Logic-Based Explainable Artificial Intelligence

Joao Marques-Silva

ICREA, Univ. Lleida, Catalunya, Spain

ESSAI, Athens, Greece, July 2024

Lecture 05

Recapitulate fourth lecture

- Monotonic classifiers vs. weighted voting games
- Advanced topics:
  - Inflated explanations
  - Probabilistic explanations
  - Constrained explanations
  - Distance-restricted explanations
  - Explanations using surrogate models
  - Certified explainability

Monotonicity & WCGs

- Every WVG $\mathcal{G}$, described by $[q; n_1, \ldots, n_m]$, can be represented as a monotonically increasing boolean classifier $\mathcal{M} = (\mathcal{F}, \{0,1\}^m, \{0,1\}, \kappa)$, such that:
  - Each voter i is mapped to a boolean feature i, such that feature i takes value 1 if voter i votes Yes; otherwise it takes value 0;
  - The classification function $\kappa : \mathbb{F} \to \{0,1\}$ is defined by:

$$\kappa(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{m} n_i x_i \geqslant q \\ 0 & \text{otherwise} \end{cases}$$

  - The target instance is $(\mathbb{1}, 1)$; and
  - Each minimal winning coalition $\mathcal{C}$ corresponds to an AXp of $\mathcal{E} = (\mathcal{M}, (\mathbb{1}, 1))$

Monotonicity & WCGs

- Every WVG $\mathcal{G}$, described by $[q; n_1, \ldots, n_m]$, can be represented as a monotonically increasing boolean classifier $\mathcal{M} = (\mathcal{F}, \{0,1\}^m, \{0,1\}, \kappa)$, such that:
  - Each voter i is mapped to a boolean feature i, such that feature i takes value 1 if voter i votes Yes; otherwise it takes value 0;
  - The classification function $\kappa : \mathbb{F} \to \{0,1\}$ is defined by:

$$\kappa(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{m} n_i x_i \geqslant q \\ 0 & \text{otherwise} \end{cases}$$

  - The target instance is $(\mathbb{1}, 1)$; and
  - Each minimal winning coalition $\mathcal{C}$ corresponds to an AXp of $\mathcal{E} = (\mathcal{M}, (\mathbb{1}, 1))$

$\therefore$ WVGs can be analyzed by studying the AXps/CXps of monotonically increasing boolean classifiers

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$
- Computing the AXps:
  - Winning coalitions must include both 1 and 2
  - We can pick 3 or, alternatively, all the other ones

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$
- Computing the AXps:
  - Winning coalitions must include both 1 and 2
  - We can pick 3 or, alternatively, all the other ones
- AXps:

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$
- Computing the AXps:
  - Winning coalitions must include both 1 and 2
  - We can pick 3 or, alternatively, all the other ones
- AXps:

$$\mathbb{A} = \{\{1, 2, 3\}, \{1, 2, 4, 5, 6, 7, 8, 9\}\}$$

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$
- Computing the AXps:
  - Winning coalitions must include both 1 and 2
  - We can pick 3 or, alternatively, all the other ones
- AXps:

$$\mathbb{A} = \{\{1,2,3\},\{1,2,4,5,6,7,8,9\}\}$$

- CXps:

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$
- Computing the AXps:
  - Winning coalitions must include both 1 and 2
  - We can pick 3 or, alternatively, all the other ones
- AXps:

$$\mathbb{A} = \{\{1, 2, 3\}, \{1, 2, 4, 5, 6, 7, 8, 9\}\}$$

- CXps:

$$\mathbb{C} = \{\{1\}, \{2\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \{3, 7\}, \{3, 8\}, \{3, 9\}, \}$$

Another WVG

- WVG: $[25; 10, 9, 7, 1, 1, 1, 1, 1, 1]$
- Computing the AXps:
  - Winning coalitions must include both 1 and 2
  - We can pick 3 or, alternatively, all the other ones
- AXps:

$$\mathbb{A} = \{\{1, 2, 3\}, \{1, 2, 4, 5, 6, 7, 8, 9\}\}$$

- CXps:

$$\mathbb{C} = \{\{1\}, \{2\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \{3, 7\}, \{3, 8\}, \{3, 9\}, \}$$

- **Q:** How should features be ranked in terms of importance?

Plan for this course – light at the end of the tunnel...

- Lecture 01 – units:
  - #01: Foundations
- Lecture 02 – units:
  - #02: Principles of symbolic XAI – feature selection
  - #03: Tractability in symbolic XAI (& myth of interpretability)
- Lecture 03 – units:
  - #04: Intractability in symbolic XAI (& myth of model-agnostic XAI)
  - #05: Explainability queries
- Lecture 04 – units:
  - #06: Advanced topics
- Lecture 05 – units:
  - #07: Principles of symbolic XAI – feature attribution (& myth of Shapley values in XAI)
  - #08: Conclusions & research directions

Unit #07

Principles of Symbolic XAI – Feature Attribution

Detour: Standard SHAP Intro (from another course...)

What are Shapley values?

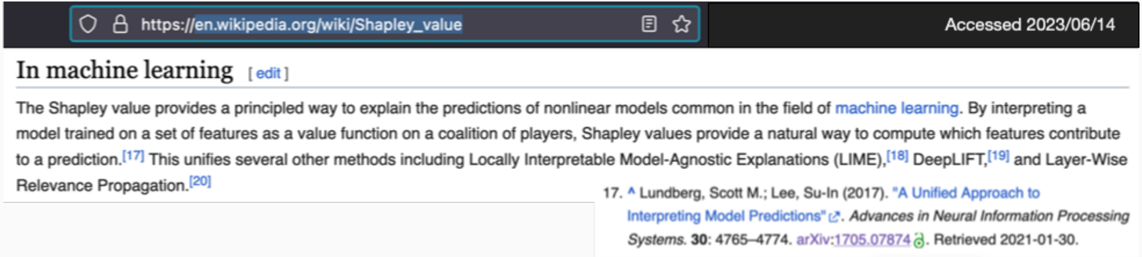
- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game

What are Shapley values?

- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game
- Application in XAI since the 2000s [LC01, SK10, SK14, DSZ16, LL17, ABBM21, VLSS21, VLSS22, ABBM23]
  - Popularized by SHAP [LL17]
  - Used for feature attribution, i.e. relative feature importance

What are Shapley values?

- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game
- Application in XAI since the 2000s [LC01, SK10, SK14, DSZ16, LL17, ABBM21, VLSS21, VLSS22, ABBM23]
  - Popularized by SHAP [LL17]
  - Used for feature attribution, i.e. relative feature importance
- Shapley values are becoming ubiquitous in XAI... – E.g. see slides from other XAI course...



https://en.wikipedia.org/wiki/Shapley_value

Accessed 2023/06/14

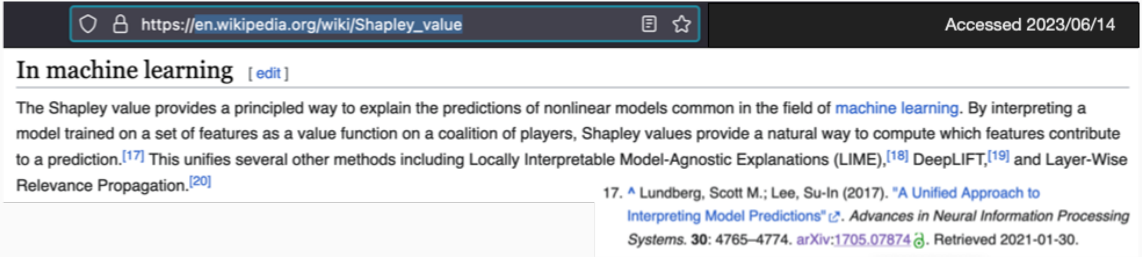
In machine learning [edit]

The Shapley value provides a principled way to explain the predictions of nonlinear models common in the field of machine learning. By interpreting a model trained on a set of features as a value function on a coalition of players, Shapley values provide a natural way to compute which features contribute to a prediction.[17] This unifies several other methods including Locally Interpretable Model-Agnostic Explanations (LIME),[18] DeepLIFT,[19] and Layer-Wise Relevance Propagation.[20]

17. ^ Lundberg, Scott M.; Lee, Su-In (2017). "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems*. **30**: 4765–4774. arXiv:1705.07874. Retrieved 2021-01-30.

What are Shapley values?

- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game
- Application in XAI since the 2000s [LC01, SK10, SK14, DSZ16, LL17, ABBM21, VLSS21, VLSS22, ABBM23]
  - Popularized by SHAP [LL17]
  - Used for feature attribution, i.e. relative feature importance
- Shapley values are becoming ubiquitous in XAI... – E.g. see slides from other XAI course...


https://en.wikipedia.org/wiki/Shapley_value

Accessed 2023/06/14

In machine learning [edit]

The Shapley value provides a principled way to explain the predictions of nonlinear models common in the field of machine learning. By interpreting a model trained on a set of features as a value function on a coalition of players, Shapley values provide a natural way to compute which features contribute to a prediction.[17] This unifies several other methods including Locally Interpretable Model-Agnostic Explanations (LIME),[18] DeepLIFT,[19] and Layer-Wise Relevance Propagation.[20]

17. ^ Lundberg, Scott M.; Lee, Su-In (2017). "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems*. **30**: 4765–4774. arXiv:1705.07874. Retrieved 2021-01-30.

- **Q:** Do Shapley values for XAI **really** provide a rigorous measure of feature importance?

