Interpretability methods in AI and a comparison with sensitivity analysis

C. Labreuche [1]

[1] Thales Research & Technology
Palaiseau, France
email: christophe.labreuche@thalesgroup.com

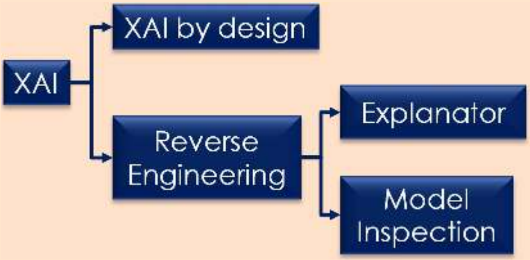
Outline

Why shall we explain decisions?

Why is “explaining” important?

- Man-Machine Interaction: Increase acceptance & trust of user
- Trustable AI: Validation and qualification for safety-critical systems

Taxonomy of XAI [Guidotti et al’18]



Explanation by Feature Attribution

Aim

Feature Attribution:

- Given an AI box with inputs and outputs,
- identify the input variables that mostly influence the outputs.

- Done by calculating the impact level of each input variable on the outputs.

Scope: numerical functions

- Filter relevant information/motivation to be presented to the user;
- Debugging mode in Machine Learning (inputs = features).

Decision setting

Decision setting

- $N = \{1, \ldots, n\}$: index set of attributes/features.
- X_i: set of values representing attribute i (for $i \in N$).
- $X = X_1 \times \cdots \times X_n$: set of alternatives/acts.
- $U : X \to \mathbb{R}$: utility representing preferences of decision maker over X
    - $U(y) > U(x)$: y is preferred to x
- Decision problems:
    - Selection: find the best element in $\mathcal{X} \subseteq X$
    - Ranking: order the elements of $\mathcal{X} \subseteq X$
    - Scoring: assign a score to each element of $\mathcal{X} \subseteq X$
    - Sorting: assign each element of $\mathcal{X} \subseteq X$ to a class $\mathcal{C}$

Outline

Outline

Why shall we explain decisions?

A simple example

Function of 3 binary variables:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$

How to explain the difference between

- $x = (0, 0, 0)$, with $u(x) = 0$
- and $y = (1, 1, 1)$, with $u(y) = 7$?

A simple problem? NO!

- A simple Gradient does not work
  - it is unstable!
- Figures shall have a meaning
- There are interactions among the inputs

Idea

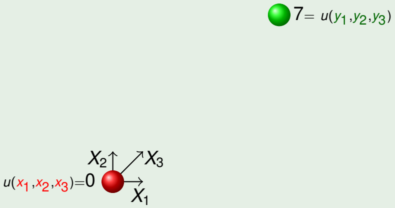
Idea of the approach

- How to isolate the contribution of each input variable?
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.

A simple example

Function of 3 binary variables, with $x = (0, 0, 0)$ and $y = (1, 1, 1)$:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$



Idea

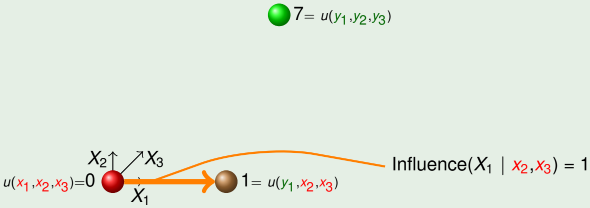
Idea of the approach

- How to isolate the contribution of each input variable?
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.

A simple example

Function of 3 binary variables, with $x = (0, 0, 0)$ and $y = (1, 1, 1)$:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$



Idea

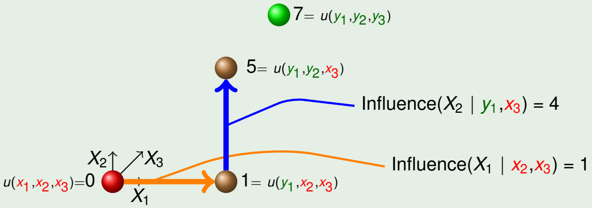
Idea of the approach

- How to isolate the contribution of each input variable?
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.

