Réseaux bayésiens et valeurs de Shapley

Mahdi Hadj Ali, Yann Le Biannic, Pierre-Henri Wuillemin

Réseaux bayésiens et valeurs de Shapley

Mahdi Hadj Ali[1,2], **Yann Le Biannic**[2], **Pierre-Henri Wuillemin**[1]
[1] LIP6 (UMR 7606), Sorbonne Université, 4 place Jussieu, 75005 Paris, France
[2] SAP, 35 rue rue d'Alsace 92300 Levallois-Perret,France
mahdi.hadj.ali@sap.com, yann.le.biannic@sap.com ,pierre-henri.wuillemin@lip6.fr

Abstract

L'interprétabilité de modèles d'apprentissage automatique est un sujet de plus en plus sensible. Des travaux récents proposent de quantifier les contributions des variables d'un modèle prédictif en s'appuyant sur les valeurs de Shapley. Dans cet article, nous recensons différentes fonctions caractéristiques utilisées dans le calcul des valeurs de Shapley, pour quantifier le pouvoir prédictif direct ou indirect des variables, ou encore leur influence causale. Nous présentons ensuite des techniques ce calcul pour évaluer et appliquer les valeurs de Shapley dans le domaine des réseaux bayésiens. Enfin, l'article propose de promouvoir les valeurs de Shapley comme une articulation entre ces deux facettes de l'apprentissage statistique que forment d'un côté, les modèles prédictifs et de l'autre, les modèles graphiques.

Introduction

Les méthodes récentes de *Machine Learning* telles que les *Random Forests* et les *Boosting Machines*, de plus en plus sophistiquées, améliorent généralement la précision des modèles construits, mais au détriment d'une difficulté d'interprétation plus grande que dans les approches plus simples tels que les régressions linéaires et les arbres de décision. Par ailleurs, l'interprétabilité de ces modèles est un sujet de plus en plus sensible dans de nombreux domaines (Burkart and Huber 2021). En effet, l'utilisation de ces modèles dans le cadre de la prise de décision automatique demande la connaissance fine des comportements afin de pouvoir justifier la décision (par exemple, dans le domaine médical de la prescription automatique, dans le domaine juridique ou dans un contexte légal) (Rieg et al. 2020).

Ainsi l'interprétabilité des modèles est aujourd'hui un domaine important de la recherche en intelligence artificielle. "L'interprétabilité est le degré auquel un humain peut comprendre ce qui cause d'une décision" (Miller 2018). Plus le modèle est simple à interpréter, plus il sera facile de comprendre pourquoi et comment certaines décisions ou prédictions ont été faites (Molnar 2019).

Les principaux points de notre article sont une présentation du framework que forment les valeurs de Shapley, les différentes valeurs de Shapley et leurs méthodes de calculs puis leurs implémentations dans les réseaux bayésiens à travers le framework *PyAgrum* (Ducamp, Gonzales, and Wuillemin 2020) et enfin comment les valeurs de Shapley peuvent aider à la construction d'un graphe à travers un exemple illustré.

Explicabilité, modèles additifs, valeurs de Shapley

Soit un problème de prédiction de la classe C, appris à partir d'une base de données composée de N variables (*features*) $\mathbf{X} = \{X_1, X_2, \cdots, X_j, \cdots, X_N\}$ et de D lignes. On notera $f(X_1, \cdots, X_n)$ la fonction calculant la valeur prédite en fonction de ces variables. Dans ce cadre, l'interprétabilité du modèle f se comprend comme une analyse de l'importance de chaque X_j dans la construction de la valeur de f. Plusieurs outils ont été développés pour répondre à ce besoin.

Par exemple, les *Partial Dependence Plots* (PDP, Friedman 2001) proposent d'examiner l'effet de la j-ème variable en étudiant la prédiction moyenne lorsque cette j-ème variable est perturbée (par exemple en mélangeant les valeurs de la colonne j dans la base).

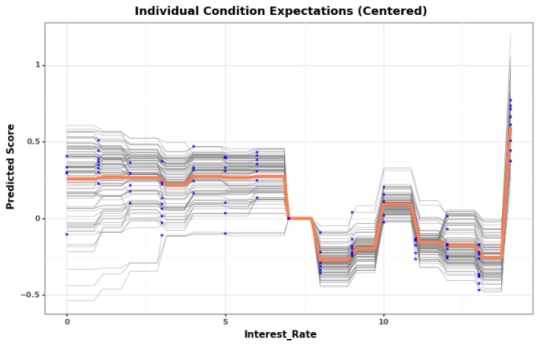


Figure 1: Diagramme ICE avec PDP en orange

Les *Individual Conditional Expectation Plots* (ICE) se basent sur la même idée que les PDP; mais correspondent à l'étude de la prédiction donnée par f à un exemple donné lorsque la j-ème variable est modifiée. Ainsi la moyenne de toutes les ICE correspond au PDP.

