Explication de dissimilarités pour la classification et l'exploration en chémoinformatique

Sébastien Ramel[1], Simon Bernard[2], Laurent Heutte[2]

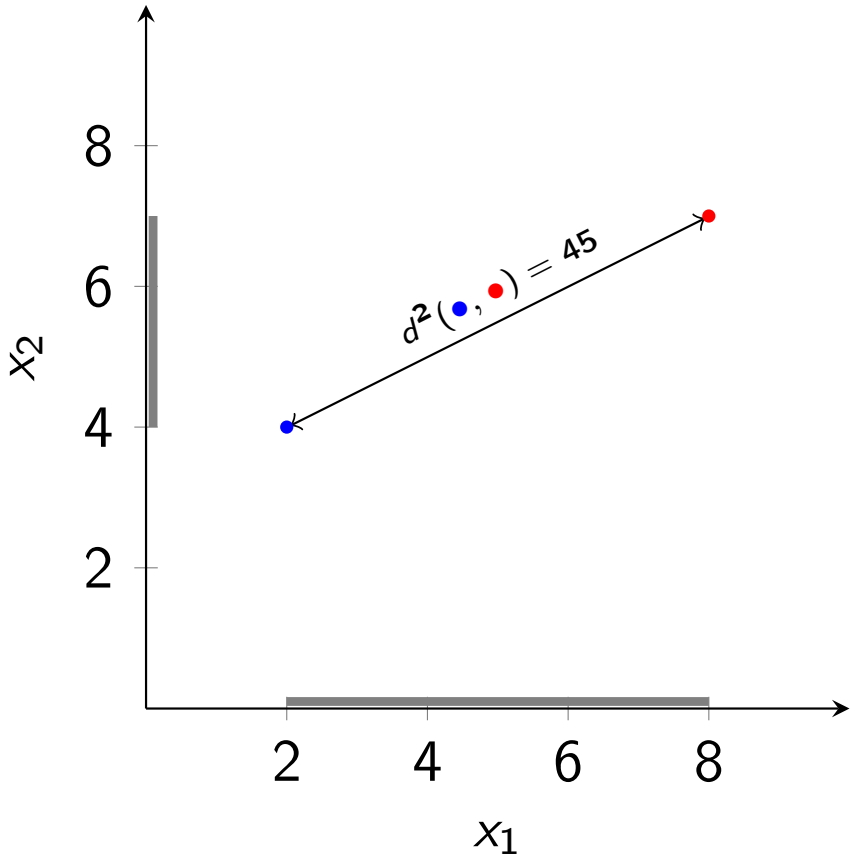
[1]Université d'Artois, LGI2A, 62400 Béthune

[2]Université de Rouen Normandie, LITIS, 76000 Rouen

3ème *réunion du projet SCHISM*
LITIS, Rouen le 16/12/2022

Explication de la distance euclidienne

attribution des caractéristiques de la distance au carré



$$
\begin{aligned}
d^2(\bullet, \bullet) &= (x_1 - x_1)^2 + (x_2 - x_2)^2 \\
&= (8 - 2)^2 + (7 - 4)^2 \\
&= 36 + 9 \\
&= 45
\end{aligned}
$$

Plan

1. Introduction
   - Contexte
   - Propositions
   - Etat de l'art
2. SHapley Additive exPlanation
   - Shapley values
   - Linear SHAP
   - Tree SHAP
3. Explication SHAP de dissimilarité
   - Contexte d'application
   - Distance de Mahalanobis
   - Dissimilarité des forêts aléatoires
4. Expériences
   - Quantitatives
   - Qualitatives
5. Conclusions

Plan

1. Introduction
   - Contexte
   - Propositions
   - Etat de l'art
2. SHapley Additive exPlanation
   - Shapley values
   - Linear SHAP
   - Tree SHAP
3. Explication SHAP de dissimilarité
   - Contexte d'application
   - Distance de Mahalanobis
   - Dissimilarité des forêts aléatoires
4. Expériences
   - Quantitatives
   - Qualitatives
5. Conclusions

Projet SCHISM [1]

Contexte et objectifs

Contexte

Analyse des relations entre la structure d'une molécule et son activité inhibitrice d'une protéine cible

Objectifs

1. Apprentissage supervisé de modèles prédictifs pour analyser la relation entre la structure moléculaire (décrites via fingerprint) et sont activité
2. Explicabilité des modèles entrainés pour permettre l'interaction avec un expert

1. Albrecht Zimmermann. *PROJET SCHISM*. 2021. url : https://schism.greyc.fr/.

Approche proposée

Décrire, prédire, expliquer la dissimilarité de paire de molécule

- Méthode proposée

  Décrire la structure topologique sous-jacente à l'activité des molécules, via des mesures de dissimilarité supervisées.

  Prédire les dissimilarités mesurées entre paires de molécule cible

  Expliquer les dissimilarités prédites selon les contributions apportées par chacune des caractéristiques.

- Applications en chémoinformatique

  Classification basée sur la dissimilarité d'une molécule test avec des représentants

  Exploration basée sur la dissimilarité d'une paire de molécule test

Dissimilarité supervisée étudiée

proximité non-linéaire des forêts aléatoires

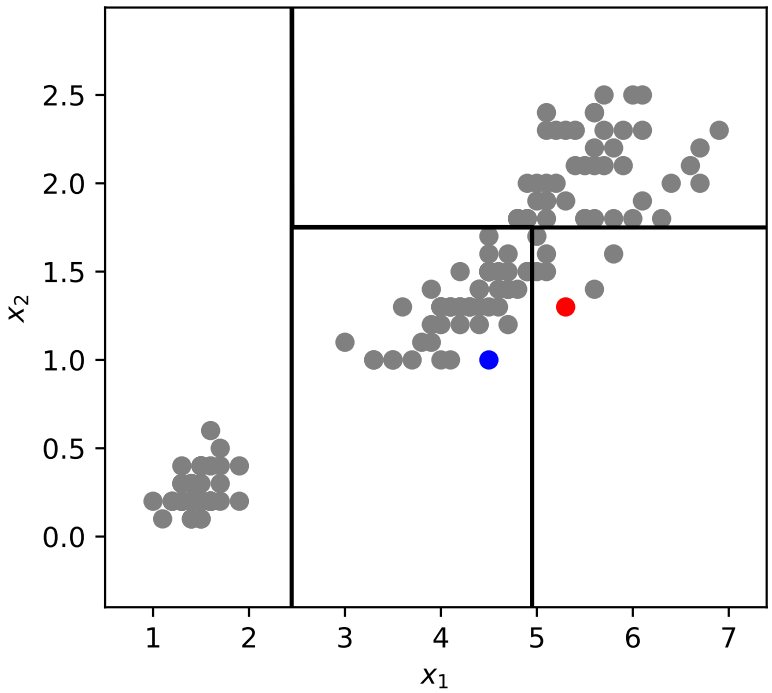
Algorithme des Forêts Aléatoires [2] (FA)

- intègre une mesure de (dis)similarité de paire d'instance, basée sur
  1. leurs positions absolues
  2. leurs appartenances aux classes
- robuste aux dimensions élevées
- fournit des outils d'analyse pour expliquer leur prédiction

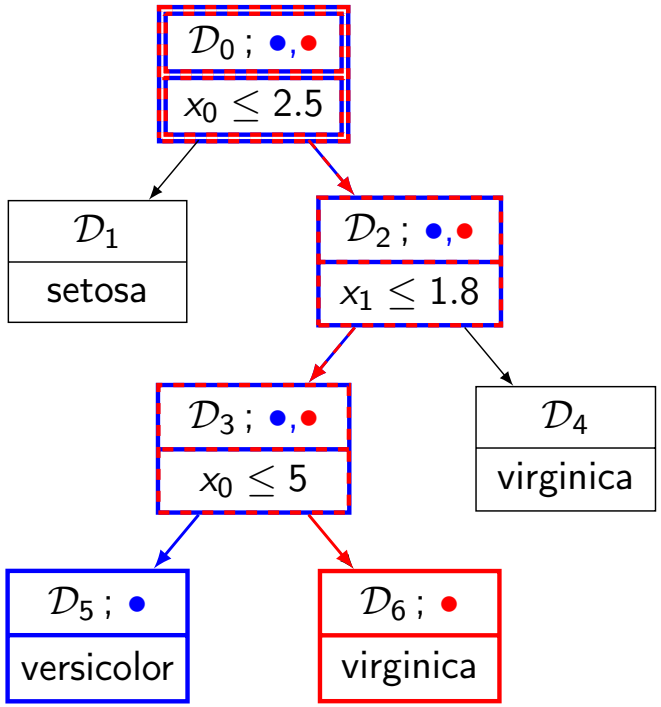
2. Leo Breiman. "Random Forests". *45* (2001), p. 5-32.

