Stability theory of game-theoretic group feature explanations for machine learning models

Alexey Miroshnikov

joint work with Konstantinous Kotsiopoulos, Khashayar Filom, Arjun Ravi Kannan

Emerging Capabilities & Data Science Research Group, Discover Financial Services

SIAM Conference on Mathematics of Data Science, October 21, 2024
Mathematics of Explainable AI with Applications to Finance and Medicine

Motivation

Introduction

- Contemporary predictive ML models are complex:

  Neural Networks (NN), Gradient Boosting Machines (GBM), Semi-supervised methods

- Interpretability is crucial for business adoption, regulatory oversight, and human acceptance and trust:

  Banking, Insurance, Healthcare

- Accuracy may come at the expense of interpretability [P. Hall, 2018].

Regulatory requirements

- ML models, and strategies that rely on ML models, are subject to laws and regulations (e.g. ECOA, EEOA).
- Financial institutions in the United States (US) are required under the ECOA to notify declined or negatively impacted applicants of the main factors that led to the adverse action.
- Common approaches: Post-hoc individualize model explanations, Self-interpretable models.

Individualized explanations

Notation

- $x \to f(x)$ ML model (classification score or regressor)
- (X, Y), where $X = (X_1, \dots, X_n)$ are features, $Y \in \mathbb{R}$ is response variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
- P_X a pushforward probability measure, $P_X(A) = \mathbb{P}(X \in A), \mathcal{B}(\mathbb{R}^n)$.

Definition

A model explainer quantifies the contribution of an observation $x = (x_1, x_2, \dots x_n) \sim X$ to the value $f(x)$. Formally:

$$\mathbb{R}^n \ni x \to E\left(x; f, X, \mathcal{J}_f\right) = (E_1, E_2, \dots E_n) \in \mathbb{R}^n$$

where the model f, the random vector X and model implementation $\mathcal{J}_f$ serve as parameters.

Games and game values

Objective: Study explanations based on game values for the marginal and conditional games.

- Cooperative game (N, v).
  - $N = \{1,2, \dots, n\}$, set of players.
  - v is utility. $v(S)$ is the worth of the coalition $S \subseteq N$.
- Game value. A map $(N, v) \to h[N, v] = \{h_i[N, v]\}_{i=1}^{n} \in \mathbb{R}^n$.

Assumption: We study game values in the marginalist form

$$h_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(S, n) \cdot \big(v(S \cup i) - v(S)\big)$$

h is linear (LN), symmetric (SM).

Example: Shapley value [Shapley, 1953]

$\varphi_i[v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \big(v(S \cup i) - v(S)\big)$ which is linear, symmetric, efficient (EF) $\sum_i \varphi_i[N, v] = v(N)$.

Other examples: Banzhaf value (1965), Owen value (1976).

Individualized explanations with deterministic games for ML models

Game theoretic approach for ML explainability has been explored in Štrumbelj & Kononenko (2014), Lundberg & Lee (2017)

Definition

Given (x, X, f) and $S \subset N = \{1,2, \dots n\}$

- $v_*^{CE}(S, x; X, f) = \mathbb{E}[f(X_S, X_{-S})|X_S = x_s]$, conditional game
- $v_*^{ME}(S, x; X, f) = \mathbb{E}[f(x_S, X_{-S})]$, marginal game

Definition

Given a game value $h[N, v]$ individualized conditional and marginal explanations are defined:

- $x \to h_*^{CE}(x) = h[N, v_*^{CE}(\cdot, x)] \in \mathbb{R}^n$, $x \to h_*^{ME}(x) = h[N, v_*^{ME}(\cdot, x)] \in \mathbb{R}^n$

Marginal vs conditional (informally)