A simple example

Function of 3 binary variables, with $x = (0, 0, 0)$ and $y = (1, 1, 1)$:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$



Idea

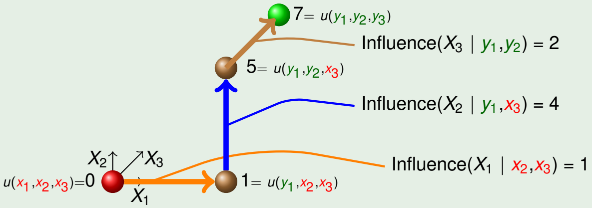
Idea of the approach

- How to isolate the contribution of each input variable?
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.

A simple example

Function of 3 binary variables, with $x = (0, 0, 0)$ and $y = (1, 1, 1)$:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$



Idea

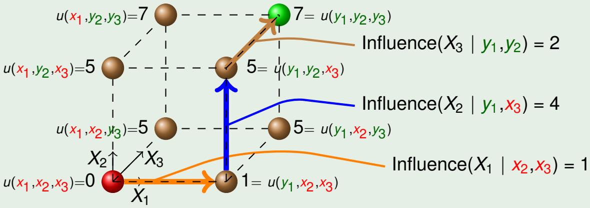
Idea of the approach

- How to isolate the contribution of each input variable?
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.

A simple example

Function of 3 binary variables, with $x = (0, 0, 0)$ and $y = (1, 1, 1)$:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$



Idea

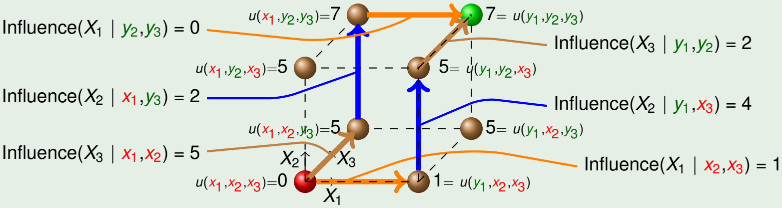
Idea of the approach

- How to isolate the contribution of each input variable?
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.

A simple example

Function of 3 binary variables, with $x = (0, 0, 0)$ and $y = (1, 1, 1)$:

$$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\ \max(x_2, x_3) + 2\ \min(x_2, x_3)$$



Conversion to Cooperative Game Theory

Feature attribution

	Game Theory	Decision
N	players	attributes
$v : 2^N \to \mathbb{R}$	game, with $v(\emptyset) = 0$	$v(S) = u(y_S, x_{N \setminus S}) - u(x)$
$\phi \in \mathbb{R}^N$	imputation	feature importance
Efficiency	$\sum_{i \in N} \phi_i = v(N) - v(\emptyset)$	

A simple example

- Approach 1: $\phi_i = \frac{v(N)}{n}$

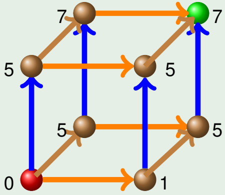
  $$\phi = (7/3, 7/3, 7/3)$$

- Approach 2: $\phi_i = v(\{i\}) - \frac{v(N) - \sum_k v(\{k\})}{n}$

  $$\phi = (-1/3, 11/3, 11/3)$$

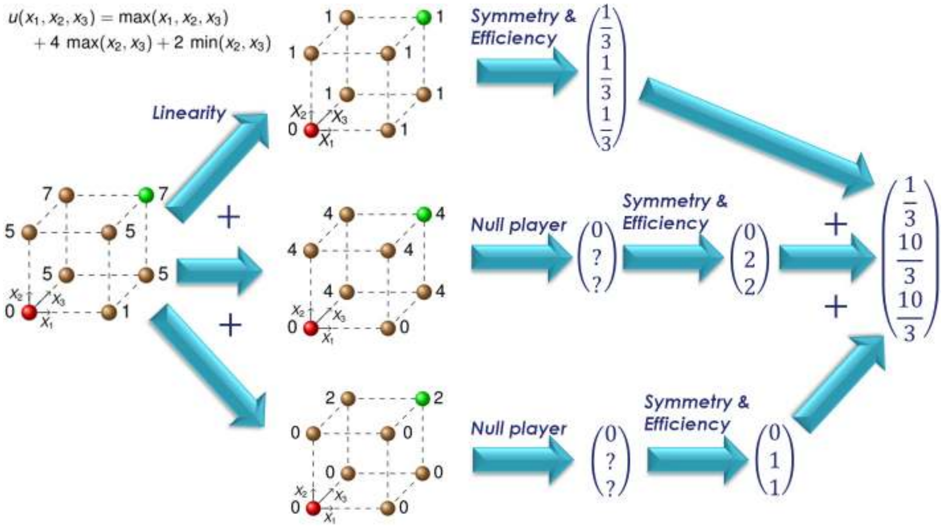
- Approach 3: $\phi_i = \frac{1}{2^{n-1}} \sum_{S \subseteq N \setminus i} (v(S \cup \{i\}) - v(S))$