Mesure de dissimilarité des FA

Illustration sur jeu de données *iris*



(a) Feuille atteinte pour chacune des instances



(b) Chemins d'un arbre de décision, suivis par chacune des instances : ● $= (4.5, 1)$ et ● $= (5.3, 1.3)$. $d($●, ●$) = 1$.

Explicabilité des forêts aléatoires

Problèmatique et contribution

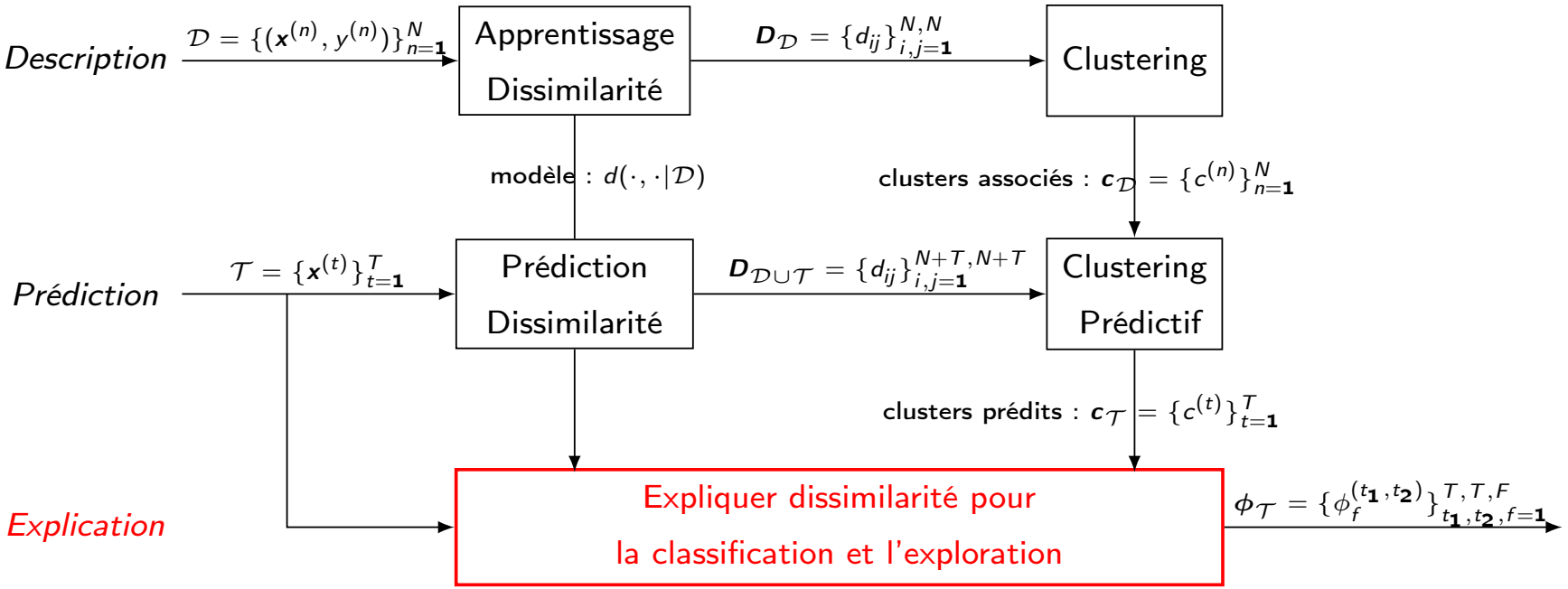
- Les outils d'analyse des prédictions des FA sont capables
  - d'analyser globalement les caractéristiques importantes en classification *e.g. Mean Decrease in Impurity, Mean Decrease in Accuracy*
  - d'expliquer localement la classification d'une instance test selon la contribution des caractéristiques *e.g. Tree SHAP (TS), Sabaas*
- Mais sont incapables d'expliquer la similarité d'une paire d'instance

$\Rightarrow$ Propositions de l'extension de la méthode SHAP à des mesures de dissimilarité et notamment à celle des FA, nommée Dissimilarity Tree SHAP (DissTS)

$\Rightarrow$ Applications à la classification basée sur des représentants et à l'exploration entre clusters

Synthèse de l'approche et de la contribution

pour la classification et l'exploration en chémoinformatique



Positionnement de la méthode proposée

vis-à-vis du clustering

Algorithme de clustering

Groupe un ensemble de points, de sorte que les points d'un groupe soient plus similaires (compacité) que les points des autres groupes (séparation)

- Consensus clutering via matrice de co-association [3]
- Clustering basé sur la mesure de proximité de forêts aléatoire non-supervisée (**urfd**) en bioinformatique [4]
- Ponderation des caractéristiques : entropy weighting k-means (ewkm) [5]

3. Dongkuan Xu et Yingjie Tian. "A comprehensive survey of clustering algorithms". *Annals of Data Science* 2.2 (2015), p. 165-193.

4. Tao Shi et Steve Horvath. "Unsupervised learning with random forest predictors". *Journal of Computational and Graphical Statistics* 15.1 (2006), p. 118-138.

5. Liping Jing, Michael K Ng et Joshua Zhexue Huang. "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data". *IEEE Transactions on knowledge and data engineering* 19.8 (2007), p. 1026-1041.

Positionnement de la méthode proposée

vis-à-vis de la classification

Classifieurs basés sur une mesure de distance

- classification par représentants : Generalized Learning Vector Quantization (glvq) [6]
- classification par k plus proche voisins (knn)
- clustering supervisé via Tree SHAP (rftskm) [7]

6. Atsushi Sato et Keiji Yamada. "Generalized learning vector quantization". *Advances in neural information processing systems* 8 (1995).

7. Scott M. Lundberg, Gabriel G. Erion et Su-In Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". (2018).

Positionnement de la méthode proposée

vis-à-vis de l'apprentissage supervisé de dissimilarité

- FA (binaire) entrainée à prédire si 2 instances (décrites via position absolue et relative) sont de même classe : (RFD-Xiong) [8]
- Distance de Mahalanobis : [Localized] Generalized Mahalanobis Learning Vector Quantization (lgmlvq, gmlvq) [9]
- Distance Euclidienne pondérée
  - [Localized] Generalized Relevance Learning Vector Quantization (lgrlv, grlvq) [10]
  - Entropy Weighting k-means (ewkm)

8. Caiming Xiong et al. "Random forests for metric learning with implicit pairwise position dependence". *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, p. 958-966.
9. Petra Schneider, Michael Biehl et Barbara Hammer. "Adaptive relevance matrices in learning vector quantization". *Neural computation* 21.12 (2009), p. 3532-3561.
10. Barbara Hammer, Frank-Michael Schleif et Thomas Villmann. "On the generalization ability of prototype-based classifiers with local relevance determination". (2005).

Positionnement de la méthode proposée

vis-à-vis de l'explication

- Méthode agnostique (au modèle)
  - Méthodes locales d'explication post-hoc
    - Local Interpretable Model-Agnostic Explanations (lime) [11]
    - SHapley Additive exPlanations (shap) [12]
  - Méthodes globales : Partial Dependence Plot (pdp)
  - Méthodes locales par selection de représentant
- Méthodes spécifiques au FA
  - globale : mean decrease in impurity (mdi), mean decrease in accuracy (mda)
  - locale : Tree SHAP (ts) [13], Sabaas

11. Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. "" Why should i trust you ?" Explaining the predictions of any classifier". *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, p. 1135-1144.
12. Scott Lundberg et Su-In Lee. "A unified approach to interpreting model predictions". ().
13. Scott Lundberg et al. "From Local Explanations to Global Understanding with

Plan

1. Introduction
   - Contexte
   - Propositions
   - Etat de l'art
2. SHapley Additive exPlanation
   - Shapley values
   - Linear SHAP
   - Tree SHAP
3. Explication SHAP de dissimilarité
   - Contexte d'application
   - Distance de Mahalanobis
   - Dissimilarité des forêts aléatoires
4. Expériences
   - Quantitatives
   - Qualitatives
5. Conclusions

