Winter values for XAI

Benedek Fulop

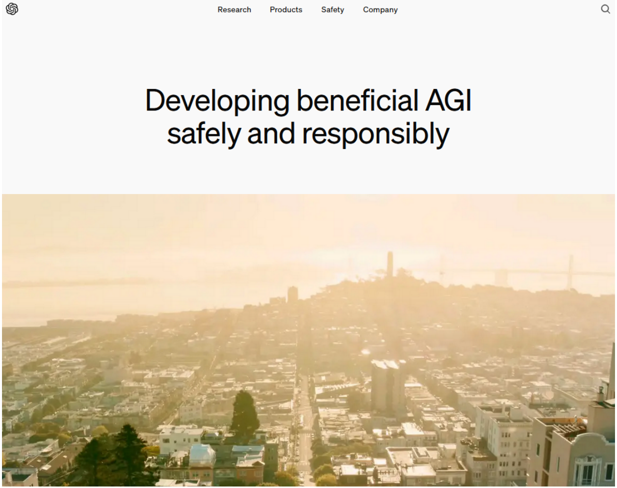
MSc Data Science and Business Analytics

June 27, 2024

Outline

- Introduction
- Methods
- Implementation
- Demonstrations
- Conclusion

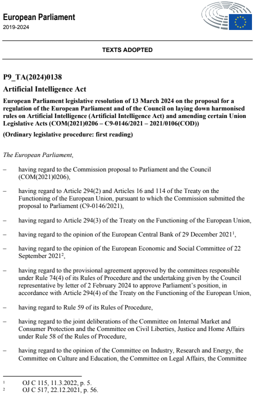
Introduction



Developing beneficial AGI safely and responsibly

: We are at 128K tokens

Introduction



European Parliament
2019-2024

TEXTS ADOPTED

P9_TA(2024)0138

Artificial Intelligence Act

European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))

(Ordinary legislative procedure: first reading)

The European Parliament,

– having regard to the Commission proposal to Parliament and the Council (COM(2021)0206),

– having regard to Article 294(2) and Articles 16 and 114 of the Treaty on the Functioning of the European Union, pursuant to which the Commission submitted the proposal to Parliament (C9-0146/2021),

– having regard to Article 294(3) of the Treaty on the Functioning of the European Union,

– having regard to the opinion of the European Central Bank of 29 December 2021[1],

– having regard to the opinion of the European Economic and Social Committee of 22 September 2021[2],

– having regard to the provisional agreement approved by the committees responsible under Rule 74(4) of its Rules of Procedure and the undertaking given by the Council representative by letter of 2 February 2024 to approve Parliament’s position, in accordance with Article 294(4) of the Treaty on the Functioning of the European Union,

– having regard to Rule 59 of its Rules of Procedure,

– having regard to the joint deliberations of the Committee on Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs under Rule 58 of the Rules of Procedure,

– having regard to the opinion of the Committee on Industry, Research and Energy, the Committee on Culture and Education, the Committee on Legal Affairs, the Committee

[1] OJ C 115, 11.3.2022, p. 5.
[2] OJ C 517, 22.12.2021, p. 56.

: Article 86: We have a right to an explanation

Introduction

BLUEPRINT FOR AN AI BILL OF RIGHTS

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

OSTP

: USA "Notice and Explanation"

Methods

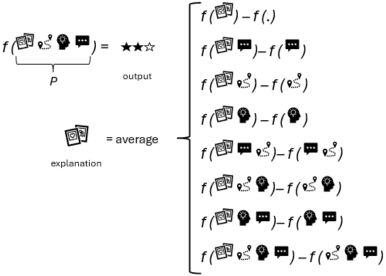
Shapley et al. (1953): Shapley value

$$\phi_i(f) = \frac{1}{|P|!} \sum_{\pi \epsilon \Pi} [f(p^i_\pi \cup i) - f(p^i_\pi)]. \tag{1}$$

- **EFFICIENCY.** $\sum_{i \epsilon N} \phi_i(f) = f(N)$.
- **SYMMETRY.** If features i and j are symmetric with respect to model f, then $\phi_i(f) = \phi_j(f)$.
- **DUMMY.** If i is a dummy feature, $f(S \cup i) - f(S) = 0$, $\forall S \subset P$, then $\phi_i(f) = 0$.
- **ADDITIVITY.** $\phi(f + g) = \phi(f) + \phi(g)$, where the model $f + g$ is defined by $(f + g)(S) = f(S) + g(S)\ \forall S$.

Methods

For example:



: Illustration of the computation of Shapley values

Methods

Issues with Shapley values:

- Computational cost $O(2^P)$
- ODD samples

The culprit: The **SYMMETRY** axiom.

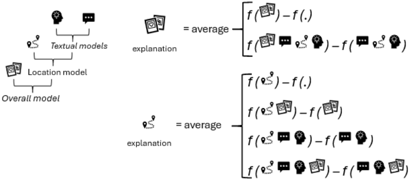
Methods

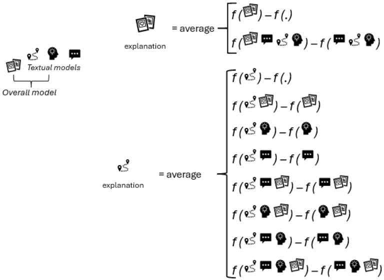
Winter (1989): Winter value

$$\phi_i(B, f) = \frac{1}{|\Pi_1|} \sum_{\pi \epsilon \Pi_1} [f(p_\pi^i \cup i) - f(p_\pi^i)] \tag{2}$$

Example level structure $P = [[1, 2], [[3], [4], [5, 6]]], [7, 8]]$
For a given level structure $B = (B_1, B_2, ..., B_m)$, define
$\Pi_m = \{\pi \in \Pi$; for each $l, j \in S \in B_m$ and $i \in N, \pi(l) < \pi(i) < \pi(j)$ implies $i \in S\}$, and
$\Pi_r = \{\pi \in \Pi_{r+1}$; for each $l, j \in S \in B_r$ and $i \in N, \pi(l) < \pi(i) < \pi(j)$ implies $i \in S\}$.