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- $\Upsilon: 2^{\mathcal{F}} \rightarrow 2^{\mathbb{F}}$ defined by, [ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}} x_i = v_i\}$$

$\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- Υ: $2^{\mathcal{F}} \rightarrow 2^{\mathbb{F}}$ defined by,

[ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}} x_i = v_i\}$$

$\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

- ϕ: $2^{\mathcal{F}} \rightarrow \mathbb{R}$ defined by,

$$\phi(\mathcal{S}) = 1/2^{|\mathcal{F} \setminus \mathcal{S}|} \sum_{\mathbf{x} \in \Upsilon(\mathcal{S})} \kappa(\mathbf{x}) \;\; = \upsilon_e(\mathcal{S})$$

$\phi(\mathcal{S})$ represents the expected value of the classifier on the points given by $\Upsilon(\mathcal{S})$

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- $\Upsilon\colon 2^{\mathcal{F}} \to 2^{\mathbb{F}}$ defined by,

[ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i\in\mathcal{S}} x_i = v_i\}$$

$\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

- $\phi\colon 2^{\mathcal{F}} \to \mathbb{R}$ defined by,

$$\phi(\mathcal{S}) = {}^{1}/_{2^{|\mathcal{F}\setminus\mathcal{S}|}} \sum\nolimits_{\mathbf{x}\in\Upsilon(\mathcal{S})} \kappa(\mathbf{x}) \; = \upsilon_e(\mathcal{S})$$

$\phi(\mathcal{S})$ represents the expected value of the classifier on the points given by $\Upsilon(\mathcal{S})$

- $\text{Sc}\colon \mathcal{F} \to \mathbb{R}$ defined by,

$$\text{Sc}(i) = \sum\nolimits_{\mathcal{S}\subseteq(\mathcal{F}\setminus\{i\})} \frac{|\mathcal{S}|!(|\mathcal{F}|-|\mathcal{S}|-1)!}{|\mathcal{F}|!} \times (\phi(\mathcal{S}\cup\{i\}) - \phi(\mathcal{S}))$$

For all subsets of features, excluding i, compute the expected value of the classifier, with and without i fixed, weighted by $\frac{1}{n}\binom{n}{|\mathcal{S}|}^{-1}$

  - **Obs:** Uniform distribution assumed; it suffices for our purposes

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- $\Upsilon: 2^{\mathcal{F}} \to 2^{\mathbb{F}}$ defined by,

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}} x_i = v_i\}$$

Marginal contribution (in SHAP lingo)!

[ABBM21, ABBM23]

$\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

- $\phi: 2^{\mathcal{F}} \to \mathbb{R}$ defined by,

$$\phi(\mathcal{S}) = 1/2^{|\mathcal{F} \setminus \mathcal{S}|} \sum_{\mathbf{x} \in \Upsilon(\mathcal{S})} \kappa(\mathbf{x}) = \upsilon_e(\mathcal{S})$$

$\phi(\mathcal{S})$ represents the expected value of the classifier on the points given by $\Upsilon(\mathcal{S})$

- Sc: $\mathcal{F} \to \mathbb{R}$ defined by,

$$\mathsf{Sc}(i) = \sum_{\mathcal{S} \subseteq (\mathcal{F} \setminus \{i\})} \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!} \times (\phi(\mathcal{S} \cup \{i\}) - \phi(\mathcal{S}))$$

For all subsets of features, excluding i, compute the expected value of the classifier, with and without i fixed, weighted by $\frac{1}{n}\binom{n}{|\mathcal{S}|}^{-1}$

- **Obs:** Uniform distribution assumed; it suffices for our purposes

How are Shapley values computed in practice?

- Exact evaluation is computationally (very) hard [VLSS21, ABBM21, VLSS22, ABBM23, HMS24]

- SHAP proposes a sample-based approach; with no guarantees of rigor [LL17]
  - Recent experiments revealed little to no correlation between Shapley values and SHAP's results [HM23a]

How are Shapley values computed in practice?

- Exact evaluation is computationally (very) hard [VLSS21, ABBM21, VLSS22, ABBM23, HMS24]
- SHAP proposes a sample-based approach; with no guarantees of rigor [LL17]
  - Recent experiments revealed little to no correlation between Shapley values and SHAP's results [HM23a]
- Polynomial-time algorithm for deterministic decomposable boolean circuits [ABBM21]
- Polynomial-time algorithm for boolean functions represented with a truth-table [HM23a]

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****."*
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****."*
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

- And [SK10] also reads:
  "When viewed together, these properties ensure that ***any effect the features might have on the classifiers output will be reflected in the generated contributions****, which effectively deals with the issues of previous general explanation methods."*

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****."*
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

- And [SK10] also reads:
  "When viewed together, these properties ensure that ***any effect the features might have on the classifiers output will be reflected in the generated contributions****, which effectively deals with the issues of previous general explanation methods."*

- **Obs:** Shapley values are defined axiomatically, i.e. **no** immediate relationship with AXp's/CXp's or with feature (ir)relevancy

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****."*
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

- And [SK10] also reads:
  "When viewed together, these properties ensure that ***any effect the features might have on the classifiers output will be reflected in the generated contributions****, which effectively deals with the issues of previous general explanation methods."*

- **Obs:** Shapley values are defined axiomatically, i.e. **no** immediate relationship with AXp's/CXp's or with feature (ir)relevancy
  - Qs: can we have **irrelevant** features with a non-zero Shapley value, and/or **relevant** features with a Shapley of zero?
    - Recall: **relevant** features occur in **some** AXp/CXp; **irrelevant** features do **not** occur in **any** AXp/CXp

Outline – Unit #07

Shapley values vs. feature (ir)relevancy – identified issues [HM23a, HM23b, HM23c, MH23, HMS24, MSH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:

Shapley values vs. feature (ir)relevancy – identified issues [HM23a, HM23b, HM23c, MH23, HMS24, MSH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,

$$\mathsf{Irrelevant}(i) \wedge (\mathsf{Sv}(i) \neq 0)$$

Shapley values vs. feature (ir)relevancy – identified issues [HM23a, HM23b, HM23c, MH23, HMS24, MSH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,

$$\text{Irrelevant}(i) \wedge (\text{Sv}(i) \neq 0)$$

  - Issue I2 occurs if,

$$\text{Irrelevant}(i_1) \wedge \text{Relevant}(i_2) \wedge (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$

Shapley values vs. feature (ir)relevancy – identified issues [HM23a, HM23b, HM23c, MH23, HMS24, MSH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,

$$\mathsf{Irrelevant}(i) \land (\mathsf{Sv}(i) \neq 0)$$

  - Issue I2 occurs if,

$$\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (|\mathsf{Sv}(i_1)| > |\mathsf{Sv}(i_2)|)$$

  - Issue I3 occurs if,

$$\mathsf{Relevant}(i) \land (\mathsf{Sv}(i) = 0)$$

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,
$$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$
  - Issue I2 occurs if,
$$\text{Irrelevant}(i_1) \land \text{Relevant}(i_2) \land (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$
  - Issue I3 occurs if,
$$\text{Relevant}(i) \land (\text{Sv}(i) = 0)$$
  - Issue I4 occurs if,
$$[\text{Irrelevant}(i_1) \land (\text{Sv}(i_1) \neq 0)] \land [\text{Relevant}(i_2) \land (\text{Sv}(i_2) = 0)]$$

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,
    $$\mathsf{Irrelevant}(i) \land (\mathsf{Sv}(i) \neq 0)$$
  - Issue I2 occurs if,
    $$\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (|\mathsf{Sv}(i_1)| > |\mathsf{Sv}(i_2)|)$$
  - Issue I3 occurs if,
    $$\mathsf{Relevant}(i) \land (\mathsf{Sv}(i) = 0)$$
  - Issue I4 occurs if,
    $$[\mathsf{Irrelevant}(i_1) \land (\mathsf{Sv}(i_1) \neq 0)] \land [\mathsf{Relevant}(i_2) \land (\mathsf{Sv}(i_2) = 0)]$$
  - Issue I5 occurs if,
    $$[\mathsf{Irrelevant}(i) \land \forall_{1 \leqslant j \leqslant m, j \neq i} (|\mathsf{Sv}(j)| < |\mathsf{Sv}(i)|)]$$

Shapley values vs. feature (ir)relevancy – identified issues [HM23a, HM23b, HM23c, MH23, HMS24, MSH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,

$$\mathsf{Irrelevant}(i) \land (\mathsf{Sv}(i) \neq 0)$$

  - Issue I2 occurs if,

$$\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (|\mathsf{Sv}(i_1)| > |\mathsf{Sv}(i_2)|)$$

  - Issue I3 occurs if,

$$\mathsf{Relevant}(i) \land (\mathsf{Sv}(i) = 0)$$

Any of these issues is a cause of (**serious**) concern per se**!**

  - Issue I4 occurs if,

$$[\mathsf{Irrelevant}(i_1) \land (\mathsf{Sv}(i_1) \neq 0)] \land [\mathsf{Relevant}(i_2) \land (\mathsf{Sv}(i_2) = 0)]$$

  - Issue I5 occurs if,

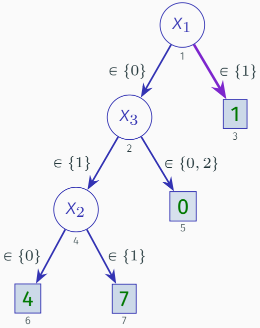
$$[\mathsf{Irrelevant}(i) \land \forall_{1 \leqslant j \leqslant m, j \neq i} (|\mathsf{Sv}(j)| < |\mathsf{Sv}(i)|)]$$

Some stats – all boolean functions with 4 variables

[HM23a, HM23b, HM23c, MH23, HMS24, MSH24]

Issue-related metric	Value	Recap issue
# of functions	65536	
# number of instances	1048576	
# of I1 issues	781696	
# of functions with I1 issues	65320	
% I1 issues / function	99.67	$[\text{Irrelevant}(i) \wedge (\text{Sv}(i) \neq 0)]$
# of I2 issues	105184	
# of functions with I2 issues	40448	
% I2 issues / function	61.72	$[\text{Irrelevant}(i_1) \wedge \text{Relevant}(i_2) \wedge (\lvert\text{Sv}(i_1)\rvert > \lvert\text{Sv}(i_2)\rvert)]$
# of I3 issues	43008	
# of functions with I3 issues	7800	
% I3 issues / function	11.90	$[\text{Relevant}(i) \wedge (\text{Sv}(i) = 0)]$
# of I4 issues	5728	
# of functions with I4 issues	2592	
% I4 issues / function	3.96	$[\text{Irrelevant}(i_1) \wedge (\text{Sv}(i_1) \neq 0)] \wedge [\text{Relevant}(i_2) \wedge (\text{Sv}(i_2) = 0)]$
# of I5 issues	1664	
# of functions with I5 issues	1248	
% I5 issues / function	1.90	$[\text{Irrelevant}(i) \wedge \forall_{1 \leqslant j \leqslant m, j \neq i} (\lvert\text{Sv}(j)\rvert < \lvert\text{Sv}(i)\rvert)]$

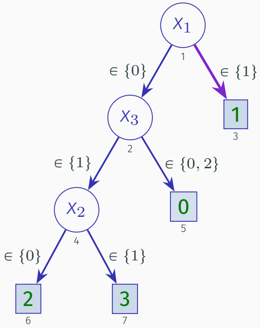
Previous results do matter! Let's go non-boolean...



