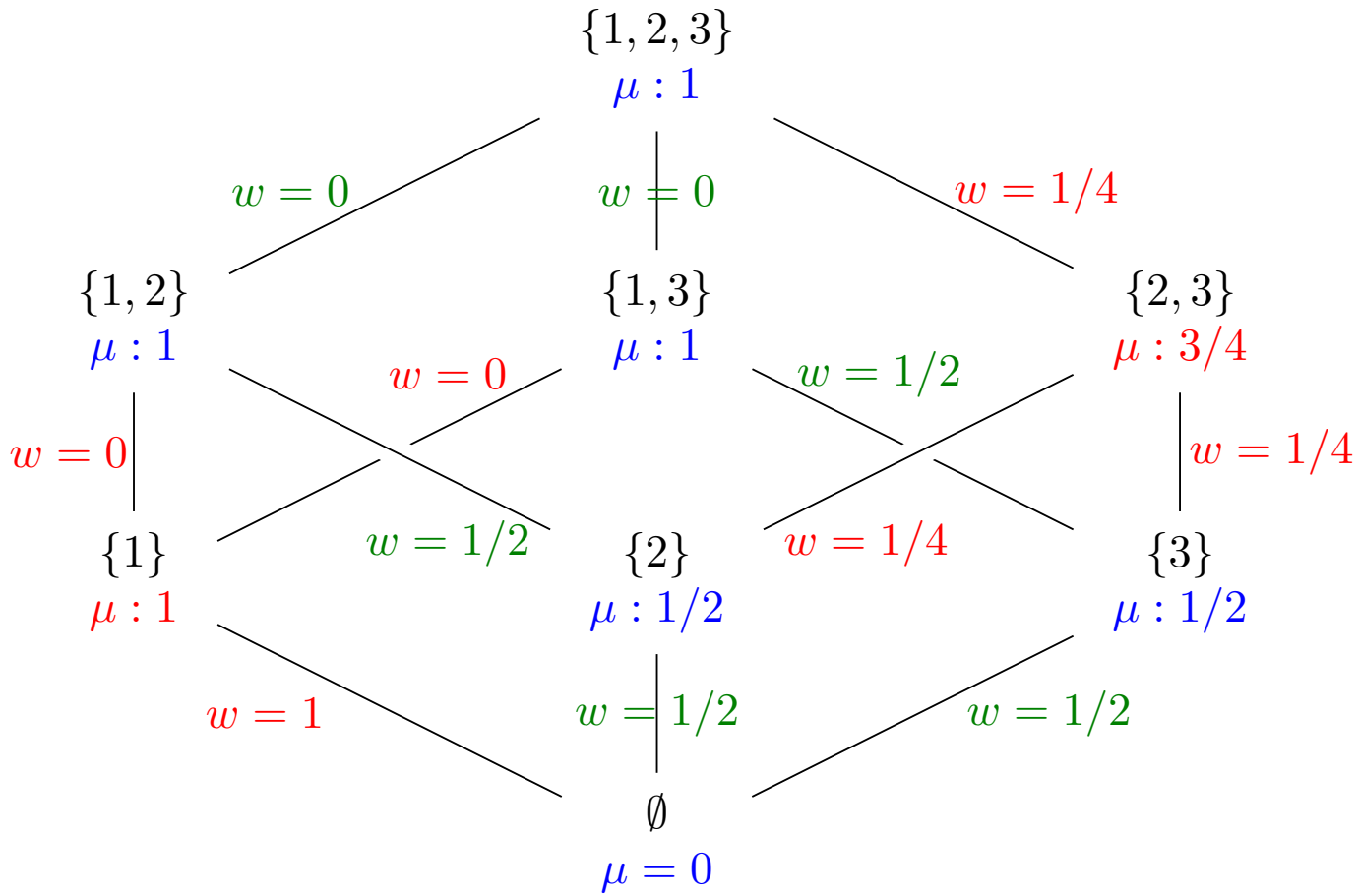
$$OWA_w(x) = \sum_{i=1}^{n} w_i x_{\sigma(i)}$$

$$C_\mu(x) = \sum_{i=1}^{n} \left(\mu(A_{\sigma(i)}) - \mu(A_{\sigma(i+1)})\right) x_{\sigma(i)}$$

- Exercice :

  à quelles conditions sur μ une intégrale de Choquet est-elle un OWA ?
  - réponse : $\mu(E)$ dépend seulement de $|E|$