  $$\phi = (1/4, 13/4, 13/4)$$



Axioms

$u(x_1, x_2, x_3) = \max(x_1, x_2, x_3) + 4\max(x_2, x_3) + 2\min(x_2, x_3)$



Characterization result

Characterization of the Shapley value [Shapley'53]

There is only one imputation ϕ which satisfies to the following properties:

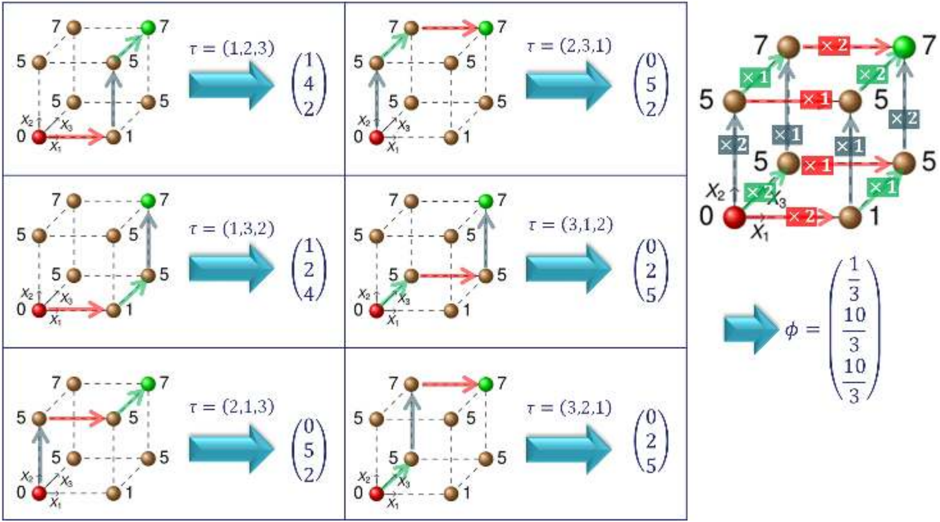
- Additivity: $\phi_i(N, v + w) = \phi_i(N, v) + \phi_i(N, w)$,
- Null player: if $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N \setminus \{i\}$, then $\phi_i(N, v) = 0$,
- Symmetry: $\phi_{\pi k}(\pi N, \pi v) = \phi_k(N, v)$ for every permutation π on N,
- Efficiency: $\sum_{i \in N} \phi_i(N, v) = v(N)$.

It is equal to:

$$\phi_i(N, v) = \mathrm{Sh}_i(N, v) := \frac{1}{n!} \sum_{\tau \in \Pi(N)} \left[v(\{\tau(1), \ldots, i\}) - v(\{\tau(1), \ldots, \tau(\tau^{-1}(i) - 1)\}) \right]$$

$$= \sum_{S \subseteq N \setminus i} \frac{(n - |S| - 1)!|S|!}{n!} \left[v(S \cup \{i\}) - v(S) \right].$$

Shapley value



Outline

Absolute Assessment In Decision

Basic Idea

Compare x to a reference r (e.g. expectation from user).