Explication locale de modèle

basée sur représentation simplifiée des instances

- Explication du modèle $f(\boldsymbol{x})$ localement à une unique instance $\boldsymbol{x}$
- Les modèles d'explication utilisent souvent une représentation simplifiée (booléenne) $\boldsymbol{x}'$ telle que $\boldsymbol{x} = h_x(\boldsymbol{x}')$
- Différentes correspondances $h_{\boldsymbol{x}}$ selon l'espace des caractéristiques
  - sac de mots : convertit x'_i en nombres d'occurence x_i du mot i si $x'_i = 1$, ou à 0 si $x_i = 0$
  - images : convertit x'_i à la valeur d'orgine x_i du pixel i si $x'_i = 1$, ou à la valeur moyenne des pixels voisins si $x'_i = 0$
- Les méthodes locales garantissent l'approximation $g(\boldsymbol{z}') \approx f(h_x(\boldsymbol{z}'))$ lorsque $\boldsymbol{z}' \approx \boldsymbol{x}'$

Explication par attribution

de caractéristique additive

Attribution additive de caractéristique

Méthodes expliquant la prédiction d'un modèle complexe f via un modèle plus simple g donné par l'aggrégation linéaire de variables booléennes

$$g(\boldsymbol{z}') = \phi_0 + \sum_{i=1}^{|\mathcal{F}|} \phi_i z_i'$$

où $\boldsymbol{z}' \in \{0,1\}^{|\mathcal{F}|}$, $|\mathcal{F}|$ est le nombre de caractéristique simplifiée et $\phi_i \in \mathbb{R}$

- z_i' représente la présence ($z_i' = 1$) ou l'absence ($z_i' = 0$) de l'observation de la i-ième caractéristique,
- ϕ_i correspond à l'effet attribué à la i-ième caractéristique
- la somme des effets approxime la prédiction : $g(\boldsymbol{z}') \approx f(h_x(\boldsymbol{z}'))$.

La méthode des valeurs Shapley

appartient à l'explication par attribution additive des caractéristiques

Valeurs Shapley

Contributions (equitables) attribuées à chacun des joueurs d'une coalition $\mathcal{F}$ représentant l'effet moyen pondéré sur le gain, de l'inclusion du joueur i, dans toutes les sous-coalitions privée de i

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!\,(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} \left(f_{S \cup \{i\}}(\boldsymbol{x}_{S \cup \{i\}}) - f_S(\boldsymbol{x}_S) \right)$$

- $f_{S \cup \{i\}}(\boldsymbol{x}_{S \cup \{i\}})$ et $f_S(\boldsymbol{x}_S)$ sont resp. le gain incluant le joueur i et le gain l'excluant. Leur différence représente le gain marginal
- $|S|!$ et $(|\mathcal{F}| - |S| - 1)!$ sont resp. le nombre de combinaisons ordonnées de joueur avant l'inclusion de i et le nombre de combinaisons après son inclusion
- $|\mathcal{F}|!$ est le nombre total de combinaisons de joueurs dans la coalition

SHapley Additive exPlanation (SHAP)

Approche des valeurs Shapley pour l'IA explicable

- Joueur $i \Leftrightarrow$ Caractéristique x_i d'une instance $\boldsymbol{x}$
- Coalition $\mathcal{F} \Leftrightarrow$ vecteur de caractéristiques $\boldsymbol{x}$
- Gain marginal $v(\mathcal{F}) \Leftrightarrow f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{Z})]$, avec
  - $f(\boldsymbol{x})$: prédiction effectuée via toutes les caractéristiques de $\boldsymbol{x}$,
  - $\mathbb{E}[f(\boldsymbol{Z})]$: prédiction moyenne sur toute la base d'apprentissage.

SHapley Additive exPlanation (SHAP)

Définition de la méthode SHAP

- Valeur SHAP d'une caractéristique x_i :

$$\varphi_i(f, \boldsymbol{x}) = \sum_{\boldsymbol{z}' \subseteq \boldsymbol{x}'} \frac{|\boldsymbol{z}'|!(|\mathcal{F}|! - |\boldsymbol{z}'| - 1)!}{|\mathcal{F}|!} \left(f(h_{\boldsymbol{x}}(\boldsymbol{z}')) - f(h_{\boldsymbol{x}}(\boldsymbol{z}' \setminus i))\right)$$

  - $|\boldsymbol{z}'|$, le nombre de composantes de $\boldsymbol{z}$ non nulle
  - $\boldsymbol{z}' \subseteq \boldsymbol{x}'$, tous les vecteurs $\boldsymbol{z}'$ où les composantes non nulles sont un sous-ensemble des composantes non nulles de $\boldsymbol{x}'$
  - $f(h_{\boldsymbol{x}}(\boldsymbol{z}')) = \mathbb{E}[f(\boldsymbol{Z})|\boldsymbol{z}_S]$ et S est l'ensemble des indexes des caractéristiques non nulles dans $\boldsymbol{z}'$

- La représentation simplifiée pour SHAP est donnée par la correspondance $h_{\boldsymbol{x}}(\boldsymbol{z}') = \boldsymbol{z}_S$ (possède des valeurs manquantes pours les caractéristique abstentes de S)

Explication SHAP d'un modèle linéaire

Linear SHAP

Linear SHAP

Etant donné un modèle linéaire : $f(\boldsymbol{x}) = \lambda_0 + \sum_{i=1}^{|\mathcal{F}|} \lambda_i x_i$, supposant des caractéristiques indépendantes les valeurs SHAP sont données par

$$\phi_i(f, \boldsymbol{x}) = \begin{cases} \lambda_i(x_i - \mathbb{E}[x_i]) & \text{si } i > 0 \\ \mathbb{E}[f(\boldsymbol{X})] & \text{sinon} \end{cases}$$

Exemple 1, Linear SHAP

Instances $\boldsymbol{X}$ tirées d'une distribution uniforme $\mathcal{U}([0,1]^2)$

- Considérons $f(\boldsymbol{x}) = x_1 + x_2$
- Supposons $\boldsymbol{X} \sim \mathcal{U}([0,1]^2)$
- On a $\mathbb{E}[f(\boldsymbol{X})] = \mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 1/2 + 1/2 = 1$
- Valeurs SHAP d'une instance $\boldsymbol{x}$

$$\phi_0(f, \boldsymbol{x}^{(t)}) = \mathbb{E}[f(\boldsymbol{X})] = 1$$
$$\phi_1(f, \boldsymbol{x}^{(t)}) = x_1 - \mathbb{E}[X_1] = x_1^{(t)} - 1/2$$
$$\phi_2(f, \boldsymbol{x}^{(t)}) = x_2 - \mathbb{E}[X_2] = x_2^{(t)} - 1/2$$

Exemple 2, Linear SHAP

Paires d'instances $(\boldsymbol{X}', \boldsymbol{X}'')$ tirées d'une même gaussienne $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

- Considérons $f(\boldsymbol{z}) = z_1 + z_2$, avec $\boldsymbol{Z} = (\boldsymbol{X}' - \boldsymbol{X}'')^2$
- Supposons $\boldsymbol{X}', \boldsymbol{X}' \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ d'ou $\sqrt{\boldsymbol{Z}} \sim \mathcal{N}(0, 2\text{tr}(\Sigma))$
- On a