DT1

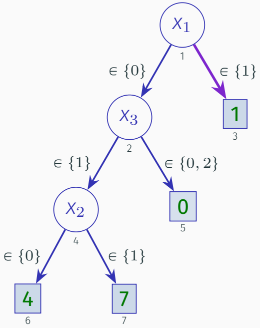
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

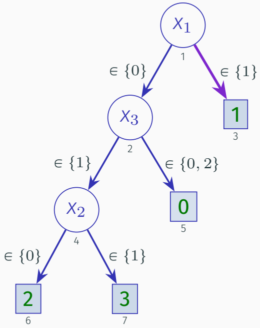
Instance $((1,1,2),1)$ – which feature matters the most for prediction 1?



DT1

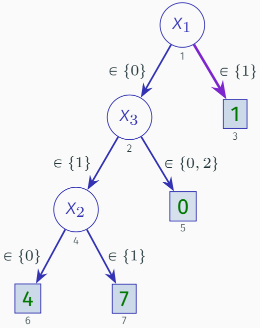
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

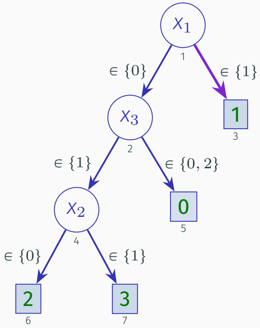
Computing XPs – make sense...



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

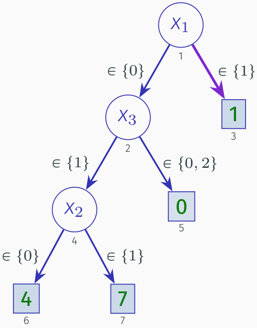
Tabular representations



DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

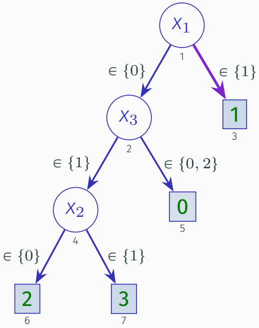
Computing XPs, AEs – also make sense...



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations

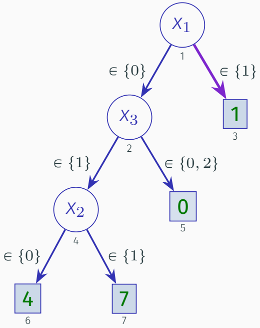


DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

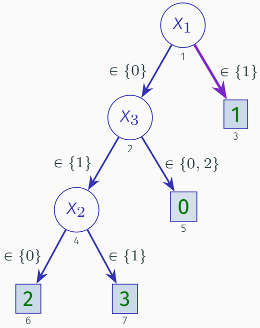
Computing XPs, AEs & Svs



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



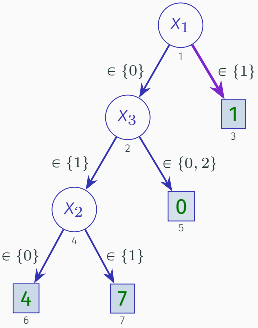
DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values			
DT	Sc(1)	Sc(2)	Sc(3)
DT1	0.000	0.083	-0.500
DT2	0.278	0.028	-0.222

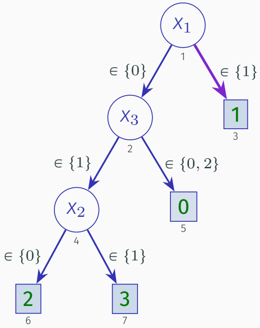
Computing XPs, AEs & Svs – what???



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations

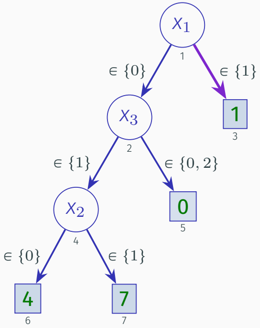


DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

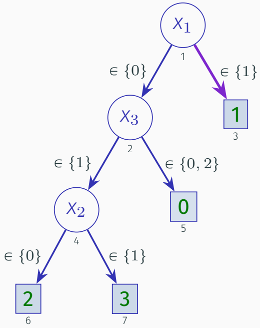
Shapley values				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

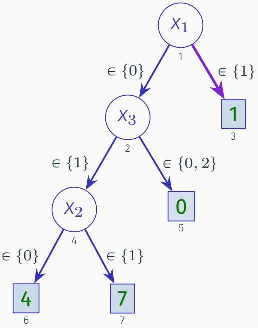
XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

Computing XPs, AEs & Svs – what???

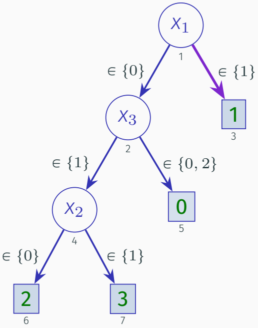
∴ Shapley values can mislead human decision-makers !



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

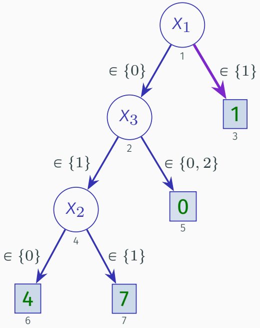
XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

Computing XPs, AEs & Svs – what???

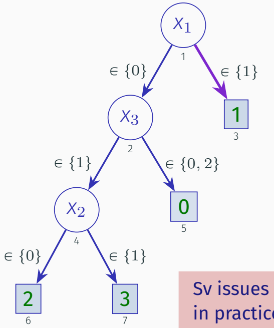
∴ Shapley values can mislead human decision-makers !



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

Sv issues also occur in practice [HM23c]

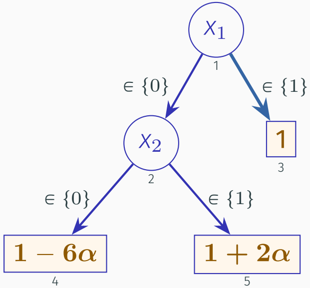
XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

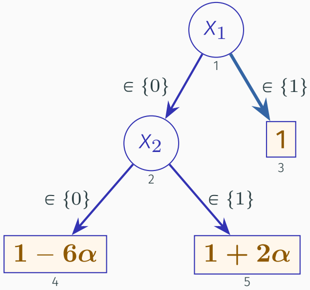
Another example – arbitrary mistakes!

[LHAMS24]



Another example – arbitrary mistakes!

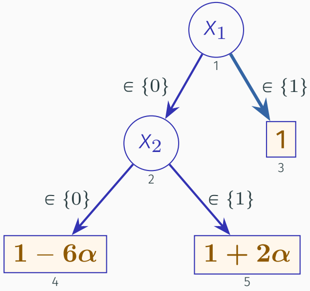
[LHAMS24]



- Instance: $((1, 1), 1)$
- **Obs:** $\alpha \neq 1$

Another example – arbitrary mistakes!

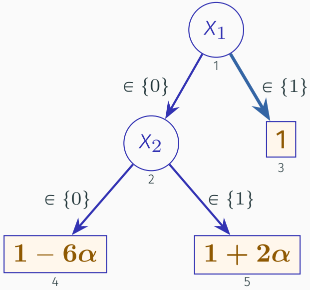
[LHAMS24]



- Instance: $((1, 1), 1)$
- **Obs:** $\alpha \mp 1$
- Sc$(1) = 0$
- Sc$(2) = \alpha$

Another example – arbitrary mistakes!

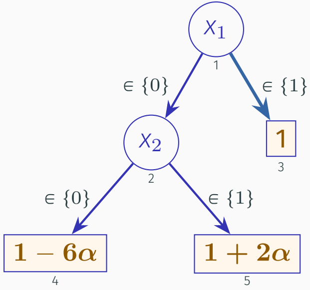
[LHAMS24]



- Instance: $((1, 1), 1)$
- **Obs:** $\alpha \neq 1$
- $\mathsf{Sc}(1) = 0$
- $\mathsf{Sc}(2) = \alpha$ (you can pick the α...)

Another example – arbitrary mistakes!

[LHAMS24]

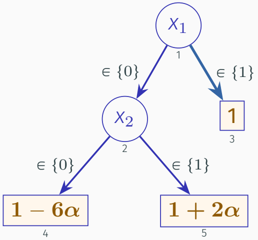


- Instance: $((1,1),1)$
- **Obs:** $\alpha \ddagger 1$
- Sc$(1) = 0$
- Sc$(2) = \alpha$ (you can pick the α...)

Example devised by O. Letoffe, PhD student at IRIT

More detail

row	x_1	x_2	$\rho(\mathbf{x})$	$\rho_a(\mathbf{x})$ $\alpha = 1/2$	$\rho_b(\mathbf{x})$ $\alpha = 1/4$
1	0	0	$1 - 6\alpha$	-2	$-1/2$
2	0	1	$1 + 2\alpha$	2	$3/2$
3	1	0	1	1	1
4	1	1	1	1	1



$\mathcal{S}$	rows($\mathcal{S}$)	$\upsilon_e(\mathcal{S})$
$\emptyset$	1, 2, 3, 4	$1 - \alpha$
$\{x_1\}$	3, 4	1
$\{x_2\}$	2, 4	$1 + \alpha$
$\{x_1, x_2\}$	4	1

$i = 1$					
$\mathcal{S}$	$\upsilon_e(\mathcal{S})$	$\upsilon_e(\mathcal{S} \cup \{1\})$	$\Delta_1(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S}) \times \Delta_1(\mathcal{S})$
$\emptyset$	$1 - \alpha$	1	α	$1/2$	$\alpha/2$
$\{2\}$	$1 + \alpha$	1	$-\alpha$	$1/2$	$-\alpha/2$
			$\mathrm{Sc}_E(1)$	$=$	0

$i = 2$					
$\mathcal{S}$	$\upsilon_e(\mathcal{S})$	$\upsilon_e(\mathcal{S} \cup \{2\})$	$\Delta_2(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S}) \times \Delta_2(\mathcal{S})$
$\emptyset$	$1 - \alpha$	$1 + \alpha$	2α	$1/2$	α
$\{1\}$	1	1	0	$1/2$	0
			$\mathrm{Sc}_E(2)$	$=$	α

Outline – Unit #07

Corrected SHAP scores & feature importance scores

[LHMS24, LHAMS24]

- Is the theory of Shapley values incorrect?