Drowning effect

Function $u(x_1, x_2) = \min(x_1, x_2)$, with $x = (0.2, 0.8)$.
Choice of reference r:

- $r = (0, 0)$ vs. x:

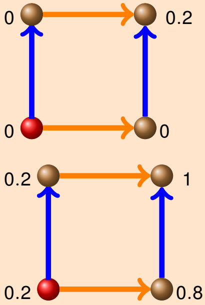
$$\phi_1 = \phi_2 = \frac{1}{2} \min(x_1, x_2) = 0.1$$

  Same importance for the two attributes $\forall x_1, x_2$!

- x vs. $r = (1, 1)$:

$$\phi_1 = \frac{1}{2} \left[1 - x_1 + x_2 - \min(x_1, x_2)\right] = 0.7$$

$$\phi_2 = \frac{1}{2} \left[1 - x_2 + x_1 - \min(x_1, x_2)\right] = 0.1$$



Outline

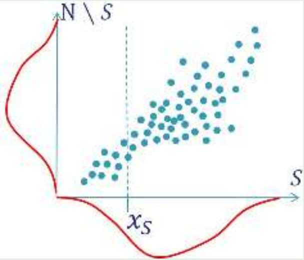
Feature attribution in Machine Learning

Notation

- $\mathcal{D}$: distribution of elements $x \in X$.
- $\mathcal{D}^{\text{JM}} = \prod_{i=1}^n \mathcal{D}_i^{\text{JM}}$, $\mathcal{D}_i^{\text{JM}}$ has the same marginal distribution than $\mathcal{D}$ over variable i
- $\mathcal{U}$: uniform distribution.

How to define the game? [Merrick, Taly'20] [Kumar et al'20]

Feature Attribution:



- Interventional distribution:
  - KernelSHAP [Lundberg, Lee'17]:
    $v(S) = \mathbb{E}_{R\sim\mathcal{D}}\left[u(x_S, R_{N\setminus S})\right] - \mathbb{E}_{R\sim\mathcal{D}}\left[u(R)\right]$
  - QII [Datta et al'16]:
    $v(S) = \mathbb{E}_{R\sim\mathcal{D}^{\text{JM}}}\left[u(x_S, R_{N\setminus S})\right] - \mathbb{E}_{R\sim\mathcal{D}^{\text{JM}}}\left[u(R)\right]$
  - IME [Strumbelj et al'10]:
    $v(S) = \mathbb{E}_{R\sim\mathcal{U}}\left[u(x_S, R_{N\setminus S})\right] - \mathbb{E}_{R\sim\mathcal{U}}\left[u(R)\right]$
- Conditional distribution: SHAP [Lundberg, Lee'17], TreeSHAP [Lundberg et al'18]

$$v(S) = \mathbb{E}_{R\sim\mathcal{D}}\left[u(x_S, R_{N\setminus S}) | R_S = x_S\right] - \mathbb{E}_{R\sim\mathcal{D}}\left[u(R)\right]$$

Outline

Shapley value in Sensibility Analysis

When variables are independent

- Functional ANOVA of $Y = u(X)$:

$$u(x) = \sum_{A\subseteq N} u_A(x_A) \ , \ u_A(x_A) = \sum_{B\subseteq A} (-1)^{|A\setminus B|} \ \mathbb{E}_{N\setminus B}(u|x_B)$$

$$\mathrm{Var}(Y) = \sum_{A\subseteq N} \mathrm{Var}_A(u_A(X_A))$$

- Sobol index $S_A = \frac{\mathrm{Var}_A(u_A(X_A))}{\mathrm{Var}(Y)}$, with $\sum_{A\subseteq N} S_A = 1$.

When variables are dependent [Owen'14]

- $\sum_{A\subseteq N} S_A \neq 1$
- Game (with $v(\emptyset) = 0$ and $v(N) = 1$)

$$v(A) = \frac{\mathrm{Var}_A[\mathbb{E}_{N\setminus A}(Y|X_A)]}{\mathrm{Var}(Y)}$$

