Local vs. Global Interpretability: A Computational Complexity Perspective

What is computationally harder - interpreting an ML model *locally* or *globally*?
The answer: It depends…

Introduction

- Can we *mathematically* assess how interpretable an ML model is?
- Previous work addresses this using *computational complexity theory* – the harder it is to generate an explanation, the less interpretable a model is.
- For example, some explanation forms can be efficiently obtained for decision trees and linear models but are intractable for neural networks.

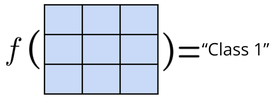
Local vs. Global Interpretability

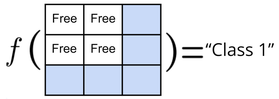
- The problem: previous work mainly analyzed the complexity of obtaining *local* explanation forms.
- Interpretability can be *local* (understanding specific decisions) or *global* (understand general behaviors)
- For example, Molnar et. al. claim that while the weights of a linear model help interpret local decisions, they don't explain global behaviors.

We analyze the computational complexity of obtaining local explanations (for a specific instance x) compared to their global variants (for *all* x in a given domain)

Local Sufficient Reason

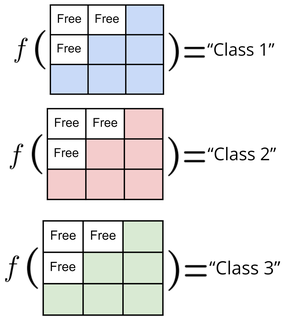
$$\forall(\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})]$$





Global Sufficient Reason

$$\forall(\mathbf{x}, \mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})]$$



Explanation forms we analyze:

- Feature selection – selecting minimum local vs. global sufficient reasons
  - Identifying local vs. global redundant or necessary features
  - A probabilistic notion of sufficiency – whether local or global

In **linear models**, feature selection is *harder to perform globally than locally:*

Explanation Type	Local	Global
Check Sufficient Reason	PTIME	coNP-Complete
Minimum Sufficient Reason	PTIME	coNP-Complete

Less globally interpretable

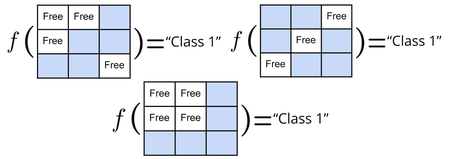
In **neural networks and decision trees** – the *local task is harder than the global one*.

	Decision Trees		Neural Networks	
Explanation Type	Local	Global	Local	Global
Minimum Sufficient Reason	NP-Complete	PTIME	Σ_2^P-Complete	coNP-Complete
Feature Redundancy	coNP-Complete	PTIME	Π_2^P-Complete	coNP-Complete

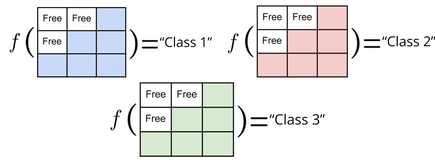
Less locally interpretable

Why does this happen?

There can be an exponential number of minimal *local* sufficient reasons:



There is one unique minimal *global* sufficient reason:



This property simplifies the global task compared to the local one. However, for linear models, the weights enable the development of efficient algorithms, which is not evident in neural networks and decision trees.

Conclusion

- **We prove that there is often a *strict complexity gap* between obtaining an explanation locally vs. globally.**
- **In some cases, such gaps justify folklore claims (e.g., linear models can be interpreted locally using their weights but not globally).**
- **In other cases, they yield unexpected outcomes (e.g., neural networks and decision trees are easier to interpret globally than locally in some contexts).**

Shahaf Bassan, Guy Amir, Guy Katz; The Hebrew University of Jerusalem