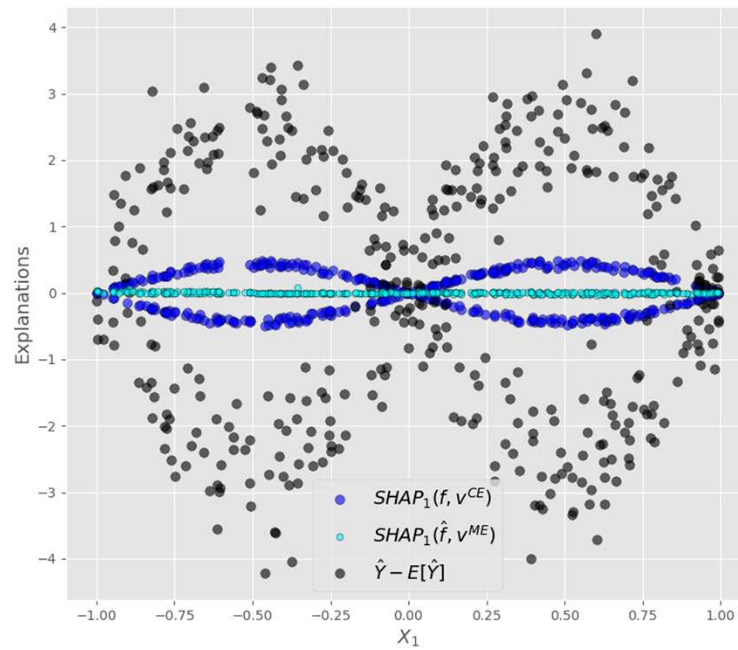
Marginal game

- v_*^{ME} explores the input-output relationship $(x, f(x))$, $x \sim X$.
- $h[N, v_*^{ME}]$ are "consistent" with the model $f(x)$

Conditional game

- v_*^{CE} explores the contribution of $x \sim X$ in the context of the observational graph $\Omega \ni \omega \to \big(X(\omega), f\big(X(\omega)\big)\big)$.
- $h[N, v_*^{CE}]$ are "consistent" with the data and $f(X)$

$$Y = f(X) = X_2 X_3 \ \mid \ X_2 = \sin(\pi X_1) + \epsilon$$



Random games and operators

In our analysis we study game values of random games.

Random games

- $v^{CE}(S;X,f) = v_*^{CE}(S,x;X,f)|_{x=X} \in (\Omega,\mathcal{F},\mathbb{P})$
- $v^{ME}(S;X,f) = v_*^{ME}(S,x;X,f)|_{x=X} \in (\Omega,\mathcal{F},\mathbb{P})$

Operators based on $\boldsymbol{h}[N,v]$

- $\bar{\mathcal{E}}^{CE}[f] = \left(\bar{\mathcal{E}}_1^{CE},\dots,\bar{\mathcal{E}}_n^{CE}\right)[f]: L^2(\mathbb{R}^n,P_X) \mapsto L^2(\Omega,\mathbb{P})^n$ by $\bar{\mathcal{E}}_i^{CE}[f] \coloneqq h_i[N,v^{CE}(\cdot;X,f)]$
- $\bar{\mathcal{E}}^{ME}[f] = (\bar{\mathcal{E}}_1^{ME},\dots,\bar{\mathcal{E}}_n^{ME})[f]: L^2\left(\mathbb{R}^n,\tilde{P}_X\right) \mapsto L^2(\Omega,\mathbb{P})^n$ by $\bar{\mathcal{E}}_i^{ME}[f] \coloneqq h_i[N,v^{ME}(\cdot;X,f)]$

where $\tilde{P}_X = \frac{1}{2^n}\sum_{S\subseteq N} P_{X_S} \otimes P_{X_{-S}}$.

Note: $\tilde{P}_X = P_X$ if features are independent.

Continuity I

Theorem [AM, Kotsiopoulos, Filom, Ravi Kannan (2022)]

- $\left(\bar{\mathcal{E}}^{CE}, L^2(P_X)\right)$ is a **well-defined bounded linear** operator such that

$$\|\bar{\mathcal{E}}^{CE}[f_1] - \bar{\mathcal{E}}^{CE}[f_2]\|_{L^2(\mathbb{P})} \leq C(w,n) \cdot \|f_1 - f_2\|_{L^2(P_X)}$$

If h is efficient then $C(w,n) = 1$.