OWA et intégrales de Choquet

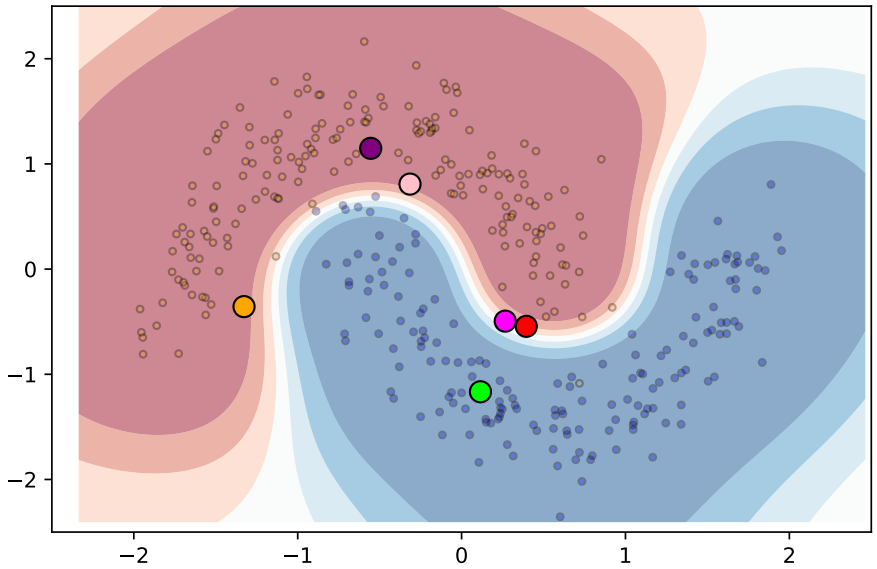


Au programme du jour

- 1. Un petit rab sur les opérateurs d'agrégation
- **2. Diversité des exemples contre-factuels**
- 3. Le fameux LIME

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



Diversité des critères

- Différentes pondérations de la moyenne pondérée (Dandl et al., 2020)
  - $\leftrightarrow$ différentes positions sur le front de Pareto

Diversité des attributs/actions

- Définition explicite de sous-espaces distincts

(Carreira et al., 21 ; Rodriguez et al., 21)

  - puis identification d'exemples contre-factuels dans chacun
  - ex : partition donnée par l'utilisateur de certains attributs

- Distance deux à deux entre les exemples contre-factuels
  - modification du problème d'optimisation

$$\{e_1^*, \ldots, e_k^*\} = \arg \min_{\{e_1,\ldots,e_k\} \subset \mathcal{E}} \left(agg \left(\sum_{i=1}^{k} c_x(e_i), \varphi(div(\{e_1, \ldots, e_k\}) \right) \right)$$

div : varier les positions $\rightarrow l_2$
ou les attributs utilisés $\rightarrow l_0$

  - génération itérative (Singh Hada and Carreira, 21 ; Russell, 19)

- Diversité des actions
  - différence d'interprétation : en terme de "recours algorithmique"

$\Longrightarrow$ p. ex. diversité des directions de modifications

Obtention de la diversité

- Génération directe ou itérative
  - optimisation ou exploration simultanée dans plusieurs directions
  - contrainte d'éloignement des générations antérieures
- Diversité explicite vs implicite
  - inclusion dans la fonction de coût : mécanisme dédié pour la génération d'explications diverses
  - non-déterminisme comme source de diversité

  (Mahajan et al., 19 ; Sharma et al, 20)
- Nombre d'exemples contre-factuels renvoyés
  - au choix de l'utilisateur
  - limité par la méthode elle-même (Becker et al. 21 ; Guidotti et al, 19)

Au programme du jour

- 1. Un petit rab sur les opérateurs d'agrégation
- 2. Diversité des exemples contre-factuels
- **3. Le fameux LIME:**
  Local Interpretable Model-agnostic Explanations

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
Proc. of the 22nd ACM SIGKDD Int. Conf.
on Knowledge Discovery and Data mining,
KDD 2016

LIME : caractéristiques

- Même objectif que les exemples contre-factuels
  - approche **post-hoc** : étant donné un modèle $f : \mathcal{X} \to \mathcal{Y}$
  - approche **locale** : étant donné une instance $x \in \mathcal{X}$

  $\Longrightarrow$ expliquer la prédiction $f(x)$

- Autre famille d'explications : à la fois
  - approche par substitution : *surrogate model*
  - score d'importance d'attribut locale : *local feature importance*