- Contribution of variable i in $\mathrm{Var}(Y)$

$$\mathrm{Sh}_i(N, v)$$

Outline

Outline

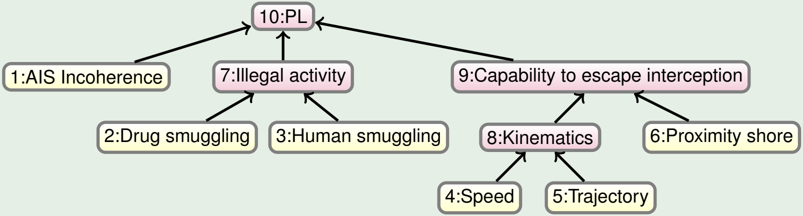
Example of application

Maritime Patrol

Mission of Maritime Patrol:

- monitor a maritime area,
- and seek for illegal activity.
- $\Rightarrow$ It evaluates in real time a Priority Level (PL) associated to each ship in this area

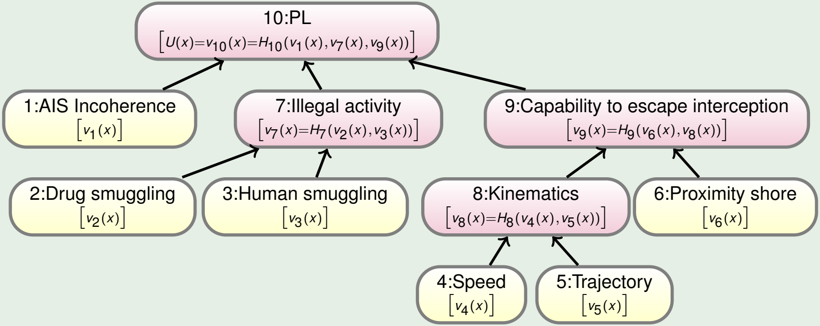
PL is intrinsically based on multiple criteria:



Hierarchical evaluation

Maritime Patrol

8. Kinematics: $8 \approx 4 \wedge 5$: $v_8(x) = 0.3 v_4(x) + 0.7 v_4(x) \wedge v_5(x)$
    - Complementarity & Speed slightly more important
10. $10 \approx 9 \wedge (1 \vee 7)$: $U(x) = v_{10}(x) = \big(v_1(x) \vee v_7(x) + v_1(x) \wedge v_9(x) + v_7(x) \wedge v_9(x)\big)/3$
    - There is suspicion of illegal activity when either 1 or 7 are satisfied;
    - We also need to have a risk of missed interception to get high PL;



Why not using the standard Shapley value?

Shapley value approach on trees

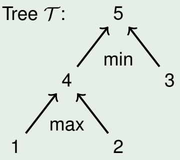
- Use the Shapley value on the leaves
- Use a recursive formulae otherwise: $I_i(x, y) = \sum_{j \in C(i)} I_j(x, y)$

Illustration

Comparison between $x = (0, 0, 0)$ and $y = (1, 1, 1)$.
Use of Shapley value on tree $\mathcal{T}$:

- $I_1(x, y) = I_2(x, y) = 1/6$, $I_3(x, y) = 2/3$
- $I_4(x, y) = I_1(x, y) + I_2(x, y) = 1/3$
- $I(x, y) = (1/6, 1/6, 2/3, 1/3, 1)$

Tree $\mathcal{T}$:



Why not using the standard Shapley value?

Shapley value approach on trees

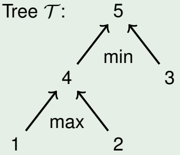
- Use the Shapley value on the leaves
- Use a recursive formulae otherwise: $I_i(x, y) = \sum_{j \in C(i)} I_j(x, y)$

Illustration

Comparison between $x = (0, 0, 0)$ and $y = (1, 1, 1)$.
Use of Shapley value on tree $\mathcal{T}$:

- $I_1(x, y) = I_2(x, y) = 1/6$, $I_3(x, y) = 2/3$
- $I_4(x, y) = I_1(x, y) + I_2(x, y) = 1/3$
- $I(x, y) = (1/6, 1/6, 2/3, 1/3, 1)$

Tree $\mathcal{T}$:



On subtree $\mathcal{T}'$:

- On $\mathcal{T}'$: $I_3(x, y) = I_4(x, y) = 1/2$
- Nodes 1 and 2 shall share equally $I_4(x, y) = 1/2$
- $I(x, y) = (1/4, 1/4, 1/2, 1/2, 1)$