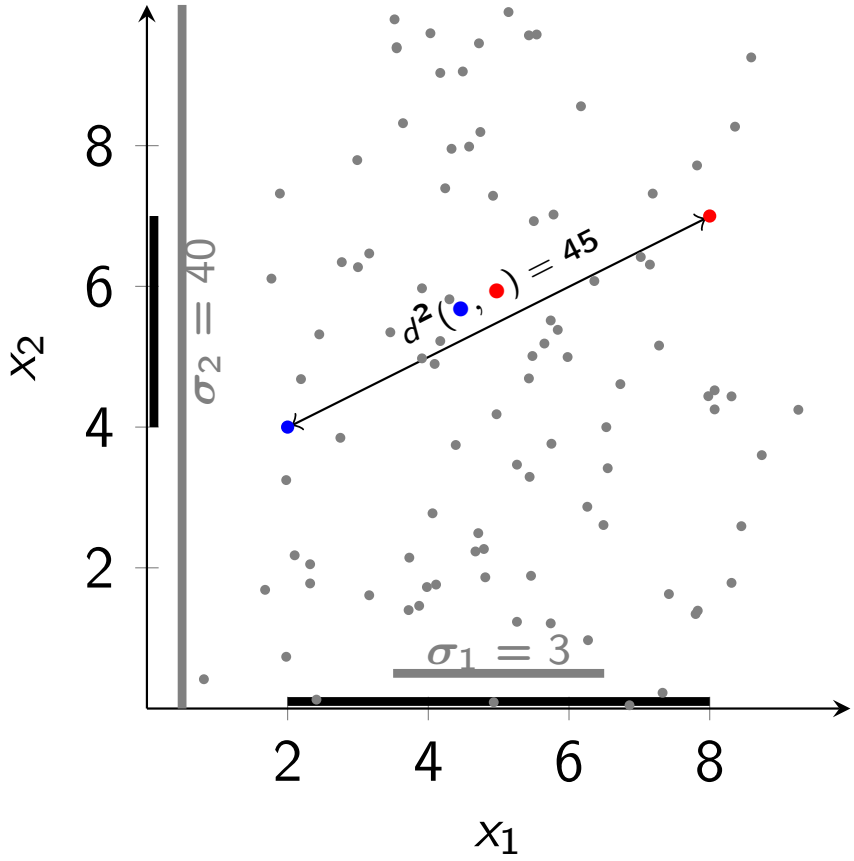
$$\begin{aligned}\mathbb{E}[f(\boldsymbol{Z})] &= \mathbb{E}[Z_1 + Z_2] \\ &= \mathbb{E}[Z_1] + \mathbb{E}[Z_2] \\ &= \mathbb{E}[\|X_1' - X_1''\|^2] + \mathbb{E}[\|X_2' - X_2''\|^2] \\ &= 2\sigma_1 + 2\sigma_2\end{aligned}$$

- Valeurs Shapley d'une paire d'instance $\boldsymbol{z}$

$$\begin{aligned}\phi_0(f, \boldsymbol{z}) &= \mathbb{E}[f(\boldsymbol{Z})] = 2(\sigma_1 + \sigma_2) \\ \phi_1(f, \boldsymbol{z}) &= z_1^{(t)} - \mathbb{E}[Z_1] = z_1 - 2\sigma_1 \\ \phi_2(f, \boldsymbol{z}) &= z_2^{(t)} - \mathbb{E}[Z_2] = z_2 - 2\sigma_2\end{aligned}$$

Exemple 3, Linear SHAP

Explication SHAP de la distance euclidienne au carré



$$
\begin{aligned}
d^2(\bullet, \bullet) &= (x_1 - x_1)^2 + (x_2 - x_2)^2 \\
&= z_1 + z_2 \\
&= 2(\sigma_1 + \sigma_2) + (z_1 - 2\sigma_1) + (z_2 - 2\sigma_2) \\
&= 2 \cdot (3 + 40) + (36 - 2 \cdot 3) + (9 - 2 \cdot 40) \\
&= 86 + 30 - 71 \\
&= 45
\end{aligned}
$$

Exemple 4, Linear SHAP

Paires d'instances $(\boldsymbol{X}', \boldsymbol{X}'')$ tirées d'un modèle de mélange gaussien GMM$(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Considérons $f(\boldsymbol{z}) = z_1 + z_2$, avec $\boldsymbol{Z} = (\boldsymbol{X}' - \boldsymbol{X}'')^2$
- Supposons $\boldsymbol{X}', \boldsymbol{X}'' \sim \text{GMM}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
- On peut montrer

$$\mathbb{E}[f(\boldsymbol{Z})] = \sum_{l,m=1}^{k,k} \pi^{(l)}\pi^{(m)} \left(\left\| \boldsymbol{\mu}^{(l)} - \boldsymbol{\mu}^{(m)} \right\|^2 + \text{tr}\left(\Sigma^{(l)} + \Sigma^{(m)} \right) \right)$$

- Valeur Shapley d'une instance

$$\phi_0(f, \boldsymbol{z}) = \mathbb{E}[f(\boldsymbol{Z})] = \sum_{l,m=1}^{k,k} \pi^{(l)}\pi^{(m)} \left(\left\| \boldsymbol{\mu}^{(l)} - \boldsymbol{\mu}^{(m)} \right\|^2 + \text{tr}\left(\Sigma^{(l)} + \Sigma^{(m)} \right) \right)$$

$$\phi_1(f, \boldsymbol{z}) = z_1 - \sum_{l,m=1}^{k,k} \pi^{(l)}\pi^{(m)} \left(\left\| \mu_1^{(l)} - \mu_1^{(m)} \right\|^2 + \sigma_1^{(l)} + \sigma_1^{(m)} \right)$$

$$\phi_2(f, \boldsymbol{z}) = z_2 - \sum_{l,m=1}^{k,k} \pi^{(l)}\pi^{(m)} \left(\left\| \mu_2^{(l)} - \mu_2^{(m)} \right\|^2 + \sigma_2^{(l)} + \sigma_2^{(m)} \right)$$

Explication SHAP de la SFA

pour la classification par représentants

ExpectedSFA($x^{(t)}$,$c^{(t)}$,S,*arbre*)
- **procedure** $G(j, w)$
  - **if** $v_j \neq$ *internal* **then**
    - **return** $w \cdot v_j$
  - **else**
    - **if** $d_j \in S$ **then**
      - **if** $x_{d_j}^{(t)} \leq t_j$ **then**
        - **return** $G(a_j, w)$
      - **else**
        - **return** $G(b_j, w)$
    - **else**
      - **return** $G(a_j, wr_{a_j}/r_j) + G(b_j, wr_{b_j}/r_j)$
- **return** $G(1, 1)$

avec

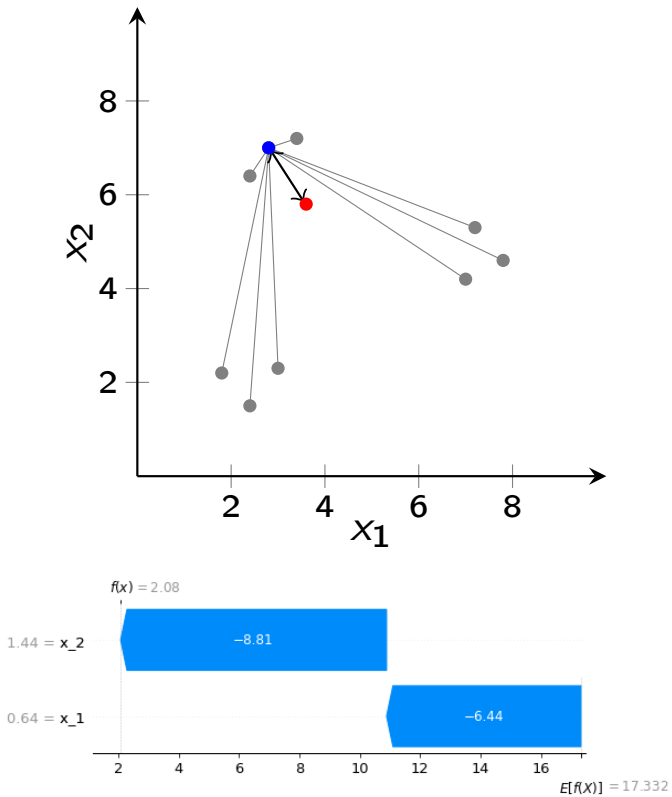
- v_j : valeurs dans feuille j. Si j nœud interne : $v_j =$ "internal"
- a_j, b_j : indexes gauche et droite du nœud interne j
- t_j : seuil du nœud interne j
- d_j : index de la caractéristique du nœud interne j
- r_j : nb. instances dans nœud j

Plan

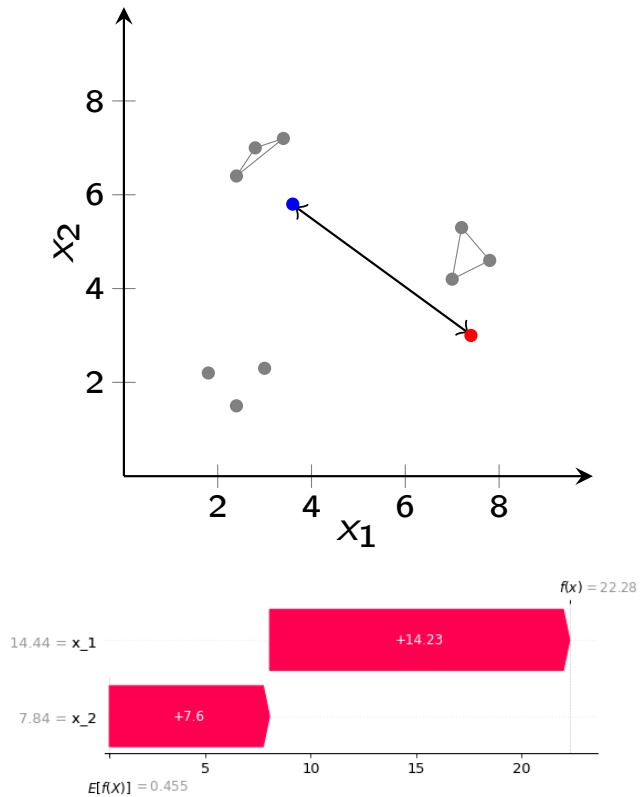
1. Introduction
   - Contexte
   - Propositions
   - Etat de l'art
2. SHapley Additive exPlanation
   - Shapley values
   - Linear SHAP
   - Tree SHAP
3. Explication SHAP de dissimilarité
   - Contexte d'application
   - Distance de Mahalanobis
   - Dissimilarité des forêts aléatoires
4. Expériences
   - Quantitatives
   - Qualitatives
5. Conclusions