Approches par substitution

$$e(x) = \arg\min_{g \in \mathcal{G}} \big(L(f, g, \pi_x) + \Omega(g) \big)$$

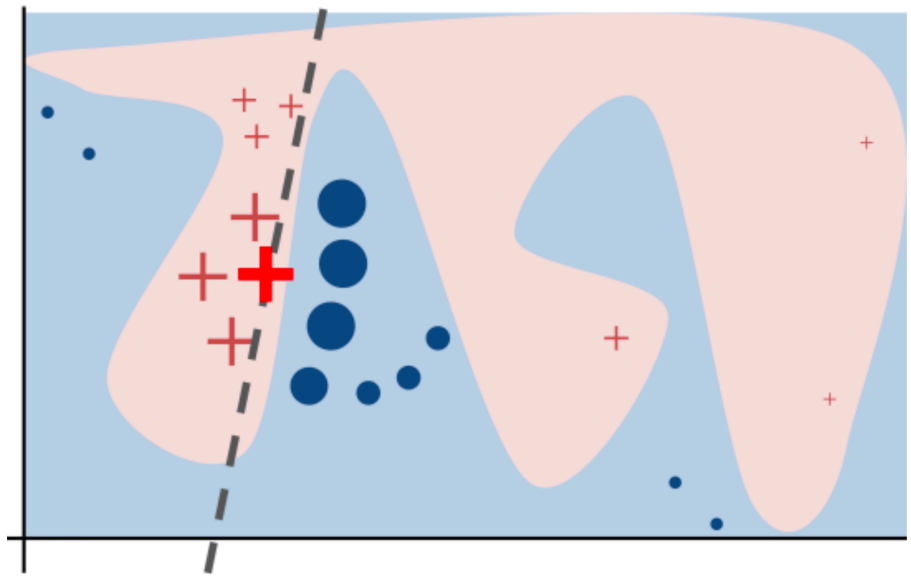
- Entrées : $f : \mathcal{X} \to \mathcal{Y}$ modèle à expliquer, x instance considérée
- $\mathcal{G}$: famille de modèles de substitution = *surrogate models*
  - contrainte : modèles interprétables
  - ex. : régression linéaire, arbres de décision de profondeur limitée, ...
- L : mesure de fidélité
  - g **doit être fidèle à** f
  - π_x : voisinage autour de x utilisé pour l'interprétation
- Ω : mesure de complexité
  - g **doit être simple**
  - ex: profondeur des arbres/ parcimonie des coefficients de régression

Cas de LIME

- f: modèle de classification probabiliste ou de régression
- $\mathcal{G} = \left\{ g(x) = \sum_{i=1}^{d} w_i x_i \right\}$: fonctions linéaires
  - explication : vecteur des w_i
  - interprétés comme scores d'importance locale
- L : moindres carrés sur $\mathcal{Z} = (z_p, f(z_p))_{p=1..m}$
  - $L(f, g, \pi_x) = \sum_{p=1}^{m} \pi_x(z_p)(f(z_p) - g(z_p))^2$
- $\Omega(g) = \sum_{i=1}^{d} w_i^2$: norme L2 des poids w_i
  - favoriser les poids faibles
- + étape de construction d'attributs interprétables

Cas de LIME : illustration

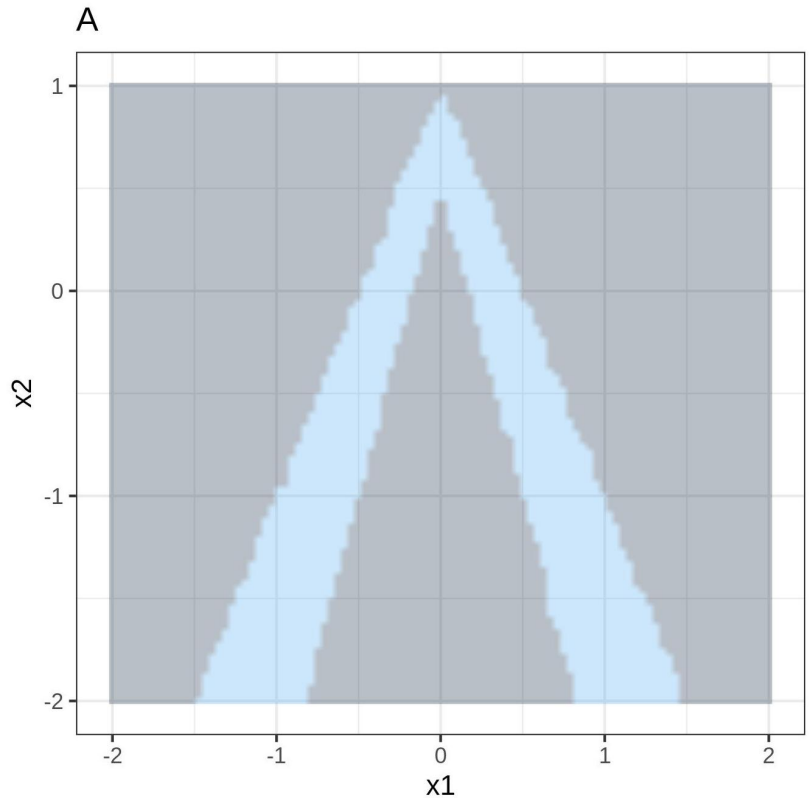
(Ribeiro et al, 2016)