Une autre idée pour l'explicabilité est de fournir un score d'importance de chaque variable ; par exemple, Breiman

(2001) propose d'échanger une variable contre du bruit puis de regarder comment la prédiction f en est impactée. D'autres méthodes existent encore, mais chacune apporte un certain nombre de contraintes qui doivent être respectées afin de s'assurer de la pertinence de l'interprétation faite (Hooker and Mentch 2019).

L'outil d'explicabilité étudié dans cet article est les valeurs de Shapley (Lundberg and Lee 2017). Cette méthode s'inscrit dans le cadre d'un modèle d'explication additif que nous détaillons ici.

Décomposition additive et valeurs de Shapley

Soit f la prédiction du modèle que l'on souhaite expliquer. Les méthodes de décomposition additive proposent un modèle d'explication g qui est une fonction linéaire de variables binaires approchant au mieux f (Ribeiro, Singh, and Guestrin 2016) :

Definition 1 ***décomposition additive*** *Soit d une ligne de la base et $\mathbf{x_d} = (x_1, \cdots, x_n)$ le vecteur des valeurs pour $X_1, \cdots, X_n$ dans la ligne d[1] (Lundberg and Lee 2017).*

La décomposition additive g est une fonction approchant f qui vérifie :

$$g(\mathbf{x}) = \phi_0 + \sum_{i=1}^{n} \phi_i(\mathbf{x})$$

avec $\phi_0 = \mathbb{E}[f(\mathbf{X})]$

L'intérêt d'une telle décomposition est la facilité d'utiliser les $\phi_i(x)$ comme des scores d'importance pour chaque variable dans la ligne d comme proposé dans LIME (Ribeiro, Singh, and Guestrin 2016). Dans ce cadre, les valeurs de Shapley proposent une méthode de calcul de ces ϕ_i.

L'utilisation des valeurs de Shapley est rapidement devenue populaire dans le domaine de l'interprétabilité. Cette approche se base sur un outil issu de la théorie des jeux coopératifs. Dans un jeu coopératif, les joueurs collaborent pour obtenir un gain. Le problème est alors de répartir le gain entre les différents acteurs (sous la forme d'une allocation). Shapley propose une méthode de répartition "équitable" des gains à la coalition des n joueurs (Shapley 1953). Pour son utilisation dans le domaine de l'explicabilité, on fait un parallèle entre la prédiction et un jeu :

- Chaque ligne d (de vecteur de valeurs $x = (x_1, \cdots, x_n)$) constitue un jeu coopératif J_d dont le gain est la valeur prédite nette : $g_d = f(x_1, \cdots, x_n) - \overline{f}$ où $\overline{f} = \mathbb{E}[f(\mathbf{X})]$ est la valeur moyenne de $f(.)$ sur l'ensemble de la base.
- Chaque sous-ensemble de variables est une coalition dans le jeu J_d.
- Enfin la fonction caractéristique de J_d est $v_d(.)$ qui vérifie $v_d(X1, \cdots, X_n) = g_d$ et $v_d(\emptyset) = 0$.

  Pour tout sous-ensemble $A \subseteq X$, $v_d(A)$ est une statistique calculée sur la sous-base vérifiant $A = x_A$ (les valeurs des variables de A sont fixés à leur valeur dans x).

[1]Pour alléger la notation, on notera $\mathbf{x}$ au lieu de $\mathbf{x_d}$ quand cela n'entraînera pas d'ambiguïté.

- Nous verrons dans la section , la caractérisation exacte de v_d qui produira différents types de valeur de Shapley.
- On peut alors calculer une valeur de Shapley pour le jeu J_d qui explique l'apport de chaque variable à la construction de g_d (voir la figure 2) :

$$\phi_i(\mathbf{x_d}) = \sum_{S \subseteq \mathbf{X}/\{X_i\}} w_X(S)\big[v_d(S \cup \{X_i\}) - v_d(S)\big] \quad (1)$$

avec

$$w_X(S) = \frac{|S|!(|\mathbf{X}| - |S| - 1)!}{|\mathbf{X}|!}$$

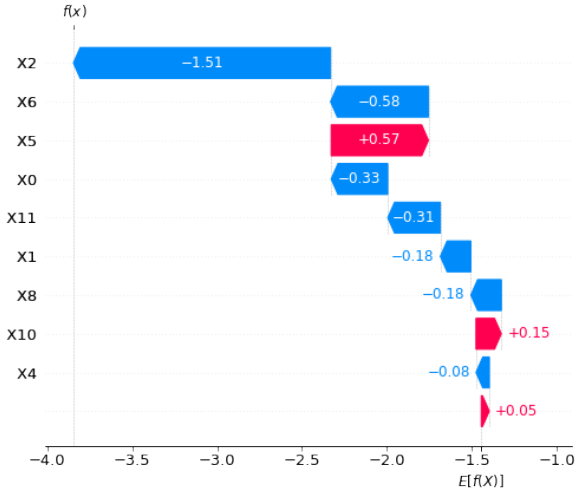


Figure 2: Diagramme *Waterfall* qui présente les contributions $\phi_i(x)$ pour chaque variable X_i permettant de construire $f(x) = logit(p(x))$ (en haut) en partant de $\overline{f}$ (en bas)[2].