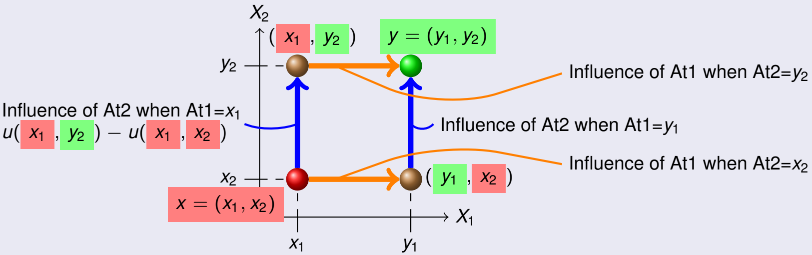
Tree $\mathcal{T}'$:

Outline

Axioms

Idea of the approach

- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Restricted Value (RV)

I_k depends only on the utility u of compound options mixing values of x, y.

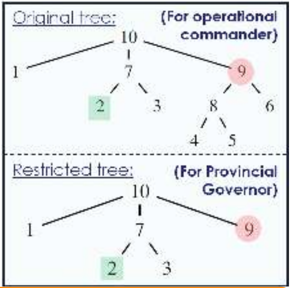
Axioms

Null Attribute (NA)

if changing x_k to y_k never changes u, then $I_k = 0$.

Consistency with Restricted Tree (CRT)

I_2 shall be the same for the original tree or a subtree where 9 becomes a leaf.



Generalized Efficiency (GE)

- General Share: $I_{10} = u(y) - u(x)$
- Decomposability: e.g. $I_9 = I_6 + I_8$

Other axioms

- Additivity (ADD): $I_k(u + u') = I_k(u) + I_k(u')$
- Restricted Equal Treatment (RET): All attributes are treated symmetrically

Are these axioms sufficient to derive I?

Theorem

There is a unique influence index satisfying **RV**, **NA**, **RET**, **ADD**, **GE** and **CRT**.

Remark

This influence index is an extension of the Shapley value on general trees.

Are these axioms sufficient to derive I?

Extended Shapley/Owen value

In order to distinguish the contribution of each attribute, we move from x to y changing one attribute at a time, following an ordering π on N:

$$x,\ (y_{\{\pi(1)\}}, x_{-\{\pi(1)\}}),\ (y_{\{\pi(1),\pi(2)\}}, x_{-\{\pi(1),\pi(2)\}}),\ \ldots,\ y.$$

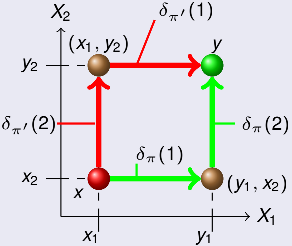
Definition:

$$I_i(x,y,T,u) = \begin{cases} \frac{1}{|\Pi(T)|} \sum_{\pi\in\Pi(T)} \delta_\pi(i) \text{ if } i \in N \\ \sum_{k\in \mathrm{Leaf}_T(i)} I_k(x,y,T,u) \text{ else} \end{cases}$$

$$\delta_\pi(i) := u(y_{S_\pi(i)}, x_{-S_\pi(i)}) - u(y_{S_\pi(i)\setminus\{i\}}, x_{-S_\pi(i)\setminus\{i\}}),\ S_\pi(\pi(k)) := \{\pi(1),\ldots,\pi(k)\}$$

Example with 2 attributes:

- Path #1, $\pi = (1,2)$:
  - for $\pi(1) = 1$: $\delta_\pi(1) = U(y_1, x_2) - U(x_1, x_2)$,
  - for $\pi(2) = 2$: $\delta_\pi(2) = U(y_1, y_2) - U(y_1, x_2)$
- Path #2, $\pi' = (2,1)$:
  - for $\pi'(1) = 2$: $\delta_{\pi'}(2) = U(x_1, y_2) - U(x_1, x_2)$,
  - for $\pi'(2) = 1$: $\delta_{\pi'}(1) = U(y_1, y_2) - U(x_1, y_2)$



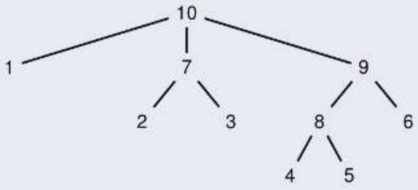
Are these axioms sufficient to derive I?