Methods


Textual models
Location model
Overall model
explanation
= average
f(.)
explanation
= average
f(.)


Textual models
Overall model
explanation
= average
f(.)
explanation
= average
f(.)

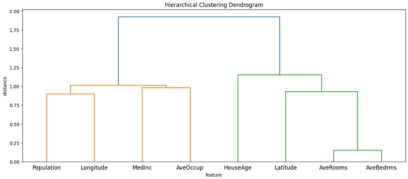
Methods

- Computational cost $O(P^2)$ for balanced clustering trees
- Related features masked together

Computation

Lundberg and Lee (2017): SHapley Additive exPlanations (SHAP): The PartitionExplainer

The main idea is to use the hierarchical correlation of features to calculate Shapley values through the hierarchy. This essentially implements the calculation of Winter values specifically for binary trees from scipy.



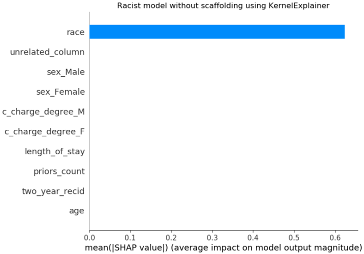
- No non-binary partitions
- No way to include domain knowledge

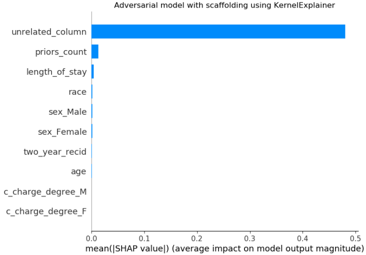
Computation

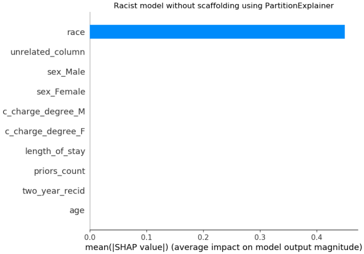
Linear approach

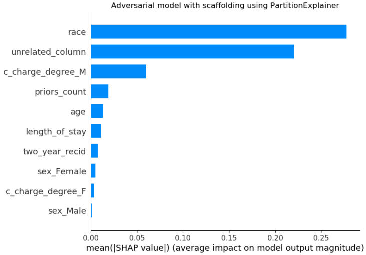
- Define a class to encode an n-ary tree structure
- Assign every node all possible permutations of sibling nodes.
- Depth first traversal of the tree collecting all permutations on path and creating all possible combinations of permutations
- Create a boolean mask for every node that encodes all leafs below a node, the features that fall into the given partition
- Combine the masks according to the combinations of permutations. This will be the operations to calculate the Winter values.
- Call the model on the unique mask combinations and calculate the average marginals as usual.

Slack et al. (2020): Fooling SHAP

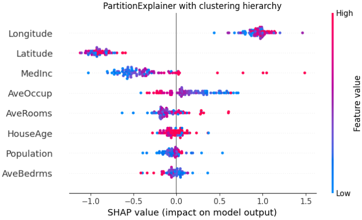

Racist model without scaffolding using KernelExplainer

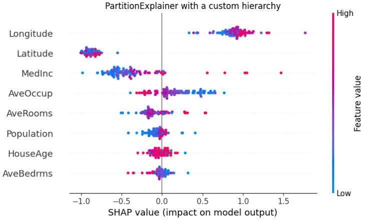

Adversarial model with scaffolding using KernelExplainer


Racist model without scaffolding using PartitionExplainer


Adversarial model with scaffolding using PartitionExplainer

Scipy clustering vs. Custom Hierarchy

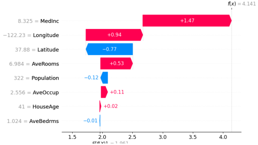

PartitionExplainer with clustering hierarchy


PartitionExplainer with a custom hierarchy

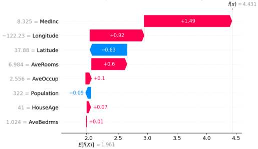
According to Winter (2002), ”how far can we depart from the Shapley value? ’Not much’, according to Young (1985)” (p. 2033).

Shapley vs. Winter

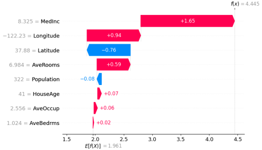
Shapley value



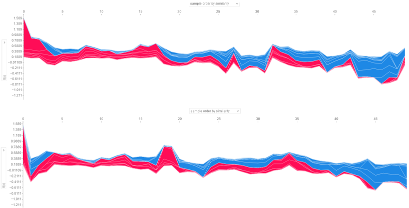
PartitionExplainer for instance 0 with the clustering method



PartitionExplainer for instance 0 with a custom hierarchy



Kernel Explainer vs. Partition Explainer



: Differences between Kernel Explainer results (top) and Winter values with custom hierarchy (bottom)

Conclusion

Current results

- We extended the PartitionExplainer to compute the Winter values for any partitions
- We formalised the interpretation of the clustering based Winter values

Future directions

- Adopt an iterative computation method for the Winter values
- Create methods for including custom hierarchies on image, graph, and text data
- Test the robustness of the Winter method to adversarial attacks

References I

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Shapley, L. S. et al. (1953). A value for n-person games.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Winter, E. (1989). A value for cooperative games with levels structure of cooperation. *International Journal of Game Theory*, 18:227–240.