Contextes d'explication SHAP de dissimilarités

Classification et exploration



(a) Classification



(b) Exploration

Distance de Mahalanobis Généralisée

Définition

Distance de Mahalanobis Généralisée (DMG)

La distance de Mahalanobis généralisée d_M entre 2 points $\boldsymbol{x}^{(l)}$ et $\boldsymbol{x}^{(m)}$, est définie par

$$d_M^2(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) = (\boldsymbol{x}^{(l)} - \boldsymbol{x}^{(m)}) M (\boldsymbol{x}^{(l)} - \boldsymbol{x}^{(m)})$$

où M est une matrice symetrique semi-définie positive arbitraire

- La matrice M est entrainée pour tenir compte des corrélations et des pondérations des caractéristiques
- Si M est diagonale de composants λ t.q. $\lambda_i \in [0, 1]$ et $\sum_i \lambda_i = 1$, la DMG est équivalente à une distance euclidienne pondérée

Distance de Mahalanobis Généralisée

exprimée selon une distance euclidienne

- M est obtenue via la décomposition spectrale $M = U\Delta U$, avec
  - U est la matrice des vecteurs propres de M
  - Δ la matrice diagonale des valeurs propres
- En posant $W = U\Delta^{1/2}$, on a

$$\begin{aligned} d_M^2(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) &= (\boldsymbol{x}^{(l)} - \boldsymbol{x}^{(m)}) WW^\intercal (\boldsymbol{x}^{(l)} - \boldsymbol{x}^{(m)}) \\ &= \left(W^\intercal (\boldsymbol{x}^{(l)} - \boldsymbol{x}^{(m)})\right)^\intercal \left(W^\intercal (\boldsymbol{x}^{(l)} - \boldsymbol{x}^{(m)})\right) \\ &= \left(\tilde{\boldsymbol{x}}^{(l)} - \tilde{\boldsymbol{x}}^{(m)}\right)^\intercal \left(\tilde{\boldsymbol{x}}^{(l)} - \tilde{\boldsymbol{x}}^{(m)}\right) \\ &= \sum_{j=1}^{|\mathcal{F}|} (\tilde{x}_i^{(l)} - \tilde{x}_i^{(m)})^2 \end{aligned}$$

où $\tilde{\boldsymbol{x}} = W^\intercal \boldsymbol{x}$

Explication SHAP de la DMG

pour la classification par représentants

- Soient $\boldsymbol{x}^{(t)}$ une instance de test et $\boldsymbol{p}^{(t)}$ son prototype le plus proche
- Les valeurs SHAP de la DMG en classification sont données par

$$\phi_0(d_M, \boldsymbol{x}^{(t)}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x}^{(i)} \in \mathcal{D}} \left\| \tilde{\boldsymbol{p}}^{(t)} - \tilde{\boldsymbol{x}}^{(i)} \right\|^2$$

$$= \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x}^{(i)} \in \mathcal{D}} \sum_{j=1}^{|\mathcal{F}|} (\tilde{p}_j^{(t)} - \tilde{x}_j^{(i)})^2$$

$$\phi_j(d_M, \boldsymbol{x}^{(t)}) = (\tilde{p}_j^{(t)} - \tilde{x}_j^{(t)})^2 - \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (\tilde{p}_j^{(t)} - \tilde{x}_j^{(i)})^2, \quad j = 1, \ldots, |\mathcal{F}|$$

Explication SHAP de la DMG

pour l'exploration entres clusters

- Soient $\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}$, des instances de test appartenant resp. aux clusters $\mathcal{C}^{(l)}, \mathcal{C}^{(m)}$
- Les valeurs SHAP de la DMG en exploration sont données par

$$\phi_0(d_M, \boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) = \kappa^{(l,m)} \left(\sum_{(\tilde{\boldsymbol{x}}', \tilde{\boldsymbol{x}}'') \in (\mathcal{C}^{(l)})^2} \left\| \tilde{\boldsymbol{x}}' - \tilde{\boldsymbol{x}}'' \right\|^2 + \sum_{(\tilde{\boldsymbol{x}}', \tilde{\boldsymbol{x}}'') \in (\mathcal{C}^{(m)})^2} \left\| \tilde{\boldsymbol{x}}' - \tilde{\boldsymbol{x}}'' \right\|^2 \right)$$

$$\phi_j(d_M, \boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) = \kappa^{(l,m)} \left(\sum_{(\tilde{\boldsymbol{x}}', \tilde{\boldsymbol{x}}'') \in (\mathcal{C}^{(l)})^2} (\tilde{x}_j' - \tilde{x}_j'')^2 + \sum_{(\tilde{\boldsymbol{x}}', \tilde{\boldsymbol{x}}'') \in (\mathcal{C}^{(m)})^2} (\tilde{x}_j' - \tilde{x}_j'')^2 \right),$$

$$j = 1, \ldots, |\mathcal{F}|$$

avec $\kappa^{(l,m)} = \left(\binom{|\mathcal{C}^{(l)}|}{2} + \binom{|\mathcal{C}^{(m)}|}{2} \right)^{-1}$, l'inverse du nombre de paires d'instance formées dans $\mathcal{C}^{(l)}$ et dans $\mathcal{C}^{(m)}$

Proximité des forêts aléatoires

Définition

Similarité d'un arbre de décision

$$s_{arbre}(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) = \begin{cases} 1 & \text{si } (\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) \text{ atteint la même feuille,} \\ 0 & \text{sinon.} \end{cases}$$

Proximité d'une FA (DFA)

$$s_{FA}(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) = \frac{1}{|FA|} \sum_{arbre=1}^{|F|} s_{arbre}(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)})$$

- Mesure de Dissimilarité d_{FA}, donnée par

$$d_{FA}(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) = 1 - s_{FA}(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)})$$

Explication SHAP de la SFA

pour la classification par représentants

ExpectedSFA($x^{(t)}$,$c^{(t)}$,S,*arbre*)

```
procedure G(j, w)
    if v_j ≠ internal then
        return w · v_j[c^(t)]
    else
        if d_j ∈ S then
            if x^(t)_{d_j} ≤ t_j then
                return G(a_j, w)
            else
                return G(b_j, w)
        else
            return
              G(a_j, wr_{a_j}/r_j) + G(b_j, wr_{b_j}/r_j)
return G(1, 1)
```

avec

- v_j : indicateurs prototypes dans feuille j. Si j nœud interne : $\boldsymbol{v}_j =$ "internal"
- a_j, b_j : indexes gauche et droite du nœud interne j
- t_j : seuil du nœud interne j
- d_j : index de la caractéristique du nœud interne j
- r_j : nb. instances dans nœud j