What is $\Pi(T)$?

$\Pi(T)$: set of orderings of elements of N for which all elements of a subtree of T are consecutive.

Example:

- $(5, 4, 6, 2, 3, 1) \in \Pi(T)$ (indicating that $\pi(1) = 5$, $\pi(2) = 4$, $\pi(3) = 6$, $\pi(4) = 2$, $\pi(5) = 3$, $\pi(6) = 1$)
- $(1, 6, 4, 5, 2, 3) \in \Pi(T)$
- $(1, 2, 3, 4, 5, 6) \in \Pi(T)$
- $(2, 3, 4, 5, 1, 6) \notin \Pi(T)$ since 1 is interleaved between attributes $\{4, 5\}$ and $\{6\}$



Computational complexity

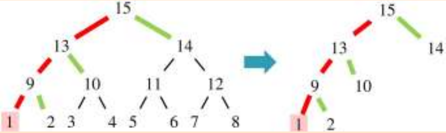
Complexity issue

Computation of I_i is exponential with n

Theorem

CRT implies that index I_i can be equivalently computed by cutting all branches not directly linking the path from node i to the root.

Example with I_1:



d	p	n	$\log_{10}\lvert\Pi(N)\rvert$	$\log_{10}\lvert\Pi(T)\rvert$	$\log_{10}\Pi(T_{[J]})$
2	2	4	1.38	0.903	0.602
2	3	9	5.559	3.112	1.556
2	4	16	13.320	6.901	2.76
2	5	25	25.19	12.47	4.158
2	6	36	41.57	20.0	5.715
3	2	8	4.605	2.107	0.903
3	3	27	28.036	10.115	2.334
3	4	64	89.1	28.984	4.14
3	5	125	209.27	64.454	6.237
3	6	216	412.0	122.86	8.571
4	2	16	13.3215	4.515	1.204
4	3	81	120.76	31.126	3.112
4	4	256	506.93	117.31	5.520
4	5	625	1477.7	324.35	8.316
4	6	1296	3473.0	740.04	11.429
5	2	32	35.42	9.332	1.505
5	3	243	475.76	94.156	3.89
5	4	1024	2639.7	470.65	6.901
5	5	3125	9566.3	1623.84	10.395
5	6	7776	26879	4443.15	14.286

Outline

Conclusion & Perspectives

Conclusion

- Shapley value is a generic tool to measure variable importance
- Extension to trees: an extensed Shapley value taking into accout the tree structure

Perspectives

- Further investigations between sensitivity analysis and interpretability

References

- [Kumar et al'20] E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. Friedler. *Problems with Shapley-value-based explanations as feature importance measures*. arxiv.org/abs/2002.11097
- [Labreuche et al'18] C. Labreuche, S. Fossier. *Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value*. IJCAI, 2018.
- [Datta et al'16] A. Datta, S. Sen, Y. Zick. *Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems*. IEEE symposium on security and privacy (SP), pp. 598-617, 2016.
- [Guidotti et al'18] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi. A survey of methods for explaining black box models. ACM Computing Surveys, 2018.
- [Lundberg, Lee'17] S. Lundberg, S. Lee. *A unified approach to interpreting model predictions*. NIPS, pp. 4765-4774, 2017.
- [Lundberg et al'18] S. Lundberg, G. Erion, S. Lee. *Consistent individualized feature attribution for tree ensembles*. arXiv:1802.03888, 2018.
- [Merrick, Taly'20] L. Merrick, A. Taly. *The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory*, arxiv.org/abs/1909.08128, 2020.
- [Owen'14] A.B. Owen. Sobol' indices and Shapley value. SIAM/ASA Journal on Uncertainty Quantification, 2014
- [Shapley'53] L.S. Shapley. A value for *n*-person games, 1953.
- [Strumbelj et al'10] E. Strumbelj, I. Kononenko. *An efficient explanation of individual classifications using game theory*. Journal of Machine Learning Research, 11, 1-18, 2010.