Procédure d'apprentissage : illustration

(Molnar, 2021)

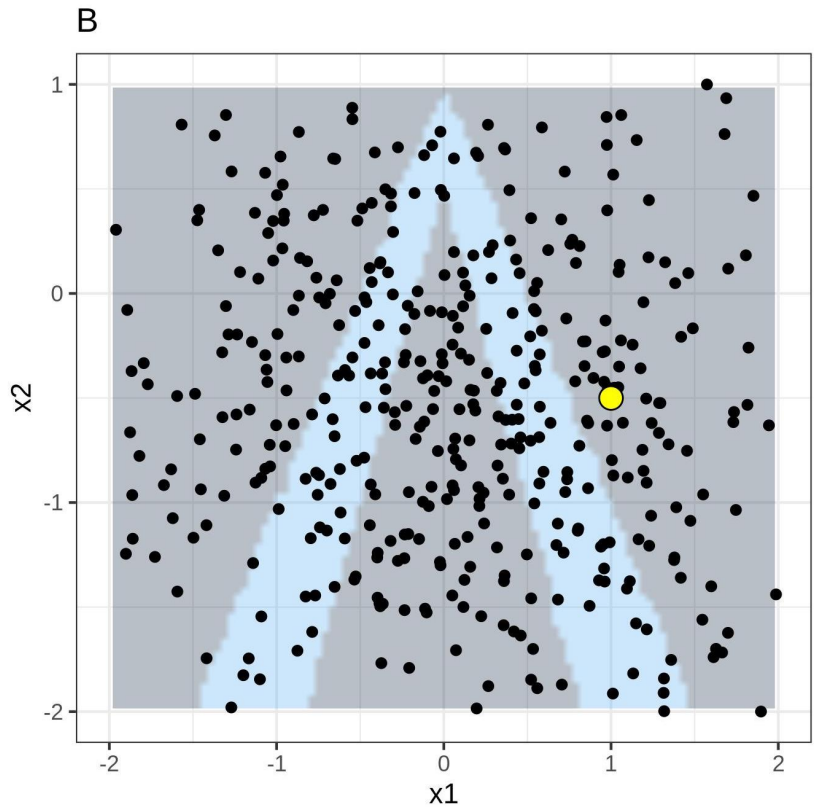
- $\mathcal{X} = \{x_1, x_2\} \subset \mathbb{R}^2$
- f: random forest
  - en foncé, classe négative
  - en clair, classe positive



Procédure d'apprentissage : illustration

(Molnar, 2020)