Afin de pouvoir comparer (et trier) globalement les variables à partir de ces calculs, on définit une notion d'importance associée aux valeurs de Shapley.

Definition 2 ***Importance d'une variable*** *L'importance de chaque variable est la moyenne de ses valeurs de Shapley sur l'ensemble des jeux de la base. (voir les diagrammes de la figure 3) :*

$$\Phi_i = \frac{1}{D} \sum_{d=1}^{D} |\phi_i(\mathbf{x_d})| \quad (2)$$

Dans une approche axiomatique, (Young 1985) a montré que seuls les valeurs de Shapley forment un modèle d'explication qui respecte la décomposition additive (définition 1) ainsi que les propriétés suivantes :

Propriété 1 *Efficience*

$$\sum_{i=1}^{n} \phi_i(x) = v(x) - v(\emptyset)$$

[2]Les diagrammes des figures 2 et 3 sont générés avec la librairie SHAP (Lundberg et al. 2020).

la somme des valeurs attribuées aux variables doit être égale à ce que la coalition de toutes les variables peut obtenir (i.e la moyenne des prédictions sur la base de données).

Propriété 2 *Symétrie*

Les contributions de deux variables i et j doivent être les mêmes si elles contribuent de manière égale à toutes les coalitions possibles. Si :

$$\forall S \subseteq \mathbf{X}/\{X_i, X_j\}, v_d(S \cup \{X_i\}) = v_d(S \cup \{X_j\})$$

alors

$$\phi_i(x) = \phi_j(x)$$

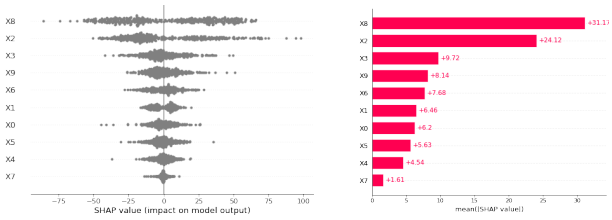
Propriété 3 *Nullité*

*Une variable X_i qui ne change pas la valeur prédite (quelle que soit la coalition de variables à laquelle elle est ajoutée) doit avoir une valeur de Shapley **nulle**.*

$$\forall S \subseteq \mathbf{X}, v_d(S \cup \{X_i\}) = v_d(S)$$

alors

$$\phi_i(x) = 0$$



(a) Diagramme *beeswarm* chaque point représente une valeur de Shapley calculée pour une ligne de la base.

(b) Diagramme d'importances. Chaque variable est associée à la moyenne des valeurs absolues du *beeswarm*.

Figure 3: Diagramme des valeurs de Shapley

Pour calculer les valeurs de Shapley, nous avons vu qu'il suffit d'un modèle de prédiction ($f(.)$). Ainsi peut-on les adapter à beaucoup de modèle dont les réseaux bayésiens. On notera aussi qu'une base de données assez conséquente peut suffire aux calculs des valeurs de Shapley. En effet, on peut calculer les différentes probabilités et statistiques nécessaires directement de manière fréquentiste sur la base de données.

Principaux types de valeurs de Shapley

Dans la section précédente, nous avons vu que la construction des valeurs de Shapley dépend largement du choix de la statistique utilisée dans les fonctions v_d. Dans cette section, nous présentons 3 types des valeurs de Shapley et les v_d associés.

Fonction de prédiction additive La fonction de prédiction d'un classifieur probabiliste est le plus souvent une distribution de probabilité sur l'ensemble des classes. L'idée même d'avoir une décomposition additive d'une telle fonction de prédiction peut alors sembler incongrue. C'est pourquoi, naturellement, on utilise régulièrement le log de cette distribution comme fonction de prédiction :

$$f(x_d) = logit\left(p(C|x_d)\right) \tag{3}$$

Les valeurs de Shapley conditionnelles et marginales

Pour pouvoir calculer les valeurs de Shapley, nous devons donc définir la fonction de caractéristique $v_d(.)$ pour un certain sous-ensemble $S \subseteq \mathbf{X}$. Par définition, cette fonction cherche à approcher une valeur de $f(x_d)$ lorsque nous ne connaissons que la valeur du sous-ensemble S. En particulier, $v_d(x) = f(x)$. À cette fin, Lundberg and Lee (2017) utilisent la sortie attendue du modèle prédictif, conditionné sur les variables $\mathbf{X_S} = \mathbf{x_S}$, nous ferons référence à cet objet comme étant les valeurs de Shapley conditionnelles :