Corrected SHAP scores & feature importance scores

[LHMS24, LHAMS24]

- Is the theory of Shapley values incorrect? **No!**

Corrected SHAP scores & feature importance scores

[LHMS24, LHAMS24]

- Is the theory of Shapley values incorrect? **No!**
- What is inadequate is the **characteristic function** used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...

Corrected SHAP scores & feature importance scores

[LHMS24, LHAMS24]

- Is the theory of Shapley values incorrect? No!
- What is inadequate is the **characteristic function** used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...
- Replace characteristic function based on expected values by new characteristic function based on AXps/WAXps [LHMS24]
  - Resulting scores are (still) Shapley values & identified issues no longer observed

Corrected SHAP scores & feature importance scores

[LHMS24, LHAMS24]

- Is the theory of Shapley values incorrect? **No!**
- What is inadequate is the **characteristic function** used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...
- Replace characteristic function based on expected values by new characteristic function based on **AXps/WAXps** [LHMS24]
  - Resulting scores are (still) Shapley values & identified issues no longer observed
- Observed tight connection between feature attribution and power indices from a priori voting power

Corrected SHAP scores & feature importance scores

[LHMS24, LHAMS24]

- Is the theory of Shapley values incorrect? No!
- What is inadequate is the **characteristic function** used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...
- Replace characteristic function based on expected values by new characteristic function based on AXps/WAXps [LHMS24]
  - Resulting scores are (still) Shapley values & identified issues no longer observed
- Observed tight connection between feature attribution and power indices from a priori voting power
  - Feature importance scores: [LHAMS24]
    - Generalize recent axiomatic aggregations [BIL+24]
    - Adapt best known power indices
    - Devise new scores for XAI

An initial compromise

[LHAMS24]

- Replace the characteristic function used for SHAP scores:

$$\upsilon_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

An initial compromise

[LHAMS24]

- Replace the characteristic function used for SHAP scores:

$$v_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x})\,|\,\mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

An initial compromise

[LHAMS24]

- Replace the characteristic function used for SHAP scores:

$$v_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

- The new characteristic function becomes:

$$v_s(\mathcal{S}) \quad := \quad \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

An initial compromise

[LHAMS24]

- Replace the characteristic function used for SHAP scores:

$$\upsilon_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

- The new characteristic function becomes:

$$\upsilon_s(\mathcal{S}) \quad := \quad \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Issues with non-boolean classifiers disappear; issues with boolean classifiers remain

An initial compromise

[LHAMS24]

- Replace the characteristic function used for SHAP scores:

$$v_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

- The new characteristic function becomes:

$$v_s(\mathcal{S}) \quad := \quad \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Issues with non-boolean classifiers disappear; issues with boolean classifiers remain
- Developed SSHAP prototype using SHAP's code base [LHMS24]

Fixing the known issues of SHAP scores

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$v_a(\mathcal{S}) \quad := \quad \left\{ \begin{array}{ll} 1 & \quad \text{if } \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1 \\ 0 & \quad \text{otherwise} \end{array} \right.$$

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$v_a(\mathcal{S}) \quad := \quad \begin{cases} 1 & \text{if } \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

  - **Recall:** $\mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1$ holds iff $\mathcal{S}$ is a WAXp

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$v_a(\mathcal{S}) \quad := \quad \begin{cases} 1 & \text{if } \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

  - **Recall:** $\mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1$ holds iff $\mathcal{S}$ is a WAXp

- Known issues of SHAP scores guaranteed not to occur

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$v_a(\mathcal{S}) \quad := \quad \begin{cases} 1 & \text{if } \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

  - **Recall:** $\mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1$ holds iff $\mathcal{S}$ is a WAXp

- Known issues of SHAP scores guaranteed not to occur

- Corrected SHAP scores reveal tight connection between XAI by feature selection (i.e. WAXps) and feature attribution

Outline – Unit #07

Exact Shapley Values for XAI

Myth #03: Shapley Values for XAI

Corrected SHAP Scores

Voting Power & Power Indices

Feature Importance Scores

Recap: weighted voting games

- General set up of weighted voting games:
  - Assembly $\mathcal{A}$ of voters, with $m = |\mathcal{A}|$
  - Each voter $i \in \mathcal{A}$ votes Yes with n_i votes; otherwise no votes are counte (and he/she votes No)
  - A coalition is a subset of voters, $\mathcal{C} \subseteq \mathcal{A}$
  - Quota q is the sum of votes required for a proposal to be approved
    - Coalitions leading to sums not less than q are **winning** coalitions
  - A weighted voting game (WVG) is a tuple $[q; n_1, \ldots, n_m]$
    - Example: $[12; 4, 4, 4, 2, 2, 1]$
  - Problem: find a measure of importance of each voter !
    - I.e. measure the a priori voting power of each voter

What are power indices?

- Power indices assign a measure of importance to each voter

What are power indices?

- Power indices assign a measure of importance to each voter
- Many power indices proposed over the years:
  - Penrose [Pen46]
  - Shapley-Shubik [SS54]
  - Banzhaf [BI65]
  - Coleman [Col71]
  - Johnston [Joh78]
  - Deegan-Packel [DP78]
  - Holler-Packel [HP83]
  - Andjiga [ACL03]
  - Responsability* [CH04, BIL+24]
  - ...

What are power indices?

- Power indices assign a measure of importance to each voter
- Many power indices proposed over the years:
  - Penrose [Pen46]
  - Shapley-Shubik [SS54]
  - Banzhaf [BI65]
  - Coleman [Col71]
  - Johnston [Joh78]
  - Deegan-Packel [DP78]
  - Holler-Packel [HP83]
  - Andjiga [ACL03]
  - Responsability* [CH04, BIL+24]
  - ...
- What characterizes power indices?
  - Account for the cases when voter is *critical* for a winning coalition
    - E.g. in previous example, Luxembourg is never critical for a winning coalition
  - Account for whether coalition is subset-minimal or cardinality-minimal

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]
  - Since the work of Shapley-Shubik [SS54], the criticality of a voter has been accounted for: *"Our definition of the power of an individual member depends on the chance he has of being critical to the success of a winning coalition."*

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]
  - Since the work of Shapley-Shubik [SS54], the criticality of a voter has been accounted for: *"Our definition of the power of an individual member depends on the chance he has of being critical to the success of a winning coalition."*
  - This means that a voter i is critical when:
    - If the voter votes Yes, then we have a winning coalition; and
    - If the voter votes No, then we have a losing coalition.

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]
  - Since the work of Shapley-Shubik [SS54], the criticality of a voter has been accounted for: *"Our definition of the power of an individual member depends on the chance he has of being critical to the success of a winning coalition."*
  - This means that a voter i is critical when:
    - If the voter votes Yes, then we have a winning coalition; and
    - If the voter votes No, then we have a losing coalition.

- Understanding (subset-)minimal winning coalitions:

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]
  - Since the work of Shapley-Shubik [SS54], the criticality of a voter has been accounted for: *"Our definition of the power of an individual member depends on the chance he has of being critical to the success of a winning coalition."*
  - This means that a voter i is critical when:
    - If the voter votes Yes, then we have a winning coalition; and
    - If the voter votes No, then we have a losing coalition.

- Understanding (subset-)minimal winning coalitions:
  - A winning coalition is subset-minimal if removing any single voter results in a losing coalition

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]
  - Since the work of Shapley-Shubik [SS54], the criticality of a voter has been accounted for: *"Our definition of the power of an individual member depends on the chance he has of being critical to the success of a winning coalition."*
  - This means that a voter i is critical when:
    - If the voter votes Yes, then we have a winning coalition; and
    - If the voter votes No, then we have a losing coalition.

- Understanding (subset-)minimal winning coalitions:
  - A winning coalition is subset-minimal if removing any single voter results in a losing coalition
  - A winning coalition is cardinality-minimal if it has the smallest cardinality among subset-minimal winning coalitions

Towards defining power indices

- Understanding criticality (used at least since 1954): [SS54]
  - Since the work of Shapley-Shubik [SS54], the criticality of a voter has been accounted for: *"Our definition of the power of an individual member depends on the chance he has of being critical to the success of a winning coalition."*
  - This means that a voter i is critical when:
    - If the voter votes Yes, then we have a winning coalition; and
    - If the voter votes No, then we have a losing coalition.

- Understanding (subset-)minimal winning coalitions:
  - A winning coalition is subset-minimal if removing any single voter results in a losing coalition
  - A winning coalition is cardinality-minimal if it has the smallest cardinality among subset-minimal winning coalitions
  - Recall that minimal winning coalitions can be obtained by computing the AXps of a monotonically increasing boolean classifier

Example power indices I

[LHAMS24]

- Necessary definitions (using formal XAI notation...):

$$\mathbb{WA}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{WAXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$
$$\mathbb{WC}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{WCXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$
$$\mathbb{A}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$
$$\mathbb{C}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$

  - Definitions of $\mathbb{WA}$, $\mathbb{WC}$, $\mathbb{A}$, and $\mathbb{C}$ mimic the ones above, but without specifying a voter

Example power indices I

[LHAMS24]

- Necessary definitions (using formal XAI notation...):

$$\mathbb{WA}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{WAXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$
$$\mathbb{WC}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{WCXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$
$$\mathbb{A}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$
$$\mathbb{C}_i(\mathcal{E}) = \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}$$

  - Definitions of $\mathbb{WA}$, $\mathbb{WC}$, $\mathbb{A}$, and $\mathbb{C}$ mimic the ones above, but without specifying a voter

- Power indices of Holler-Packel and Deegan-Packel:

[HP83, DP78]

$$\mathsf{Sc}_H(i;\mathcal{E}) = \sum\nolimits_{\mathcal{S}\in\mathbb{A}_i(\mathcal{E})} \left(1/|\mathbb{A}(\mathcal{E})|\right)$$
$$\mathsf{Sc}_D(i;\mathcal{E}) = \sum\nolimits_{\mathcal{S}\in\mathbb{A}_i(\mathcal{E})} \left(1/(|\mathcal{S}| \times |\mathbb{A}(\mathcal{E})|)\right)$$

Example power indices I

[LHAMS24]

- Necessary definitions (using formal XAI notation...):

$$\begin{aligned}
\mathbb{WA}_i(\mathcal{E}) &= \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{WAXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\} \\
\mathbb{WC}_i(\mathcal{E}) &= \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{WCXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\} \\
\mathbb{A}_i(\mathcal{E}) &= \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\} \\
\mathbb{C}_i(\mathcal{E}) &= \{\mathcal{S} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{S};\mathcal{E}) \land i \in \mathcal{S}\}
\end{aligned}$$