Explication SHAP de la DFA

pour l'exploration entre clusters

ExpectedSFA($x^{(l)}$,$x^{(m)}$,$c^{(l)}$,$c^{(m)}$,S,*arbre*)

procedure $G(j, w, c)$
- **if** $v_j \neq$ *internal* **then**
  - **return** $w \cdot 1$; /* the pair reaches a leaf, they are similar */
- **else**
  - **if** $d_j \in S$ **then**
    - **if** $(x_{d_j}^{(l)} \leq t_j)$ **and** $(x_{d_j}^{(m)} \leq t_j)$ **then**
      - **return** $P(a_j, w)$
    - **else if** $(x_{d_j}^{(l)} > t_j)$ **and** $(x_{d_j}^{(m)} > t_j)$ **then**
      - **return** $P(b_j, w)$
    - **else**
      - **return** $w \cdot 0$; /* the node splits the pair, they are dissimilar */
  - **else**
    - **return** $G(a_j, w\binom{r_{a_j}^c+1}{2}/\binom{r_j^c+1}{2}) + G(b_j, w\binom{r_{b_j}^c+1}{2}/\binom{r_j^c+1}{2})$

return $\frac{1}{2}G(1, 1, c^{(l)}) + \frac{1}{2}G(1, 1, c^{(m)})$

Explication SHAP de la DFA

Quelques précisions

- r_j^c est le nombre d'instance du cluster $\mathcal{C}^c$ dans le noeud j
- Procédures **ExpectedSFA** permettent de calculer $\mathbb{E}[d_{FA}|\boldsymbol{z}_S]$ dans la méthode SHAP
- Valeurs SHAP expliquant la similarité d'un arbre

$$\left(\phi_0(s_{FA}), \phi_1(s_{FA}), \ldots, \phi_{|\mathcal{F}|}(s_{FA})\right)$$

- Valeurs SHAP expliquant la dissimilarité

$$\left(1 - \phi_0(s_{FA}), -\phi_1(s_{FA}), \ldots, -\phi_{|\mathcal{F}|}(s_{FA})\right)$$

Plan

1 Introduction
- Contexte
- Propositions
- Etat de l'art

2 SHapley Additive exPlanation
- Shapley values
- Linear SHAP
- Tree SHAP

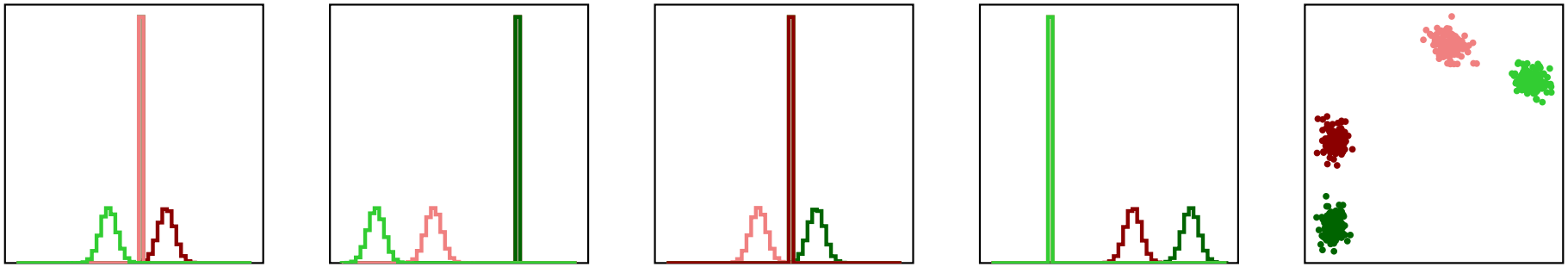
3 Explication SHAP de dissimilarité
- Contexte d'application
- Distance de Mahalanobis
- Dissimilarité des forêts aléatoires

4 Expériences
- Quantitatives
- Qualitatives

5 Conclusions

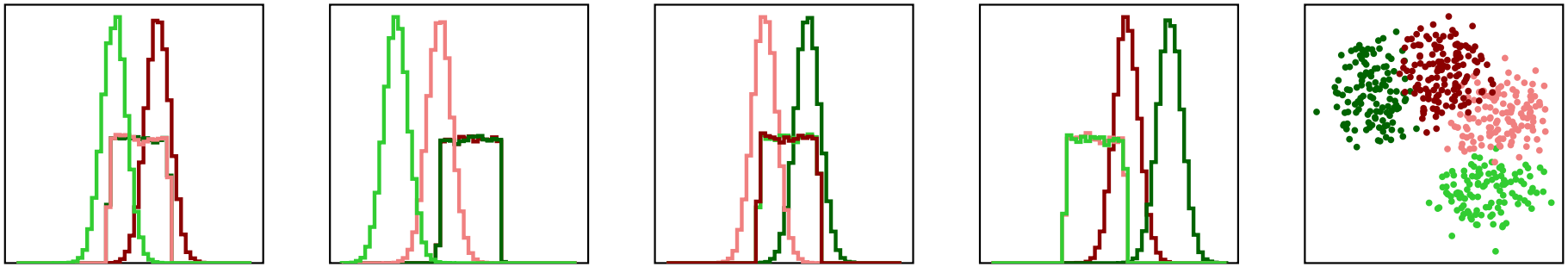
Données simulées (Preuve de concept)

4 dimensions, 2 classes, 4 clusters



(a) x_1 (b) x_2 (c) x_3 (d) x_4 (e) mds

Figure – Données précises



(a) x_1 (b) x_2 (c) x_3 (d) x_4 (e) mds

Figure – Données bruitées

Performance des explications SHAP

de dissimilarités pour la classification

- Soit $\mathcal{F}^{(i)} = \{f_1^{(i)}, f_2^{(i)}\}$, la paire de caractéristiques les plus importantes (attributions SHAP minimum) identifiées pour la prédiction d'une instance $\boldsymbol{x}^{(i)}$
- Soit $\mathcal{E}^{(i)} = \{e_1^{(i)}, e_2^{(i)}\}$, la vérité relative à la prédiction de $\boldsymbol{x}^{(i)}$
- Indicateurs de performance des explications mesurés sur $\mathcal{T}$

$$Acc_{exact} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \mathbb{1}\{\mathcal{F}^{(i)} = \mathcal{E}^{(i)}\}$$

$$Acc_{moy} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{|\mathcal{F}^{(i)} \cap \mathcal{E}^{(i)}|}{|\mathcal{E}^{(i)}|}$$

Performance des explications SHAP

de dissimilarités pour l'exploration entre clusters

- Soit $\mathcal{F}^{(v,w)} = \{f_1^{(v,w)}, f_2^{(v,w)}, f_3^{(v,w)}, f_4^{(v,w)}\}$, les caractéristiques ordonées par niveau d'importance (attributions SHAP maximum) décroissant pour $d(\boldsymbol{x}^{(v)}, \boldsymbol{x}^{(w)})$
- Soient $\mathcal{E}^{(v)} = \{e_1^{(v)}, e_2^{(v)}\}$ et $\mathcal{E}^{(w)} = \{e_1^{(w)}, e_2^{(w)}\}$, les vérités relatives aux predictions de $\boldsymbol{x}^{(v)}$ et $\boldsymbol{x}^{(w)}$
- Notons $\mathcal{E}^{(v\cap w)} = \mathcal{E}^{(v)} \cap \mathcal{E}^{(w)}$ et $\mathcal{E}^{v\cup w} = \mathcal{E}^{(v)} \cup \mathcal{E}^{(w)}$
- Indicateurs de performance des explications mesurés sur $\mathcal{T}$

$$Acc_{exact} = \binom{k}{2}^{-1} \sum_{l,m=1}^{k,k} \left(\frac{1}{|\mathcal{C}^{(l)}| \cdot |\mathcal{C}^{(m)}|} \sum_{v,w=1}^{|\mathcal{C}^{(l)}|,|\mathcal{C}^{(m)}|} \left(\frac{2}{3} I_1 + \frac{1}{3} I_2 \right) \right)$$

$$I_1 = \mathbb{1}\left\{ \mathcal{F}^{(v,w)}_{1,|\mathcal{E}^{(v\cap w)}|} = \mathcal{E}^{(v\cap w)} \right\}$$

$$I_2 = \mathbb{1}\left\{ \mathcal{F}^{(v,w)}_{|\mathcal{E}^{(v\cap w)}|,|\mathcal{E}^{(v\cap w)}|+|\mathcal{E}^{(v\cup w)}\setminus\mathcal{E}^{(v\cap w)}|} = \{\mathcal{E}^{(v\cup w)}\setminus\mathcal{E}^{(v\cap w)}\} \right\}$$

Performances prédictives

Taux de reconnaissance de classes

Clusters	Précis		Bruités	
#Classes / Modèles	2	4	2	4
lvq	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.02	0.95 ± 0.01
gmlvq	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.02	0.96 ± 0.02
grlvq	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.02	0.95 ± 0.02
lgrlvq	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.04	0.83 ± 0.09
rf	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.02	0.95 ± 0.02
rftskm	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.02	0.69 ± 0.03
rfdkmed	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.01	0.94 ± 0.02