$$\begin{aligned} v_d^{cond}(S) &= \mathbb{E}\big[f(X_{\overline{S}}, x_S)|X_S = x_S\big] \\ &= \int f(X_{\overline{S}}, x_S)p(X_{\overline{S}}|X_S = x_S)dX_{\overline{S}} \end{aligned} \tag{4}$$

où $\overline{S} = \mathbf{X} \setminus S$

Toutefois, il s'avère que le calcul exact de ces valeurs de Shapley est extrêmement coûteux en temps. En utilisant les propriétés des modèles additifs (définition 1) et en supposant l'indépendance des variables, Lundberg and Lee 2017 proposent une approximation plus facile à calculer. On y fera référence comme étant les valeurs de Shapley marginales, la fonction caractéristique $v_d(S)$ devient alors :

$$v_d^m(S) = \mathbb{E}\big[f(X_{\overline{S}}, x_S)\big] \tag{5}$$

Il est à noter que, par exemple, Aas, Jullum, and Løland (2020) soutiennent et illustrent que les valeurs de Shapley marginales peuvent conduire à des interprétations erronées lorsque les variables sont fortement corrélées. Ils proposent donc de conserver la notion de valeurs de Shapley conditionnelles (équation 4) et de plutôt se focaliser sur des méthodes d'approximations de cette fonction.

Les valeurs de Shapley interventionnelles et causales

Datta, Sen, and Zick (2016), Janzing, Minorics, and Blöbaum (2019), Sundararajan Mukund (2020), proposent d'introduire la causalité dans le calcul des valeurs de Shapley, en remplaçant le conditionnement observationnel par un conditionnement par intervention. En arguant que, si le but est d'expliquer causalement le modèle, les entrées du modèle peuvent être formellement distinguées des variables du monde réel, on peut alors considérer ces entrées comme indépendantes. Ainsi, les valeurs de Shapley interventionnelles se simplifient en valeurs de Shapley marginales et l'utilisation d'un cadre tel que le *do-calculus* de Pearl (Pearl 2012) n'est plus nécessaire. Même si l'argument semble faible, il est tout de même mis en avant également par Lundberg and Lee (2017) lors du développement du framework *interventional Tree SHAP* (Lundberg et al. 2020).

Une autre méthode proposée pour introduire de la causalité dans les valeurs de Shapley est celle de Frye, Rowat, and Feige (2020). L'idée est d'incorporer de la connaissance causale du monde réel en limitant les permutations possibles sur les variables. Le raisonnement est le suivant :

si d'après la connaissance causale externe, on sait que X_i est l'ancêtre causal et détermine totalement X_j, on pourrait alors vouloir attribuer toute la contribution prédictive à X_i et aucune à X_j. Pour ce faire, on ne considère que les permutations qui sont cohérentes avec cet ordre causal. Les auteurs considèrent toutefois qu'un conditionnement interventionnel a le désavantage de pouvoir extrapoler hors du *manifold* des données (i.e. sur une zone de données que le modèle n'a jamais exploré). C'est pourquoi, en accord avec Aas, Jullum, and Løland (2020), ils proposent d'utiliser néanmoins un conditionnement observationnel.

Enfin, non content de considérer la causalité lors de l'interprétation des valeurs de Shapley, Heskes et al. (2020) proposent de prendre en compte un modèle causal exhaustif qui leur permet d'introduire les valeurs de Shapley **causale**. La fonction caractéristique v s'écrit alors :

$$\begin{aligned} v_d^{caus}(S) &= \mathbb{E}\big[f(X_{\overline{S}}, x_S)|do(X_S = x_S)\big] \\ &= \int f(X_{\overline{S}}, x_S)p(X_{\overline{S}}|do(X_S = x_S))dX_{\overline{S}}. \end{aligned} \tag{6}$$

où la notation $do()$ est bien celle du *do-calculs* de Pearl (Pearl 2012).