  - Definitions of $\mathbb{WA}$, $\mathbb{WC}$, $\mathbb{A}$, and $\mathbb{C}$ mimic the ones above, but without specifying a voter

- Power indices of Holler-Packel and Deegan-Packel:

[HP83, DP78]

$$\begin{aligned}
\mathsf{Sc}_H(i;\mathcal{E}) &= \textstyle\sum_{\mathcal{S}\in\mathbb{A}_i(\mathcal{E})} \left(1/|\mathbb{A}(\mathcal{E})|\right) \\
\mathsf{Sc}_D(i;\mathcal{E}) &= \textstyle\sum_{\mathcal{S}\in\mathbb{A}_i(\mathcal{E})} \left(1/(|\mathcal{S}| \times |\mathbb{A}(\mathcal{E})|)\right)
\end{aligned}$$

  - **Obs:** One *only* needs the AXps

Example power indices II

- Additional definitions:

$$\text{Crit}(i, \mathcal{S}; \mathcal{E}) \quad := \quad \text{WAXp}(\mathcal{S}; \mathcal{E}) \land \neg\text{WAXp}(\mathcal{S} \backslash \{i\}; \mathcal{E})$$

Example power indices II

- Additional definitions:

$$\text{Crit}(i, \mathcal{S}; \mathcal{E}) \quad := \quad \text{WAXp}(\mathcal{S}; \mathcal{E}) \wedge \neg\text{WAXp}(\mathcal{S}\backslash\{i\}; \mathcal{E})$$

- Power indices of Shapley-Shubik, Banzhaf and Johnston: [SS54, BI65, Joh78]

$$\begin{aligned}
\text{Sc}_S(i; \mathcal{E}) &= \sum\nolimits_{\mathcal{S}\subseteq\mathcal{F} \wedge \text{Crit}(i,\mathcal{S};\mathcal{E})} \left({}^{1}\!/\!\left(|\mathcal{F}| \times \binom{|\mathcal{F}|-1}{|\mathcal{S}|-1} \right) \right) \\
\text{Sc}_B(i; \mathcal{E}) &= \sum\nolimits_{\mathcal{S}\subseteq\mathcal{F} \wedge \text{Crit}(i,\mathcal{S};\mathcal{E})} \left({}^{1}\!/2^{|\mathcal{F}|-1} \right) \\
\text{Sc}_J(i; \mathcal{E}) &= \sum\nolimits_{\mathcal{S}\subseteq\mathcal{F} \wedge \text{Crit}(i,\mathcal{S};\mathcal{E})} \left({}^{1}\!/\Delta(\mathcal{S}) \right)
\end{aligned}$$

Example power indices II

- Additional definitions:

$$\mathsf{Crit}(i,\mathcal{S};\mathcal{E}) \quad := \quad \mathsf{WAXp}(\mathcal{S};\mathcal{E}) \land \neg\mathsf{WAXp}(\mathcal{S}\backslash\{i\};\mathcal{E})$$

- Power indices of Shapley-Shubik, Banzhaf and Johnston: [SS54, BI65, Joh78]

$$\begin{aligned}
\mathsf{Sc}_S(i;\mathcal{E}) &= \sum\nolimits_{\mathcal{S}\subseteq\mathcal{F}\land\mathsf{Crit}(i,\mathcal{S};\mathcal{E})} \left({}^1/\left(|\mathcal{F}|\times\binom{|\mathcal{F}|-1}{|\mathcal{S}|-1}\right)\right)\\
\mathsf{Sc}_B(i;\mathcal{E}) &= \sum\nolimits_{\mathcal{S}\subseteq\mathcal{F}\land\mathsf{Crit}(i,\mathcal{S};\mathcal{E})} \left({}^1/2^{|\mathcal{F}|-1}\right)\\
\mathsf{Sc}_J(i;\mathcal{E}) &= \sum\nolimits_{\mathcal{S}\subseteq\mathcal{F}\land\mathsf{Crit}(i,\mathcal{S};\mathcal{E})} \left({}^1/\Delta(\mathcal{S})\right)
\end{aligned}$$

  - One needs the WAXps to find critical voters...

Example #01

- WVG: $[9; 9, 2, 2, 2, 2, 1, 1]$

Example #01

- WVG: $[9; 9, 2, 2, 2, 2, 1, 1]$
- AXps:

$$\begin{array}{lllll} 1 & & & & \\ 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 7 \end{array}$$

Example #01

- WVG: $[9; 9, 2, 2, 2, 2, 1, 1]$
- AXps:

$$\begin{array}{lllll} 1 & & & & \\ 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 7 \end{array}$$

- Holler-Packel scores: $\langle 0.333, 0.667, 0.667, 0.667, 0.667, 0.333, 0.333 \rangle$
- Banzhaf scores (normalized): $\langle 0.813, 0.040, 0.040, 0.040, 0.040, 0.013, 0.013 \rangle$
- Shapley-Shubik scores: $\langle 0.810, 0.043, 0.043, 0.043, 0.043, 0.010, 0.010 \rangle$
- Different relative orders of voter importance... which ones seem more realistic?

Example #02

- WVG: $[16; 10, 6, 4, 2, 2]$

Example #02

- WVG: $[16; 10, 6, 4, 2, 2]$

- AXps:

$$\begin{array}{ccc} 1 & 2 & \\ 1 & 3 & 4 \\ 1 & 3 & 5 \end{array}$$

Example #02

- WVG: $[16; 10, 6, 4, 2, 2]$
- AXps:

$$\begin{array}{ccc} 1 & 2 & \\ 1 & 3 & 4 \\ 1 & 3 & 5 \end{array}$$

- Deegan-Packel scores: $\langle 0.389, 0.167, 0.222, 0.111, 0.111 \rangle$
- Banzhaf scores (normalized): $\langle 0.524, 0.238, 0.143, 0.048, 0.048 \rangle$
- Shapley-Shubik scores: $\langle 0.617, 0.200, 0.117, 0.033, 0.033 \rangle$
- Different relative orders of voter importance... which ones seem more realistic?

Example #03

- WVG: $[6; 4, 2, 1, 1, 1, 1]$

Example #03

- WVG: $[6; 4, 2, 1, 1, 1, 1]$

- AXps:

$$\begin{array}{ccccc} 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 4 & & \\ 1 & 4 & 5 & & \\ 1 & 4 & 6 & & \\ 1 & 3 & 6 & & \\ 1 & 5 & 6 & & \\ 1 & 2 & & & \\ 1 & 3 & 5 & & \end{array}$$

Example #03

- WVG: $[6; 4, 2, 1, 1, 1, 1]$
- AXps:

$$\begin{array}{ccccc} 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 4 \\ 1 & 4 & 5 \\ 1 & 4 & 6 \\ 1 & 3 & 6 \\ 1 & 5 & 6 \\ 1 & 2 \\ 1 & 3 & 5 \end{array}$$

- Deegan-Packel scores: $\langle 0.312, 0.087, 0.150, 0.150, 0.150, 0.150 \rangle$
- Banzhaf scores (normalized): $\langle 0.542, 0.125, 0.083, 0.083, 0.083, 0.083 \rangle$
- Shapley-Shubik scores: $\langle 0.533, 0.133, 0.083, 0.083, 0.083, 0.083 \rangle$
- Different relative orders of voter importance… which ones seem more realistic?

Example #04

- WVG: $[21; 12, 9, 4, 4, 1, 1, 1]$

Example #04

- WVG: $[21; 12, 9, 4, 4, 1, 1, 1]$
- AXps:

$$\begin{array}{cccc} 1 & 2 & & \\ 1 & 3 & 4 & 5 \\ 1 & 3 & 4 & 6 \\ 1 & 3 & 4 & 7 \end{array}$$

Example #04

- WVG: $[21; 12, 9, 4, 4, 1, 1, 1]$
- AXps:

$$\begin{array}{cccc} 1 & 2 & & \\ 1 & 3 & 4 & 5 \\ 1 & 3 & 4 & 6 \\ 1 & 3 & 4 & 7 \end{array}$$

- Deegan-Packel scores: $\langle 0.312, 0.125, 0.188, 0.188, 0.062, 0.062, 0.062 \rangle$
- Banzhaf scores (normalized): $\langle 0.481, 0.309, 0.086, 0.086, 0.012, 0.012, 0.012 \rangle$
- Shapley-Shubik scores: $\langle 0.574, 0.257, 0.074, 0.074, 0.007, 0.007, 0.007 \rangle$
- Different relative orders of voter importance... which ones seem more realistic?

Outline – Unit #07

From power indices to feature importance scores

- A Feature Importance Score (FIS) is a measure of feature importance in XAI, parameterizable on an explanation problem and a chosen characteristic function
  - Explanation problem: $(\mathcal{M}, (\mathbf{v}, q))$
  - Define characteristic function using explanation problem (more next slide)

- Obs: Can adapt (generalized) power indices as templates for feature importance scores

- Obs: Can devise new templates and/or new FISs

Some examples (1 of 2)

- More notation:

$$\Delta_i(\mathcal{S};\mathcal{E},\upsilon) = \upsilon(\mathcal{S};\mathcal{E}) - \upsilon(\mathcal{S}\backslash\{i\};\mathcal{E})$$

  - Can use **any** characteristic function, including those presented earlier in this lecture

Some examples (1 of 2)

- More notation:

$$\Delta_i(\mathcal{S};\mathcal{E},\upsilon) = \upsilon(\mathcal{S};\mathcal{E}) - \upsilon(\mathcal{S}\setminus\{i\};\mathcal{E})$$

  - Can use **any** characteristic function, including those presented earlier in this lecture

- Some templates:
  - Shapley-Shubik:

$$\mathsf{TSc}_S(i;\mathcal{E},\upsilon) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},\upsilon)}{|\mathcal{F}| \times \binom{|\mathcal{F}|-1}{|\mathcal{S}|-1}}\right)$$

  - Banzhaf:

$$\mathsf{TSc}_B(i;\mathcal{E},\upsilon) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},\upsilon)}{2^{|\mathcal{F}|-1}}\right)$$

Some examples (1 of 2)

- More notation:

$$\Delta_i(\mathcal{S};\mathcal{E},\upsilon) = \upsilon(\mathcal{S};\mathcal{E}) - \upsilon(\mathcal{S}\setminus\{i\};\mathcal{E})$$

  - Can use **any** characteristic function, including those presented earlier in this lecture