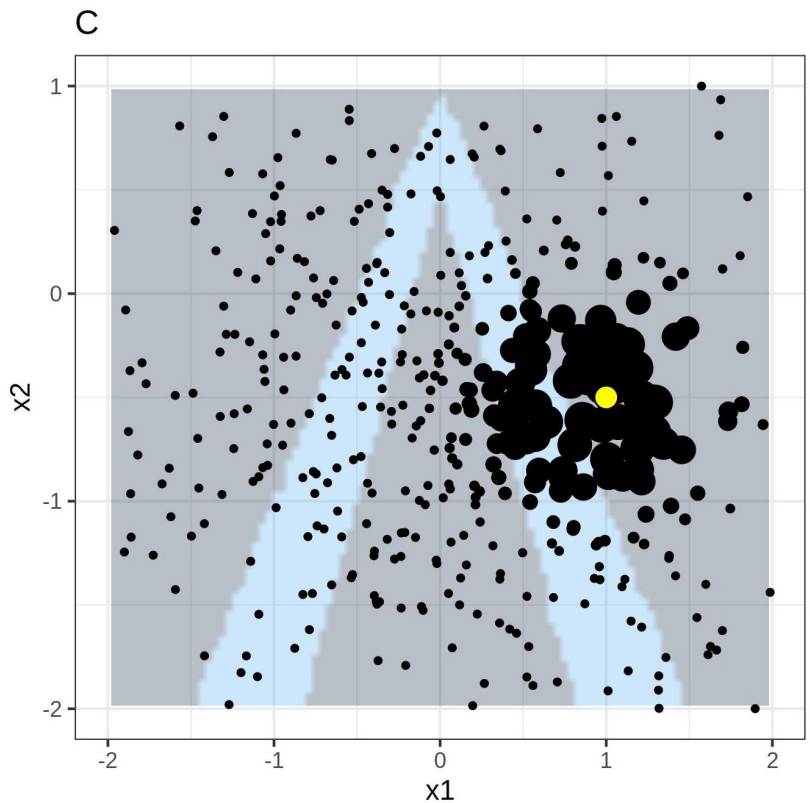
- Point jaune : instance d'intérêt x
- Construction de $\mathcal{Z} = \{(z_p, f(z_p))\}$
  - génération autour de la moyenne des données
  - selon une distribution normale



Procédure d'apprentissage : illustration

(Molnar, 2020)

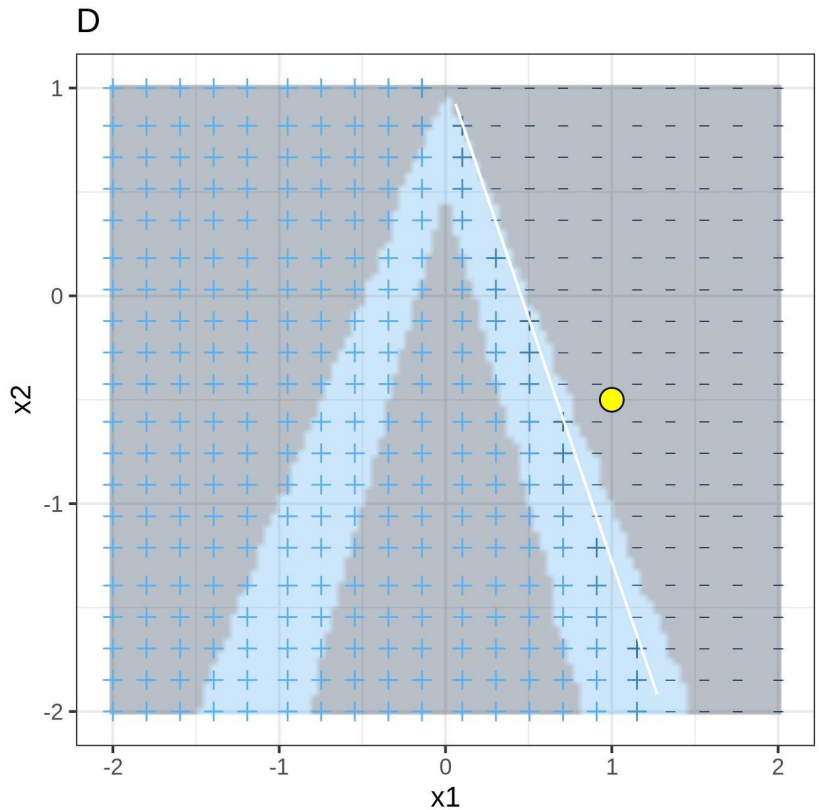
- Calcul des poids $\pi_x(z_p)$
  - selon la distance à x



Procédure d'apprentissage : illustration

(Molnar, 2020)

- Apprentissage du modèle de substitution g
  - droite en blanc : frontière de décision de g $p(class = 1) = 0.5$



Présentation des résultats

(Ribeiro et al, 2016)

Discussion sur l'échantillonnage

- Caractéristique
  - global, centré autour du centre des données
  - attribut par attribut
- Avantages
  - plus de chance de tirer des exemples de l'autre classe
- Inconvénients
  - besoin de connaissances sur les données
  - composante globale de l'explication
- Mitigé
  - réalisme des données générées ?