Les auteurs utilisent les distributions de probabilité interventionnelles, comme proposé par (Janzing, Minorics, and Blöbaum 2019), pour calculer le résultat attendu du modèle. Cette méthode prend en compte l'effet qu'une variable a sur la sortie via d'autres variables d'entrée, mais ne tient pas compte des effets de *confunding* des autres variables. On notera que dans le cas où il n'y a pas de chemins causaux entre les variables d'entrée, l'opérateur do dans les valeurs de Shapley causales mènent aux valeurs de Shapley marginales. Cependant, lorsque les chemins causaux sont présents, cela conduit à un résultat différent, car les auteurs attribuent une valeur aux variables qui causent une différence dans la production et n'attribuent pas aux variables qui sont accidentellement corrélées avec d'autres variables qui influencent la production. Cette interprétation causale permet de distinguer les effets directs et indirects de chaque caractéristique sur la prédiction d'un modèle :

$$\begin{aligned} \delta_i &= v_d^{caus}(S \cup \{X_i\}) - v_d^{caus}(S) \\ \delta_i &= \mathbb{E}\left[f(X_{\overline{S}}, x_{\underline{S}\cup i})|do(X_{\underline{S}\cup i} = x_{\underline{S}\cup i})\right] \\ &\quad - \mathbb{E}\left[f(X_{\overline{S}\cup i}, x_{\underline{S}})|do(X_{\underline{S}} = x_{\underline{S}})\right] \text{ (total effect)} \\ &= \mathbb{E}\left[f(X_{\overline{S}}, x_{\underline{S}\cup i})|do(X_{\underline{S}} = x_{\underline{S}})\right] \\ &\quad - \mathbb{E}\left[f(X_{\overline{S}\cup i}, x_{\underline{S}})|do(X_{\underline{S}} = x_{\underline{S}})\right] \text{ (direct effect)} \\ &\quad + \mathbb{E}\left[f(X_{\overline{S}}, x_{\underline{S}\cup i})|do(X_{\underline{S}\cup i} = x_{\underline{S}\cup i})\right] \\ &\quad - \mathbb{E}\left[f(X_{\overline{S}}, x_{\underline{S}\cup i})|do(X_{\underline{S}} = x_{\underline{S}})\right] \text{ (indirect effect)} \end{aligned} \tag{7}$$

où $\underline{S} = S \cap Anc(C)$ (i.e. les causes de C inclus dans S).

L'effet direct mesure le changement attendu dans la prédiction lorsque la caractéristique stochastique X_i est remplacée par sa valeur de caractéristique x_i , sans changer la distribution des autres caractéristiques *'hors-coalition'*.

L'effet indirect mesure la différence d'espérance lorsque la distribution des autres caractéristiques *'hors-coalition'* change en raison de l'intervention supplémentaire $do(X_i = x_i)$.

Valeurs de Shapley dans les réseaux bayésiens

Il est facile de faire d'un réseau bayésien (BN) une fonction de prédiction en s'intéressant plus particulièrement à une variable cible C et en spécialisant les requêtes d'inférence sur l'estimation de la distribution de C a posteriori. Il est donc possible de croiser les méthodes issues des BNs et les méthodes des classifieurs probabilistes, comme les valeurs de Shapley.

Le calcul des valeurs de Shapley tels que présentés dans la section précédente sont clairement coûteuses ; principalement en temps puisqu'il faut itérer pour chaque variable sur l'ensemble des sous-ensembles de $\mathbf{X}$. Donc, une complexité de l'ordre de $n.2^n.K(n)$ où $K(n)$ est la complexité de l'évaluation de $v_d(\mathbf{x})$.

Dans le cadre du *Machine Learning*, des méthodes plus rapides et approchées ont pu être proposées comme le framework SHAP (Lundberg and Lee 2017) qui s'appuie sur des spécificités du calcul pour certains modèles comme, par exemple, pour les *Random Forest* avec *TreeShap* (Lundberg et al. 2020).

Dans le cadre des réseaux bayésiens, le calcul peut également être optimisé fortement à plusieurs niveaux. Par exemple, certains types de valeurs de Shapley seront forcément nuls pour certaines variables en fonction de la topologie du BN. Une autre optimisation aisée vient du fait qu'une analyse graphique peut permettre d'identifier des termes de l'équation 1 qui s'annulent sans avoir à les calculer : $[v_d(S \cup \{X_i\}) - v_d(S)]$ est en effet nul si X_i est indépendant de C conditionnellement à S.

Une autre approximation possible consiste à remarquer que le calcul d'une espérance de probabilité peut être optimisée par inférence (Madsen and Jensen 1999) contrairement au calcul d'une espérance de log de probabilité. L'approximation consiste donc à remplacer l'espérance d'un log par le log de l'espérance et permet un calcul rapide d'une approximation de $v_d^{cond}(x)$ (l'erreur revient exactement à l'écart de Jensen). Expérimentalement, on note que cet écart est faible par rapport à la valeur des importances obtenues, et que donc, l'approximation n'a pas d'impact sur le rangement ultérieur des variables dans l'ordre d'importance (voir diagramme d'importance Figure 3b).

Enfin, des schémas d'approximation basés sur un échantillonnage sur toutes les permutations permettraient également de produire des versions de ces valeurs approchées par échantillonnage de Monte-Carlo. Nous n'avons actuellement pas implémentés ces algorithmes d'échantillonnage.