- Some templates:
  - Shapley-Shubik:

$$\mathsf{TSc}_S(i;\mathcal{E},\upsilon) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},\upsilon)}{|\mathcal{F}| \times \binom{|\mathcal{F}|-1}{|\mathcal{S}|-1}}\right)$$

  - Banzhaf:

$$\mathsf{TSc}_B(i;\mathcal{E},\upsilon) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},\upsilon)}{2^{|\mathcal{F}|-1}}\right)$$

- Can use other templates

Some examples (1 of 2)

- More notation:

$$\Delta_i(\mathcal{S};\mathcal{E},\upsilon) = \upsilon(\mathcal{S};\mathcal{E}) - \upsilon(\mathcal{S}\setminus\{i\};\mathcal{E})$$

  - Can use **any** characteristic function, including those presented earlier in this lecture

- Some templates:
  - Shapley-Shubik:

$$\mathsf{TSc}_S(i;\mathcal{E},\upsilon) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},\upsilon)}{|\mathcal{F}|\times\binom{|\mathcal{F}|-1}{|\mathcal{S}|-1}}\right)$$

  - Banzhaf:

$$\mathsf{TSc}_B(i;\mathcal{E},\upsilon) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},\upsilon)}{2^{|\mathcal{F}|-1}}\right)$$

- Can use other templates

- Can devise FISs without exploiting existing templates

Some examples (2 of 2)

- Recall WAXp based characteristic function:

$$v_a(\mathcal{S}) \quad := \quad \begin{cases} 1 & \text{if } \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

Some examples (2 of 2)

- Recall WAXp based characteristic function:

$$v_a(\mathcal{S}) \;:=\; \begin{cases} 1 & \text{if } \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

- Some FISs:
  - Shapley-Shubik:

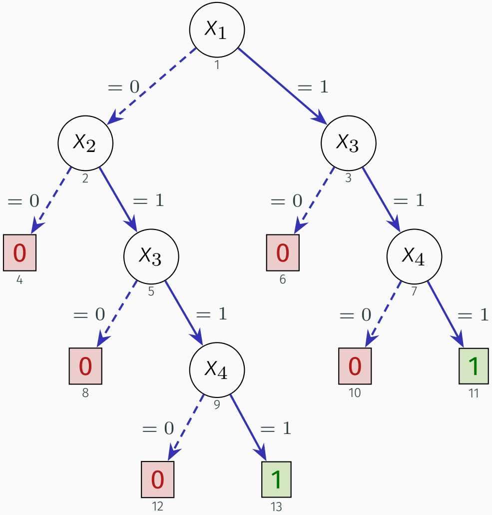
$$\mathsf{Sc}_S(i;\mathcal{E}) := \mathsf{TSc}_S(i;\mathcal{E},v_a) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},v_a)}{|\mathcal{F}| \times \binom{|\mathcal{F}|-1}{|\mathcal{S}|-1}}\right)$$

  - Banzhaf:

$$\mathsf{Sc}_B(i;\mathcal{E}) := \mathsf{TSc}_B(i;\mathcal{E},v_a) := \sum_{\mathcal{S}\in\{\mathcal{T}\subseteq\mathcal{F}\,|\,i\in\mathcal{T}\}} \left(\frac{\Delta_i(\mathcal{S};\mathcal{E},v_a)}{2^{|\mathcal{F}|-1}}\right)$$

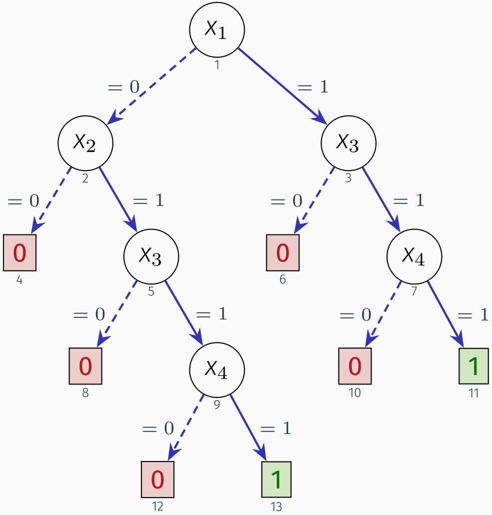
A concrete example

- AXps: $\{\{1,3,4\},\{2,3,4\}\}$
- Feature attribution:



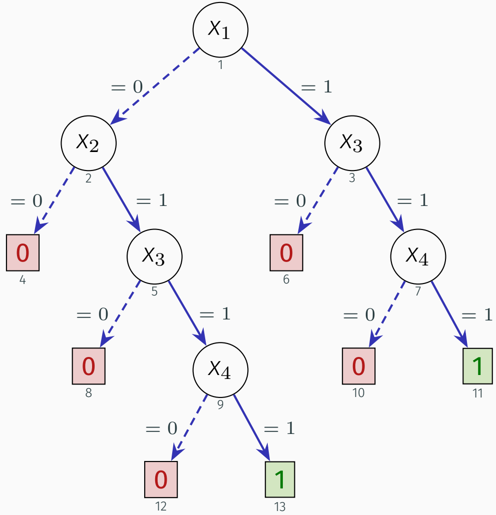
A concrete example

- AXps: $\{\{1,3,4\},\{2,3,4\}\}$
- Feature attribution:
  - SS: $\langle 0.083, 0.083, 0.417, 0.417 \rangle$



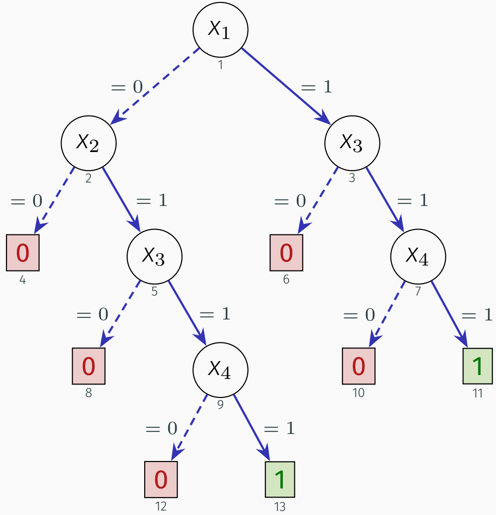
A concrete example

- AXps: $\{\{1,3,4\},\{2,3,4\}\}$
- Feature attribution:
  - SS: $\langle 0.083, 0.083, 0.417, 0.417\rangle$
  - B (norm.): $\langle 0.125, 0.125, 0.375, 0.375\rangle$



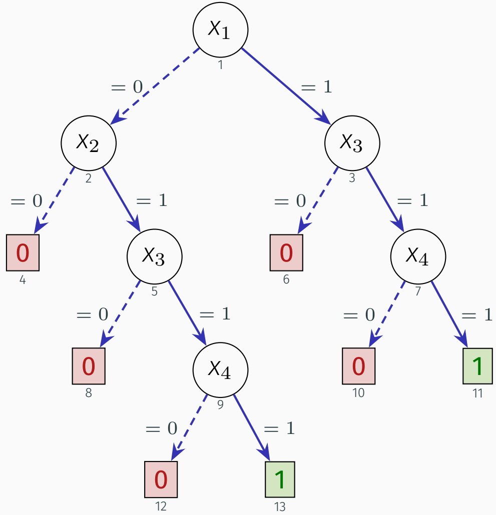
A concrete example

- AXps: $\{\{1,3,4\},\{2,3,4\}\}$
- Feature attribution:
  - SS: $\langle 0.083, 0.083, 0.417, 0.417 \rangle$
  - B (norm.): $\langle 0.125, 0.125, 0.375, 0.375 \rangle$
  - J (norm.): $\langle 0.111, 0.111, 0.389, 0.389 \rangle$



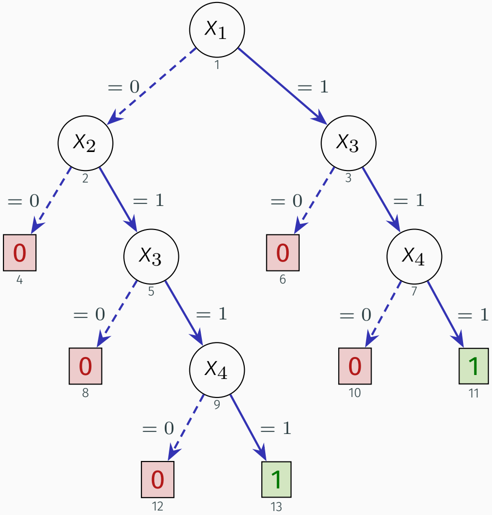
A concrete example

- AXps: $\{\{1,3,4\},\{2,3,4\}\}$
- Feature attribution:
  - SS: $\langle 0.083, 0.083, 0.417, 0.417\rangle$
  - B (norm.): $\langle 0.125, 0.125, 0.375, 0.375\rangle$
  - J (norm.): $\langle 0.111, 0.111, 0.389, 0.389\rangle$
  - HP: $\langle 0.167, 0.167, 0.333, 0.333\rangle$



A concrete example

- AXps: $\{\{1,3,4\},\{2,3,4\}\}$
- Feature attribution:
  - SS: $\langle 0.083, 0.083, 0.417, 0.417\rangle$
  - B (norm.): $\langle 0.125, 0.125, 0.375, 0.375\rangle$
  - J (norm.): $\langle 0.111, 0.111, 0.389, 0.389\rangle$
  - HP: $\langle 0.167, 0.167, 0.333, 0.333\rangle$
  - DP: $\langle 0.167, 0.167, 0.333, 0.333\rangle$



Questions?