Table – Taux de reconnaissance des classes

Performances prédictives

Taux de reconnaissance de clusters

Clusters	Précis		Bruités	
#Classes / Modèles	2	4	2	4
glvq	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.03	0.95 ± 0.01
gmlvq	0.84 ± 0.20	1.00 ± 0.00	0.84 ± 0.05	0.96 ± 0.02
grlvq	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.03	0.95 ± 0.02
lgrlvq	1.00 ± 0.00	1.00 ± 0.00	0.71 ± 0.13	0.83 ± 0.09
rf	–	–	–	–
rftskm	1.00 ± 0.00	1.00 ± 0.00	0.76 ± 0.08	0.69 ± 0.03
rfdkmed	1.00 ± 0.00	1.00 ± 0.00	0.82 ± 0.04	0.94 ± 0.02

Table – Taux de reconnaissance des clusters

Performances explicatives en prédiction

Taux de caractéristiques importantes Acc_{exact} selectionnées via SHAP

Clusters	Précis		Bruités	
#Classes / Modèles	2	4	2	4
glvq	0.00 ± 0.00	0.00 ± 0.00	0.16 ± 0.12	0.07 ± 0.06
gmlvq	0.25 ± 0.43	0.00 ± 0.00	0.19 ± 0.30	0.01 ± 0.03
grlvq	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.03	0.00 ± 0.00
lgrlvq	0.05 ± 0.22	0.00 ± 0.00	0.21 ± 0.30	0.38 ± 0.41
rf	0.95 ± 0.22	1.00 ± 0.01	0.44 ± 0.33	0.29 ± 0.21
rftskm	0.00 ± 0.00	0.05 ± 0.22	0.00 ± 0.01	0.01 ± 0.02
rfdkmed	0.90 ± 0.30	0.99 ± 0.01	0.54 ± 0.22	0.39 ± 0.23

Table – Taux de caractéristiques importantes correctement selectionnées Acc_{exact} selon la décomposition SHAP

Performances explicatives en prédiction

Taux de caractéristiques importantes Acc_{exact} via distance euclidienne au carré

Clusters	Précis		Bruités	
#Classes / Modèles	2	4	2	4
glvq	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.07	0.32 ± 0.09
gmlvq	0.18 ± 0.30	0.08 ± 0.09	0.20 ± 0.20	0.22 ± 0.22
grlvq	0.00 ± 0.00	0.00 ± 0.00	0.25 ± 0.12	0.21 ± 0.15
lgrlvq	0.08 ± 0.18	0.19 ± 0.37	0.04 ± 0.07	0.07 ± 0.17
rf	–	–	–	
rftskm	0.00 ± 0.00	0.05 ± 0.22	0.04 ± 0.07	0.09 ± 0.09
rfdkmed	–	–	–	–

Table – Taux de caractéristiques importantes correctement selectionnées Acc_{exact} en classification, selon la décomposition de la distance euclidienne au carré

Performances explicatives en exploration

Taux de caractéristiques importantes Acc_{exact} via SHAP

Clusters	Précis		Bruités	
Modèles \ #Classes	2	4	2	4
glvq	0.65 ± 0.46	0.64 ± 0.46	0.50 ± 0.40	0.55 ± 0.35
gmlvq	0.41 ± 0.46	0.66 ± 0.47	0.58 ± 0.44	0.57 ± 0.43
grlvq	0.66 ± 0.47	0.66 ± 0.47	0.54 ± 0.41	0.55 ± 0.39
rftskm	0.44 ± 0.48	0.33 ± 0.47	0.45 ± 0.42	0.54 ± 0.40
rfdkmed	0.89 ± 0.31	0.68 ± 0.46	0.61 ± 0.42	0.61 ± 0.44

Table – Taux de caractéristiques importantes correctement selectionnées Acc_{exact} en exploration, selon la décomposition SHAP

Performances explicatives en exploration

Taux de caractéristiques importantes Acc_{exact} via distance euclidienne au carré

Clusters	Précis		Bruités	
#Classes / Modèles	2	4	2	4
glvq	0.66 ± 0.47	0.65 ± 0.46	0.48 ± 0.40	0.51 ± 0.36
gmlvq	0.41 ± 0.46	0.66 ± 0.47	0.57 ± 0.44	0.58 ± 0.43
grlvq	0.66 ± 0.47	0.67 ± 0.47	0.51 ± 0.41	0.52 ± 0.39
rftskm	0.44 ± 0.48	0.33 ± 0.47	0.46 ± 0.41	0.54 ± 0.40
rfdkmed	–	–	–	–

Table – Taux de caractéristiques importantes correctement selectionnées Acc_{exact} en exploration, selon la décomposition de la distance euclidienne au carré

Données simulées (Jing) [14]

Plusieurs dimensions, classes et clusters

- Données synthétique dont la structure des clusters et la parcimonie des données sont contrôlées
  - Données d'un cluster concentrées dans un sous-ensemble de dimensions pertinentes sélectionnées aléatoirement
    (taille contrôlée par *subspace ratio* s)
  - Autres dimensions non pertinentes contiennent des valeurs nulles et quelques valeurs positives générées aléatoirement
    (taux piloté contrôlé *sparsity* ϵ),
  - Les dimensions pertinentes des clusters peuvent se chevaucher
    (taille contrôlée par *overlap ratio* ρ).

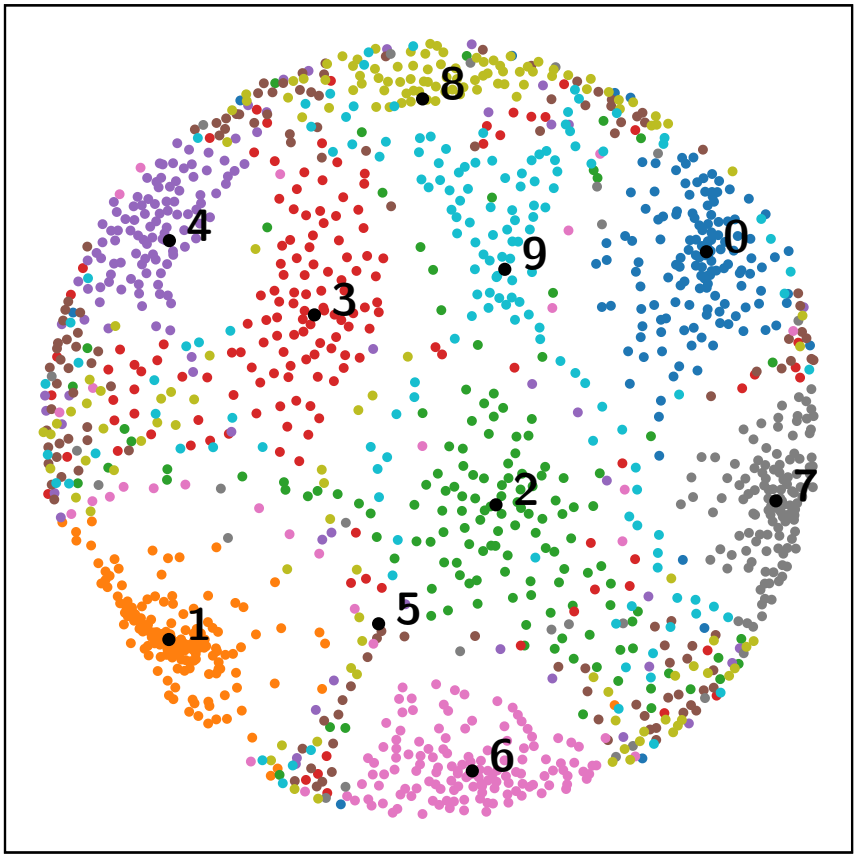
14. Jing, Ng et Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data".