Dans cette section, nous allons présenter les différentes approximations et méthodes que nous proposons pour le calculer de valeurs de Shapley dans les réseaux bayésiens. Ces améliorations ont été implémentées dans le cadre de la librairie pyAgrum (Ducamp, Gonzales, and Wuillemin 2020). Nous avons étudié le calcul de valeurs de Shapley dans un BN dans les cas conditionnel, marginal et causal.

Valeurs de Shapley conditionnelles

Pour ce type de valeurs de Shapley, le calcul utilise les optimisations proposées ci-dessus mais ne possède aucune spécificité permettant d'améliorer encore les performances de l'algorithme. On peut remarquer qu'une variable non directement connectée à la variable cible (C) peut avoir une valeur de Shapley conditionnelle non nulle.

Valeurs de Shapley marginales

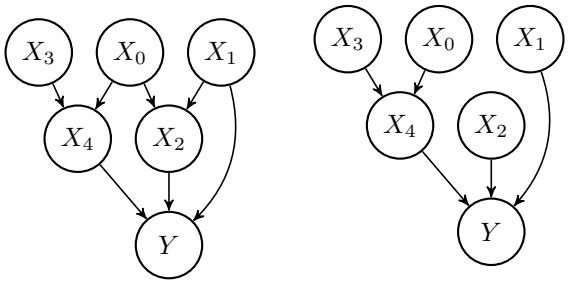
Pour les valeurs de Shapley marginales, nous avons basé notre approche sur la valeur de Shapley interventionnelle déjà existante dans le cadre SHAP (Lundberg et al. 2020). Nous avons juste adapté le calcul à la prédiction faite dans les réseaux bayésiens.

Il est à noter que par définition des valeurs de Shapley marginales, ces valeurs sont nulles pour toutes les variables qui ne font pas partie de la couverture de Markov de la variable d'intérêt C. En effet, dans l'équation 1, les termes $v_d^m(S \cup \{X\}) - v_d^m(S)$ seront nuls quand X ne fait pas partie de la couverture de Markov de C.

L'implémentation utilise bien évidemment cette propriété pour ne calculer que les contributions des variables de la couverture de Markov.

Valeurs de Shapley causales

D'une manière pragmatique, le calcul des valeurs de Shapley causale prend pour hypothèse l'exhaustivité du modèle causal (donc un modèle sans variable latente). C'est une hypothèse forte qui est donc utilisée dans cette implémentation. Sous cette hypothèse, l'opérateur *do* revient à *mutiler* le graphe causal (Pearl and Mackenzie 2018). En effet, il suffit de couper les arcs entre les parents et la variable sur laquelle on veut effectuer l'intervention. Les calculs se feront alors effectivement dans le BN mutilé ci-contre (voir figure 4b).



(a) Graphe initial

(b) Nouveau graphe avec une intervention sur X_2

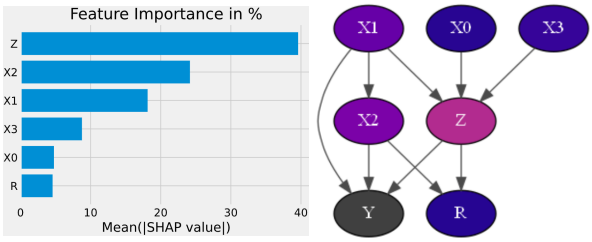
Figure 4: Exemple de mutilation de graphe

En répétant l'opération sur chaque variable, on peut alors calculer la valeur de Shapley *causale* comme décrit dans la section précédente .

Il est à noter également ici que les valeurs de Shapley causale pour les nœuds qui ne sont pas des ancêtres de la variable d'intérêt C seront nulles. La connaissance du graphe causal nous permet donc de ne calculer ces valeurs que pour les nœuds pertinents.

Visualisation

Une fois les valeurs de Shapley calculées, quelle que soit la méthode choisie et on peut assigner une couleur à chaque variable. Il suffit de prendre l'importance de la variable (voir figure 5a), comme décrit dans l'équation 2 et de la convertir en couleur (cf. figure 5).



(a) Importance de chaque variable calculée dans un BN.

(b) Affichage des importances de chaque variable sur le BN.

Figure 5: Visualisation sur le graphe du BN des importances de chaque variable.

L'ensemble des valeurs de Shapley peut donc être calculé avec plus ou moins de facilité dans les réseaux bayésiens. Ces calculs ont tous été implémentés dans le module `explain` de pyAgrum[3]. Leur optimisation n'est pas terminée et il reste certainement de grande marge d'améliorations.