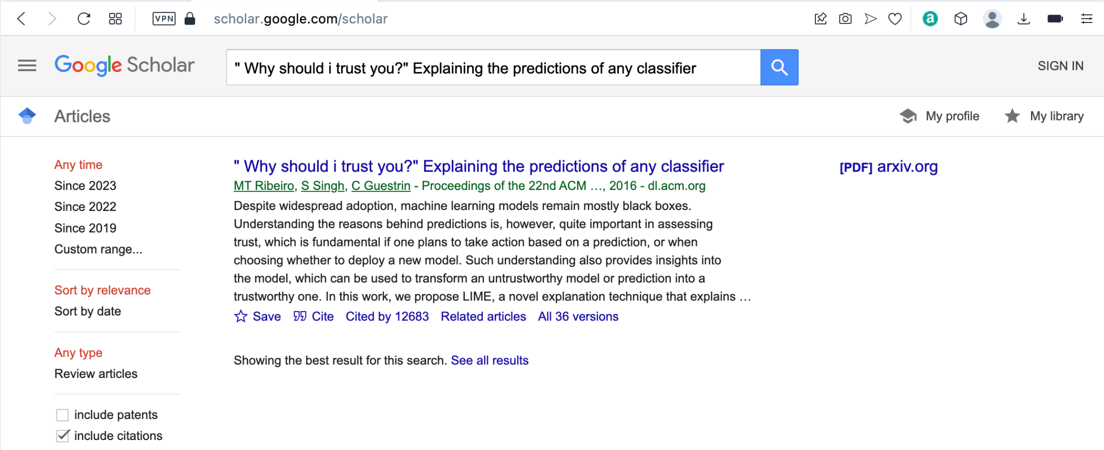
Unit #08

Conclusions & Research Directions

Outline – Unit #08

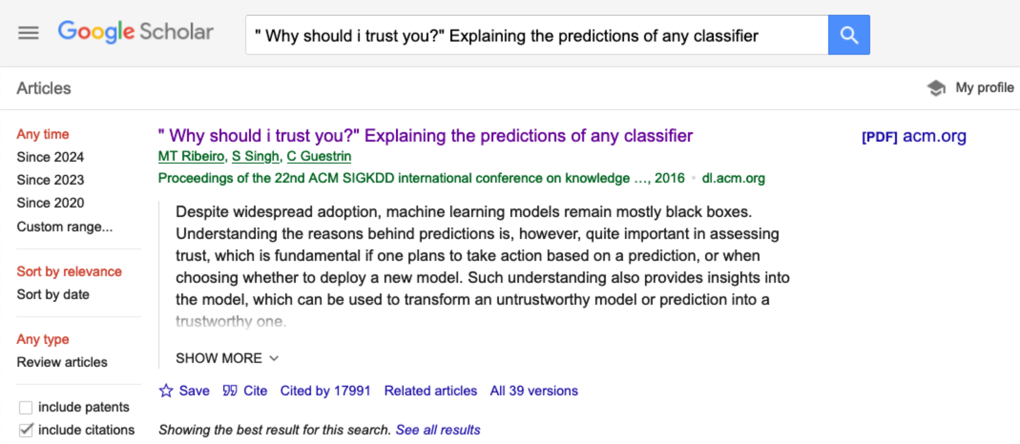
Can heuristic XAI's myths be stopped?

LIME on 2023/05/31:



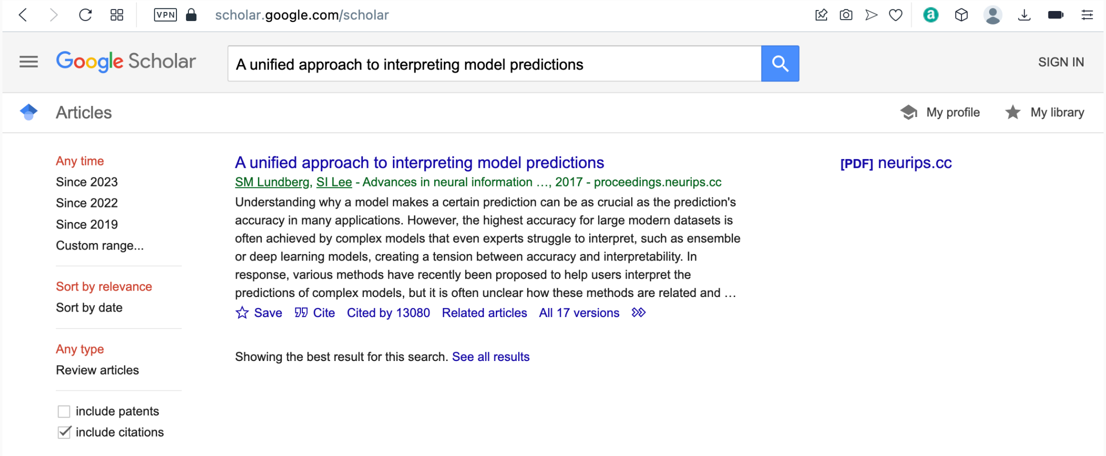
Can heuristic XAI's myths be stopped?

LIME on 2024/07/02:

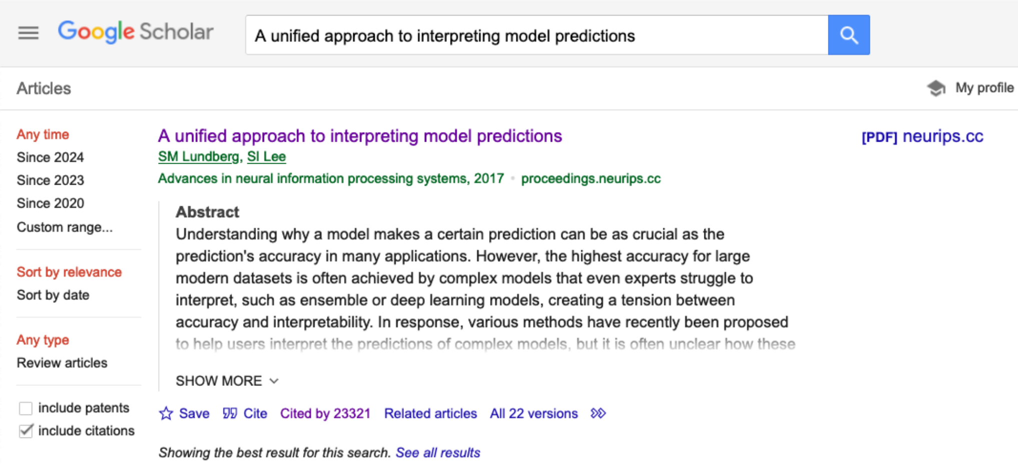


Can heuristic XAI's myths be stopped?

SHAP on 2023/05/31:



SHAP on 2024/07/02:



What's the bottom line?

What’s the bottom line?

- (Heuristic) XAI research experiences a persistent “*Don’t Look Up*” moment...

What's the bottom line?

- (Heuristic) XAI research experiences a persistent "*Don't Look Up*" moment...

BTW, there are a multitude of proposed uses of LIME/SHAP in medicine... ⚠

Some unsettling works...

- For DTs:
  - One AXp in polynomial-time [IIM20, HIIM21, IIM22]
  - All CXps in polynomial-time [HIIM21, IIM22]

Some unsettling works...

- For DTs:
  - One AXp in polynomial-time [IIM20, HIIM21, IIM22]
  - All CXps in polynomial-time [HIIM21, IIM22]

Declarative Reasoning on Explanations Using Constraint Logic Programming

Abstract. Explaining opaque Machine Learning (ML) models is an increasingly relevant problem. Current explanation in AI (XAI) methods suffer several shortcomings, among others an insufficient incorporation of background knowledge, and a lack of abstraction and interactivity with the user. We propose REASONX, an explanation method based on Constraint Logic Programming (CLP). REASONX can provide declarative, interactive explanations for decision trees, which can be the ML models under analysis or global/local surrogate models of any black-box model. Users can express background or common sense knowledge using linear constraints and MILP optimization over features of factual and contrastive instances, and interact with the answer constraints at different levels of abstraction through constraint projection. We present here the architecture of REASONX, which consists of a Python layer, closer to the user, and a CLP layer. REASONX's core execution engine is a Prolog meta-program with declarative semantics in terms of logic theories.

arXiv:2309.00422v1 [cs.AI] 1 Sep 2023

Some unsettling works...

- For DTs:
  - One AXp in polynomial-time [IIM20, HIIM21, IIM22]
  - All CXps in polynomial-time [HIIM21, IIM22]

HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA240183

Exploring Large Language Models Capabilities to Explain Decision Trees

Some unsettling works...

- For DTs:
  - One AXp in polynomial-time [IIM20, HIIM21, IIM22]
  - All CXps in polynomial-time [HIIM21, IIM22]

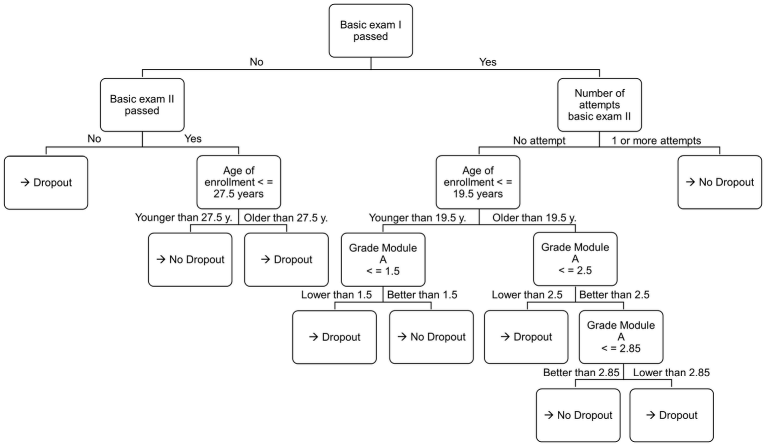
Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658953

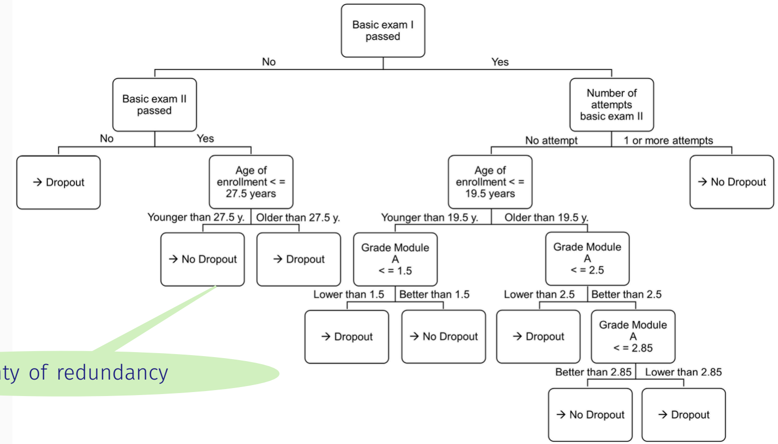
Some unsettling works...

- For DTs:
  - One AXp in polynomial-time [IIM20, HIIM21, IIM22]
  - All CXps in polynomial-time [HIIM21, IIM22]



Some unsettling works...