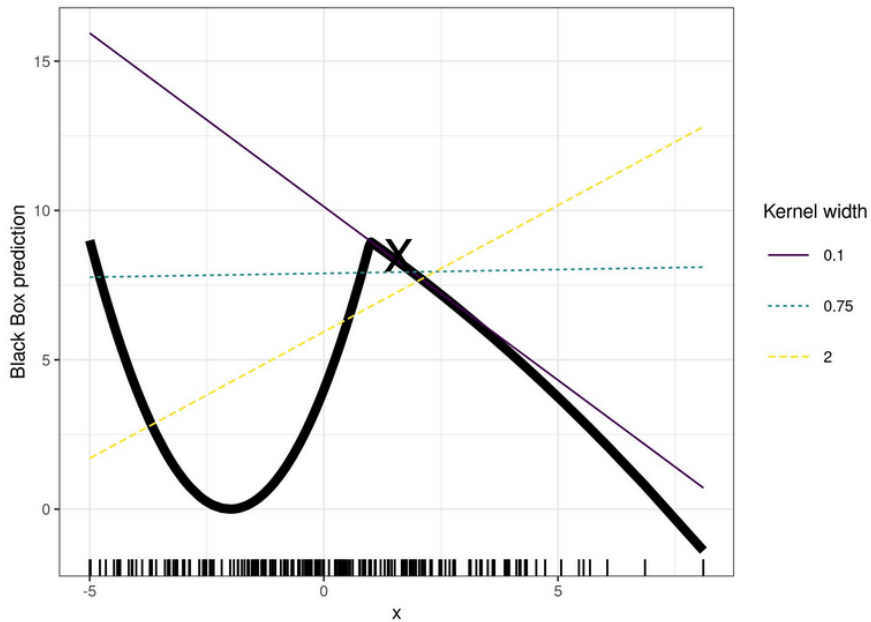
Discussion sur la définition du voisinage

- Dans l'implémentation de lime
  - après normalisation des données

$$\pi_x(z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \text{ avec } \sigma = 0.75\sqrt{d}$$

- Effet du choix de σ

(Molnar, 2021)



Cas des données textuelles

- Procédure de génération des données $\mathcal{Z}$ différente
  - suppression aléatoire de mots du texte d'intérêt
  - représentation binaire : 1 si mot conservé, 0 si supprimé

- Calcul de la prédiction : $prob = P_f(\texttt{class} = 1)$
- Calcul du poids de l'exemple généré π_z

$$weight = 1 - \frac{\#\text{ mots supprimés}}{\#\text{ mots initialement}}$$

Exemple : détection de spam

- Donnée d'intérêt x

texte	class
For Christmas Song visit my channel! ;)	1

- Variations z_p

For	Christmas	Song	visit	my	channel!	;)	P_f	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

- Apprentissage d'un modèle linéaire g pour prédire P_f

Exemple : détection de spam

- Donnée d'intérêt x

texte	class
For Christmas Song visit my channel! ;)	1

- Poids des caractéristiques

label_prob	feature	feature_weight
0.9939024	channel!	6.180747
0.9939024	For	0.000000
0.9939024	;)	0.000000

$\Longrightarrow$ Le mot "channel!" indique une forte probabilité de spam

Cas des images

- Procédure de génération des données $\mathcal{Z}$ différente
  - des perturbations au niveau des pixels seraient insuffisantes
  - de nombreux pixels contribuent à la prédiction d'une classe
- Principe
  - segmentation de l'image en régions de couleurs similaires : superpixels
  - "allumer et éteindre" les superpixels
- Explication directement sur l'image
  - en rouge : les régions qui font baisser la probabilité de la classe
  - en vert : celles qui la font augmenter

Exemple

- Modèle : Google's Inception V3 neural network
- à gauche : donnée d'intérêt
- au ilieu : classe prédite : "Bagel" (proba. 77%)
- à droite : classe prédite : "Strawberry" (proba. 4%)

Autre exemple célèbre

(Ribeiro et al, 2016)

- Reconnaissance d'images de loup vs chien husky

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

$\Longrightarrow$ détection de biais dans les données d'apprentissage

En guise de conclusion : éléments de discussion sur LIME

- Avantages
  - post-hoc : même si on change le modèle de prédiction, on peut utiliser le même modèle local interprétable pour l'explication
  - par construction les explications sont sélectives et elles peuvent être contrastives
  - LIME s'applique aux données tabulaires, aux textes et aux images
- Inconvénients
  - définition du voisinage
  - échantillonnage : effets des corrélations entre attributs ?
  - instabilité des explications