Complémentarité *Machine Learning* et *BNs*

On oppose facilement les deux types d'apprentissage depuis une base de donnée : d'un côté, le *Machine Learning* qui s'attache principalement à améliorer la précision dans une prédiction ; et de l'autre, les approches *Model-based* (comme les BNs) qui privilégient l'extraction de connaissance et la compréhension du processus dont est issue la base ; la prédiction ne faisant que découler de cette modélisation. Comme on l'a vu précédemment, l'un des intérêts des valeurs de Shapley est de se trouver à l'interface entre les 2 types d'apprentissages ; apportant une information quantitative dans la représentation qualitative de la structure d'un BN et apportant une information qualitative dans la représentation quantitative du modèle prédictif. Il en découle un lien entre le domaine prédictif et le domaine explicatif que nous avons l'intention d'explorer le plus complètement possible. Cette section propose des réflexions préliminaires sur ce sujet.

Du *Model-based* au *Machine Learning*

Les valeurs de Shapley (et les importances) sont calculables pour l'ensemble des *features* de la base (les variables X_i). Toutefois, la connaissance d'un modèle graphique (voire causal) permet d'affiner ces calculs selon 2 modes principaux : (i) le modèle graphique permet l'optimisation de certains calculs comme nous l'avons vu précédemment ; mais surtout (ii) le modèle graphique permet de sélectionner les

[3] https://pyagrum.readthedocs.io/en/latest/lib.explain.html#dealing-with-shapvalues.

variables pour lesquelles il est intéressant de mener ces calculs. En effet, une analyse purement prédictive peut amener à des interprétations erronées. On peut notamment distinguer le cas courant où l'on donne toutes les variables sans aucune distinction au modèle prédictif ; cette méthode va donc donner des variables qui peuvent être les conséquences de la variable intérêt C comme entrée du modèle (dont la tâche est de prédire C). On pourrait alors retrouver des conséquences de C avec une importance élevée alors que ces dernières n'ont aucun apport prédictif.

L'apport du *Machine Learning* pour les *BNs*

Une première idée d'un apport des valeurs de Shapley pour les BNs part d'un constat simple portant sur les valeurs de Shapley marginales. En effet, ces dernières donnent une valeur nulle aux variables qui n'ont pas de relations directes avec la variable cible. Ainsi en entrainant un modèle pour la prédiction d'une variable cible C, tel qu'un *Random Forest*, puis en calculant les valeurs de Shapley marginales, on peut retrouver les nœuds de la couverture de Markov de C.

Récupérer les liens directs d'une variable permet donc de retrouver ses voisins proches. Il en découle un algorithme qui répéterait cette opération, pour chaque variable et permettant donc de retrouver partiellement le graphe dans son ensemble. Cette méthode serait à rapprocher des algorithmes d'apprentissages de réseau bayésien basés sur les couvertures de Markov (Bui and Jun 2012; Gao and Ji 2017).

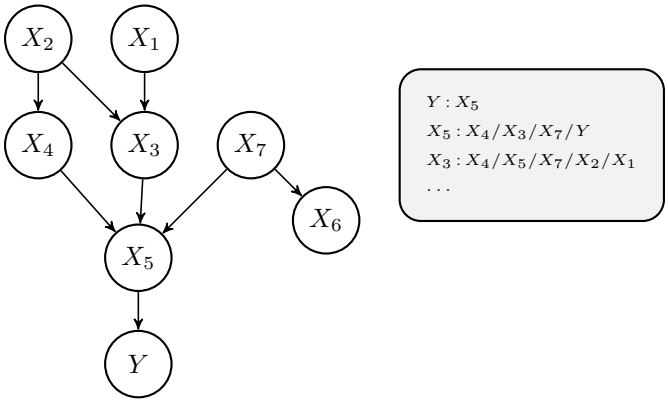


Figure 6: Un réseau bayésien et la découverte des couvertures de Markov qui pourrait être réalisée à partir des valeurs de Shapley marginales calculées (par exemple) grâce à un modèle de *Random Forest*.

Dans le graphe 6, on peut imaginer de créer un modèle qui sert à la prédiction de Y, théoriquement les valeurs de Shapley marginales nous permettent de retrouver le lien avec X_5. Puis on entraîne un modèle sur $X5$, qui nous donne les relations décrites dans le rectangle gris en dessous. De proche en proche on pourra retrouver les différentes couvertures de Markov de chaque variable, une fois celles-ci connues nous pouvons utiliser un algorithme type PC (Spirtes, Glymour, and Scheines 1993) ou MIIC (Verny et al. 2017) afin de retrouver l'orientation des liens dans chacune de ces couvertures de Markov.