- $\left(\bar{\mathcal{E}}^{ME}, L^2(\tilde{P}_X)\right)$ is a **well-defined bounded linear** operator such that

$$\|\bar{\mathcal{E}}^{ME}[f_1] - \bar{\mathcal{E}}^{ME}[f_2]\|_{L^2(\mathbb{P})} \leq \tilde{C}(w,n) \cdot \|f_1 - f_2\|_{L^2(\tilde{P}_X)}$$

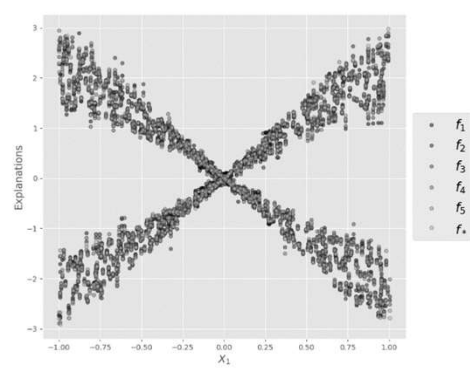
Note: $f_1(X) \approx f_2(X)$ in $L^2(\mathbb{P}) \Rightarrow h[v^{CE}(f_1)] \approx h[v^{CE}(f_2)]$ in $L^2(\mathbb{P})$.

Example: Rashomon effect on marginal explanations

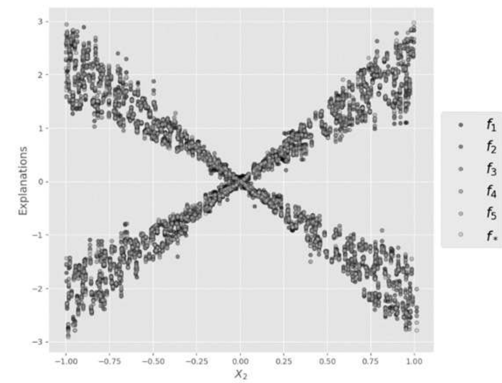
Synthetic model

$$
\begin{aligned}
&Z \sim Unif(-1,1)\\
&X_1 = Z + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0, 0.05),\\
&X_2 = \sqrt{2}\sin(Z(\pi/4)) + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, 0.05),\\
&X_3 \sim Unif\big([-1,-0.5] \cup [0.5,1]\big).
\end{aligned}
$$

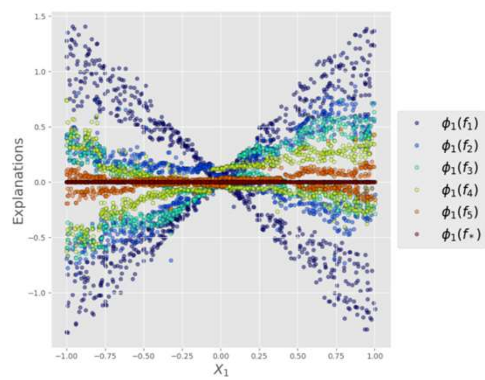
$$Y = f_*(X_1, X_2, X_3) + \epsilon_3 = 3X_2X_3 + \epsilon_3$$



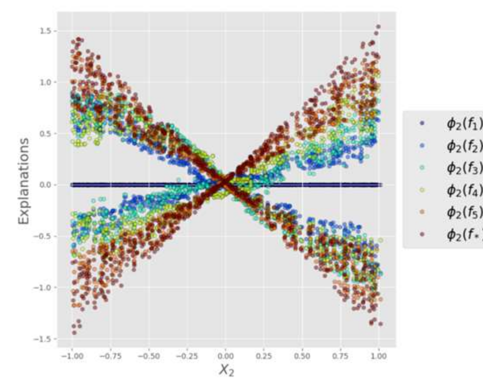
(a) Predictions vs X_1.



(b) Predictions vs X_2.



(a) Explanations φ_1 vs X_1.



(b) Explanations φ_2 vs X_2.

Continuity II

Questions regarding the marginal operator:

- Can the marginal operator be well-defined and bounded on a space equipped with $L^2(P_X)$-norm?
- Is there any relationship between boundedness and dependencies?

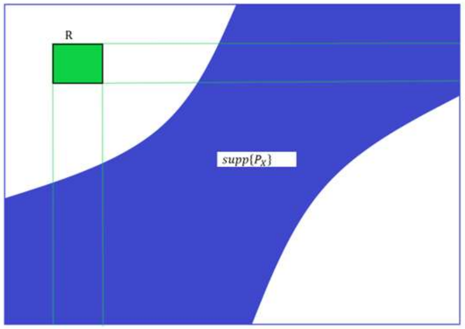
To answer these questions it is necessary to consider the two cases:

1. $\tilde{P}_X \ll P_X$ i.e. $\tilde{P}_X$ is absolutely continuous w.r.t. P_X
2. $\tilde{P}_X$ is not absolutely continuous w.r.t. P_X