Données simulées (Jing)

- Plusieurs types de clusters gaussiens isotropiques sont proposés
  - Type I Toutes les dimensions pertinentes d'un cluster ont la même importance et donc une même variance (=0.9)
  - Type II Variances des clusters (dans dimensions pertinentes) sélectionnées aléatoirement parmi 3 intervalles
    1. [0.1, 0.2]
    2. [1, 2]
    3. [5, 6]

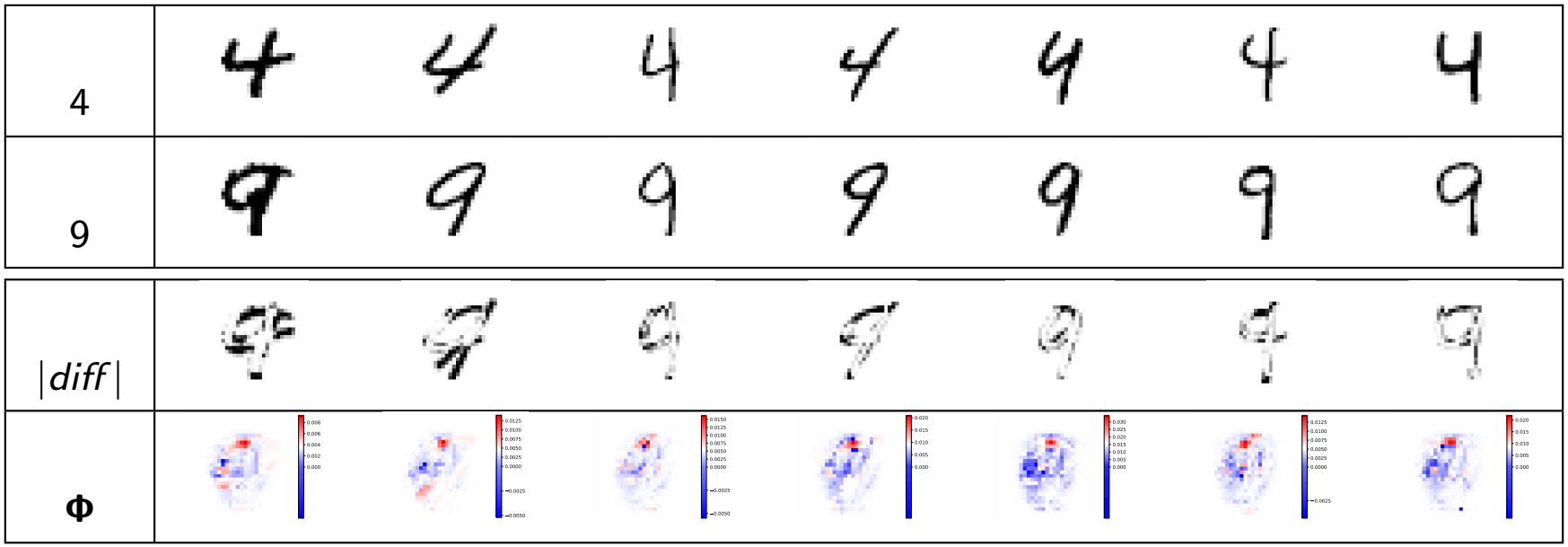
Donnée MNIST

Positionnement 2D via RFD



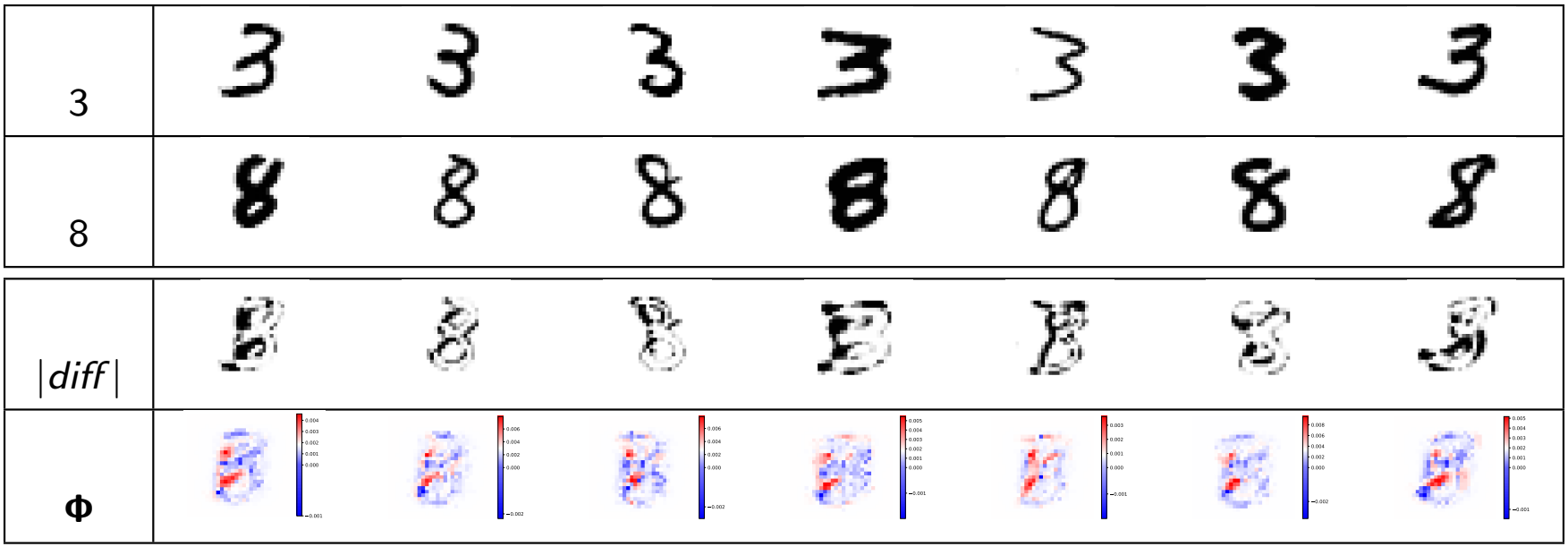
Similarity tree SHAP en exploration

Dissimilarité entre les instances des clusters 4 et 9



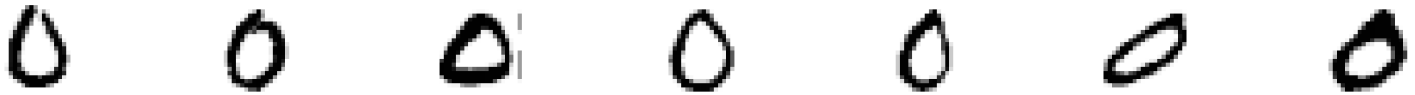
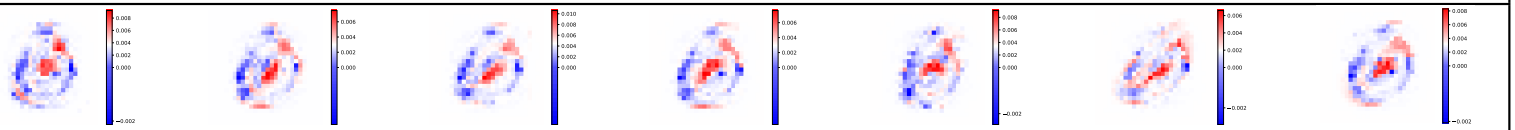
Similarity tree SHAP en exploration

Dissimilarité entre les instances des clusters 3 et 8



Similarity tree SHAP en exploration

Dissimilarité entre les instances des clusters 0 et 6

0	
6	
$\|diff\|$	
Φ	

Similarity tree SHAP en classification

Dissimilarité entre les instances du cluster 4 et son prototype

coming soon

Jeu de données *fingerprint*

accessibles

- Quelques jeux de données de molécules décrites selon la représentation *fingerpint*
  1. RSCTC 2010 Discovery Challenge (OpenML)
     "*Sélection de caractéristiques pertinentes pour l'analyse des données de puces à ADN (DNA-microarray) dans le diagnostic médical*"
  2. QSAR androgen receptor Data Set (UCI)
     "*Discrimination des molécules liantes/positives (199) et non liantes/négatives (1488) concernant la liaison au récepteur des androgènes*"
  3. QSAR oral toxicity Data Set (UCI)
     " *discrimination des molécules très toxiques/positives (741) et peu toxiques/négatives (8251) concernant la toxicité systémique orale aiguë*"

Plan

1. Introduction
   - Contexte
   - Propositions
   - Etat de l'art
2. SHapley Additive exPlanation
   - Shapley values
   - Linear SHAP
   - Tree SHAP
3. Explication SHAP de dissimilarité
   - Contexte d'application
   - Distance de Mahalanobis
   - Dissimilarité des forêts aléatoires
4. Expériences
   - Quantitatives
   - Qualitatives
5. Conclusions

Conclusions

Synthèse et perspectives

- Synthèse
  - Décrire la structure de l'activité de molécules en utilisant la mesure (supervisée) de (dis)similarité d'une FA
  - Prédire le cluster de molécules test étant donné la description inférée à l'apprentissage.
  - Proposition d'une variante de SHAP pour expliquer les dissimilarités en classification et en exploration
- Perspectives
  1. Déduire Dissimilarity Tree SHAP pour des FA non supervisée en clustering
  2. Etendre les outils d'analyse pour le clustering hiérachique

Merci de votre attention