- For DTs:
  - One AXp in polynomial-time [IIM20, HIIM21, IIM22]
  - All CXps in polynomial-time [HIIM21, IIM22]



Plenty of redundancy

Outline – Unit #08

Conclusions

- Covered logic-based (aka symbolic, aka formal) XAI & its recent progress:
  - Abductive & contrastive explanations
  - Reviewed their computation in practice
  - Duality & enumeration
  - Other explainability queries – feature necessity & relevancy

Conclusions

- Covered logic-based (aka symbolic, aka formal) XAI & its recent progress:
  - Abductive & contrastive explanations
  - Reviewed their computation in practice
  - Duality & enumeration
  - Other explainability queries – feature necessity & relevancy
- Showed that formal XAI **disproves** some myths of (heuristic) XAI:
  - Explainability using intrinsic interpretability is a **myth**
  - The rigor of model-agnostic explanations is a **myth**
  - The rigor of SHAP scores as a measure of relative feature importance is a **myth**

Conclusions

- Covered logic-based (aka symbolic, aka formal) XAI & its recent progress:
  - Abductive & contrastive explanations
  - Reviewed their computation in practice
  - Duality & enumeration
  - Other explainability queries – feature necessity & relevancy
- Showed that formal XAI **disproves** some myths of (heuristic) XAI:
  - Explainability using intrinsic interpretability is a **myth**
  - The rigor of model-agnostic explanations is a **myth**
  - The rigor of SHAP scores as a measure of relative feature importance is a **myth**
- Demonstrated tight connection between (rigorous) feature selection and (rigorous) feature attribution in XAI

Conclusions

- Covered logic-based (aka symbolic, aka formal) XAI & its recent progress:
  - Abductive & contrastive explanations
  - Reviewed their computation in practice
  - Duality & enumeration
  - Other explainability queries – feature necessity & relevancy
- Showed that formal XAI **disproves** some myths of (heuristic) XAI:
  - Explainability using intrinsic interpretability is a **myth**
  - The rigor of model-agnostic explanations is a **myth**
  - The rigor of SHAP scores as a measure of relative feature importance is a **myth**
- Demonstrated tight connection between (rigorous) feature selection and (rigorous) feature attribution in XAI
- Symbolic XAI exhibits links with many fields of research:
  machine learning, artificial intelligence, formal methods, automated reasoning, optimization, computational social choice (& game theory), etc.

Research directions

Research directions

- Scalabilitty, scalability, and scalability

Research directions

- Scalabilitty, scalability, and scalability
- Probabilitistic explanations

Research directions

- Scalabilitty, scalability, and scalability
- Probabilitistic explanations
- Distance-restricted explanations

Research directions

- Scalabilitty, scalability, and scalability
- Probabilitistic explanations
- Distance-restricted explanations
- Rigorous feature attribution

Research directions

- Scalabilitty, scalability, and scalability
- Probabilitistic explanations
- Distance-restricted explanations
- Rigorous feature attribution
- Preferred explanations

Research directions

- Scalabilitty, scalability, and scalability
- Probabilitistic explanations
- Distance-restricted explanations
- Rigorous feature attribution
- Preferred explanations
- Certified XAI tools

Research directions

- Scalabilitty, scalability, and scalability
- Probabilitistic explanations
- Distance-restricted explanations
- Rigorous feature attribution
- Preferred explanations
- Certified XAI tools
- New topics from discussions with participants of ESSAI'24 – **Thank you!**

Research directions

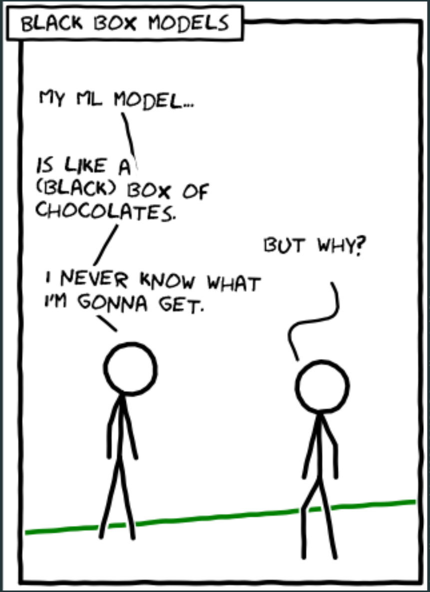
- Scalabilitty, scalability, and scalability
- Probabilitistic explanations
- Distance-restricted explanations
- Rigorous feature attribution
- Preferred explanations
- Certified XAI tools
- New topics from discussions with participants of ESSAI'24 – **Thank you!**
- ... And trying to curb the massive momentum of (heuristic) XAI **myths**!

What this course covered

- Lecture 01 – units:
  - #01: Foundations
- Lecture 02 – units:
  - #02: Principles of symbolic XAI – feature selection
  - #03: Tractability in symbolic XAI (& myth of interpretability)
- Lecture 03 – units:
  - #04: Intractability in symbolic XAI (& myth of model-agnostic XAI)
  - #05: Explainability queries
- Lecture 04 – units:
  - #06: Advanced topics
- Lecture 05 – units:
  - #07: Principles of symbolic XAI – feature attribution (& myth of Shapley values in XAI)
  - #08: Conclusions & research directions

Q & A

Acknowledgment: joint work with X. Huang, Y. Izza, O. Létoffé, A. Ignatiev, N. Narodytska, M. Cooper, N. Asher, A. Morgado, J. Planes, et al.



https://arxiv.org/abs/1901.01686 & http://cmx.io/edit/

References i

[ABBM21] Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet.
The tractability of SHAP-score-based explanations for classification over deterministic and decomposable boolean circuits.
In *AAAI*, pages 6670–6678, 2021.

[ABBM23] Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet.
On the complexity of SHAP-score-based explanations: Tractability via knowledge compilation and non-approximability results.
J. Mach. Learn. Res., 24:63:1–63:58, 2023.

[ACL03] Nicolas-Gabriel Andjiga, Fréderic Chantreuil, and Dominique Lepelley.
La mesure du pouvoir de vote.
Mathématiques et sciences humaines. Mathematics and social sciences, (163), 2003.

[BI65] John F Banzhaf III.
Weighted voting doesn't work: A mathematical analysis.
Rutgers L. Rev., 19:317, 1965.

[BIL+24] Gagan Biradar, Yacine Izza, Elita Lobo, Vignesh Viswanathan, and Yair Zick.
Axiomatic aggregations of abductive explanations.
In *AAAI*, pages 11096–11104, 2024.

References ii

[CH04] Hana Chockler and Joseph Y Halpern.
Responsibility and blame: A structural-model approach.
Journal of Artificial Intelligence Research, 22:93–115, 2004.

[Col71] James S Coleman.
Control of collectivities and the power of a collectivity to act.
In Bernhardt Lieberman, editor, *Social choice*, chapter 2.10. Gordon and Breach, New York, 1971.

[DP78] John Deegan and Edward W Packel.
A new index of power for simple n-person games.
International Journal of Game Theory, 7:113–123, 1978.

[DSZ16] Anupam Datta, Shayak Sen, and Yair Zick.
Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems.
In *IEEE S&P*, pages 598–617, 2016.

[HIIM21] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On efficiently explaining graph-based classifiers.
In *KR*, November 2021.
Preprint available from `https://arxiv.org/abs/2106.01350`.

[HM23a] Xuanxiang Huang and João Marques-Silva.
The inadequacy of Shapley values for explainability.
CoRR, abs/2302.08160, 2023.

[HM23b] Xuanxiang Huang and Joao Marques-Silva.
A refutation of shapley values for explainability.
CoRR, abs/2309.03041, 2023.

[HM23c] Xuanxiang Huang and Joao Marques-Silva.
Refutation of shapley values for XAI – additional evidence.
CoRR, abs/2310.00416, 2023.

[HMS24] Xuanxiang Huang and Joao Marques-Silva.
On the failings of Shapley values for explainability.
International Journal of Approximate Reasoning, page 109112, 2024.

[HP83] Manfred J Holler and Edward W Packel.
Power, luck and the right index.
Journal of Economics, 43(1):21–29, 1983.

[IIM20] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On explaining decision trees.
CoRR, abs/2010.11034, 2020.

[IIM22] Yacine Izza, Alexey Ignatiev, and João Marques-Silva.
On tackling explanation redundancy in decision trees.
J. Artif. Intell. Res., 75:261–321, 2022.

[Joh78] Ronald John Johnston.
On the measurement of power: Some reactions to Laver.
Environment and Planning A, 10(8):907–914, 1978.

[LC01] Stan Lipovetsky and Michael Conklin.
Analysis of regression in game theory approach.
Applied Stochastic Models in Business and Industry, 17(4):319–330, 2001.

[LHAMS24] Olivier Létoffé, Xuanxiang Huang, Nicholas Asher, and Joao Marques-Silva.
From SHAP scores to feature importance scores.
CoRR, abs/2405.11766, 2024.

[LHMS24] Olivier Létoffé, Xuanxiang Huang, and Joao Marques-Silva.
On correcting SHAP scores.
CoRR, abs/2405.00076, 2024.

[LL17] Scott M. Lundberg and Su-In Lee.
A unified approach to interpreting model predictions.
In *NIPS*, pages 4765–4774, 2017.

[MH23] Joao Marques-Silva and Xuanxiang Huang.
Explainability is NOT a game.
CoRR, abs/2307.07514, 2023.

[MSH24] Joao Marques-Silva and Xuanxiang Huang.
Explainability is *Not* a game.
Commun. ACM, 67(7):66–75, jul 2024.

[Pen46] Lionel S Penrose.
The elementary statistics of majority voting.
Journal of the Royal Statistical Society, 109(1):53–57, 1946.

[Sha53] Lloyd S. Shapley.
A value for n-person games.
Contributions to the Theory of Games, 2(28):307–317, 1953.

[SK10] Erik Strumbelj and Igor Kononenko.
An efficient explanation of individual classifications using game theory.
J. Mach. Learn. Res., 11:1–18, 2010.

[SK14] Erik Strumbelj and Igor Kononenko.
Explaining prediction models and individual predictions with feature contributions.
Knowl. Inf. Syst., 41(3):647–665, 2014.

[SS54] Lloyd S Shapley and Martin Shubik.
A method for evaluating the distribution of power in a committee system.
American political science review, 48(3):787–792, 1954.

[VLSS21] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu.
On the tractability of SHAP explanations.
In *AAAI*, pages 6505–6513, 2021.

[VLSS22] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu.
On the tractability of SHAP explanations.
J. Artif. Intell. Res., 74:851–886, 2022.