Conclusion

Les valeurs de Shapley héritées de la théorie des jeux s'imposent de plus en plus comme un outil populaire d'explication de modèle de *Machine Learning*. Malgré le fait que leurs interprétations soient aisées, elles induisent des temps de calculs coûteux. Ce problème peut être partiellement levé notamment grâce à certaines approximations faites sur le modèle prédictif. Les réseaux bayésiens peuvent être utilisés comme des modèles prédictifs et sont donc aptes à permettre le calcul des valeurs de Shapley. Cet article propose quelques pistes d'optimisations de ces calculs en utilisant la topologie et les indépendances encodés dans le BN. Par ailleurs, cet article propose de considérer les valeurs de Shapley comme une articulation entre le domaine du prédictif et le domaine de la modélisation et d'en tirer des nouvelles techniques en utilisant cet outil pour tirer parti simultanément de ces deux domaines. Il illustre cette proposition par deux exemples : (i) effectuer une sélection des variables à intégrer dans un modèle explicatif ou alors (ii) identifier les couvertures de Markov dans le cadre d'un algorithme d'apprentissage de la structure d'un BN. Les travaux futurs envisagés consistent bien évidemment à continuer d'explorer l'utilisation des valeurs de Shapley dans ce cadre unificateur entre prédiction et modélisation, mais aussi d'étendre ce type d'analyse à d'autres outils issus du domaine de l'explication dans les modèles prédictifs.

Remerciements

Ce travail a été effectué dans le cadre d'une thèse CIFRE (no2020/1640) soutenue par SAP et l'ANRT (Association Nationale de la Recherche et de la Technologie).

References

Aas, K.; Jullum, M.; and Løland, A. 2020. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values.

Breiman, L. 2001. Random Forests. *Machine Learning* 45(1): 5–32. ISSN 1573-0565. doi:10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Bui, A. T.; and Jun, C.-H. 2012. Learning Bayesian network structure using Markov blanket decomposition. *Pattern Recognition Letters* 33(16): 2134–2140. ISSN 0167-8655. doi:https://doi.org/10.1016/j.patrec.2012.06.013. URL https://www.sciencedirect.com/science/article/pii/S0167865512002012.

Burkart, N.; and Huber, M. F. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research* 70: 245–317. doi:10.1613/jair.1.12228. URL https://doi.org/10.1613/jair.1.12228.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. IEEE. ISBN 9781509008247. doi:10.1109/SP.2016.42.

Ducamp, G.; Gonzales, C.; and Wuillemin, P.-H. 2020. aGrUM/pyAgrum : a toolbox to build models and algorithms for Probabilistic Graphical Models in Python. In *10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, 609–612. Skørping, Denmark. URL https://hal.archives-ouvertes.fr/hal-03135721.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5): 1189 – 1232. doi:10.1214/aos/1013203451.

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability.

Gao, T.; and Ji, Q. 2017. Efficient score-based Markov Blanket discovery. *International Journal of Approximate Reasoning* 80: 277–293. ISSN 0888-613X. doi:https://doi.org/10.1016/j.ijar.2016.09.009. URL https://www.sciencedirect.com/science/article/pii/S0888613X1630161X.

Heskes, T.; Sijben, E.; Bucur, I. G.; and Claassen, T. 2020. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models.

Hooker, G.; and Mentch, L. 2019. Please Stop Permuting Features: An Explanation and Alternatives. *arXiv:1905.03151 [cs, stat]* URL http://arxiv.org/abs/1905.03151. ArXiv: 1905.03151.

Janzing, D.; Minorics, L.; and Blöbaum, P. 2019. Feature relevance quantification in explainable AI: A causal problem.

Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1): 2522–5839.

Madsen, A. L.; and Jensen, F. V. 1999. Lazy propagation: A junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence* 113(1): 203–245. ISSN 0004-3702. doi:10.1016/S0004-3702(99)00062-4.

Miller, T. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences.

Molnar, C. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Pearl, J. 2012. The Do-Calculus Revisited.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why*. New York: Basic Books. ISBN 978-0-465-09760-9.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Rieg, T.; Frick, J.; Baumgartl, H.; and Buettner, R. 2020. Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. *PLOS ONE* 15(12): 1–20. doi:10.1371/journal.pone.0243615. URL https://doi.org/10.1371/journal.pone.0243615.

Shapley, L. S. 1953. *17. A Value for n-Person Games*, 307–318. Princeton University Press. doi:10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*, volume 81. ISBN 978-1-4612-7650-0. doi:10.1007/978-1-4612-2748-9.

Sundararajan Mukund, A. N. 2020. The Many Shapley Values for Model Explanation 119: 9269–9278. URL http://proceedings.mlr.press/v119/sundararajan20b.html.

Verny, L.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology* 13(10): e1005662. doi:10.1371/journal.pcbi.1005662. URL https://doi.org/10.1371/journal.pcbi.1005662.

Young, H. P. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14(2): 65–72. ISSN 1432-1270. doi:10.1007/BF01769885. URL https://doi.org/10.1007/BF01769885.