Lemma [AM, Kotsiopoulos, Filom, Ravi Kannan (2022)]

- The marginal game (v^{ME}, H_X) on $H_X = \left(L^2(\tilde{P}_X)/H_X^0, \|\cdot\|_{L^2(P_X)}\right)$ is well-defined if and only if $\tilde{P}_X \ll P_X$.
- If $\tilde{P}_X \ll P_X$, $H_X = \left(L^2(\tilde{P}_X), \|\cdot\|_{L^2(P_X)}\right)$
- If $\tilde{P}_X \ll P_X$ then $r_X := \frac{d\,\tilde{P}_X}{d\,P_X} \in L^1(P_X)$ controls the amount of dependencies in the sense of:

$$W_1\left(\tilde{P}_X, P_X\right) \leq \int |x| \cdot |r_X(x) - 1|\, P_X(dx)$$



Continuity II

Theorem [AM, Kotsiopoulos, Filom, Ravi Kannan (2023,revised)]

Suppose $\tilde{P}_X \ll P_X$

- Bounded case. Suppose $r_X \in L^\infty(P_X)$. Then $(\bar{\mathcal{E}}^{ME}, H_X)$ is a **well-defined bounded linear** operator satisfying

$$\left\| \bar{\mathcal{E}}_i^{ME}[f] \right\|_{L^2(\mathbb{P})} \leq \left(1 + \|r_X - 1\|_{L^\infty(P_X)}\right) \cdot C_i(w) \cdot \|f\|_{L^2(P_X)}$$

- Unbounded case.

Let $S \subset N$. Suppose that there exists $T \subseteq S$ and $Q \subseteq -S$ such that

$$\sup\left\{ \frac{[P_{X_T} \otimes P_{X_Q}](A \times B)}{P_{(X_T,X_Q)}(A \times B)} \cdot P_{X_Q}(B),\ A \in \mathcal{B}(\mathbb{R}^{|T|}),\ B \in \mathcal{B}(\mathbb{R}^{|Q|}), P_{(X_T,X_Q)}(A \times B) > 0 \right\} = \infty. \quad \text{(UG)}$$

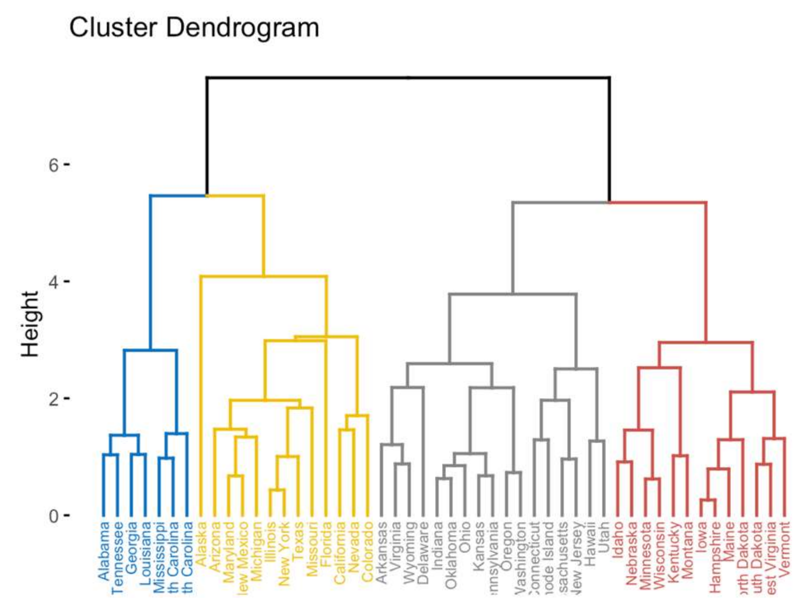
Then the map $f \in H_X \mapsto v^{ME}(S; X, f) \in L^2(\mathbb{P})$ is unbounded.

Suppose (UG) *holds with $T = \{i\}$ and $Q = \{j\}$ for two distinct indices $i, j \in \{1, 2, ..., n\}$ and that the game value weights $w(S, n) > 0$ for each proper subset $S \subset N$. Then $(\bar{\mathcal{E}}_i^{ME}, H_X)$, $(\bar{\mathcal{E}}_j^{ME}, H_X)$, and $(\bar{\mathcal{E}}^{ME}, H_X)$ are unbounded linear operators.*

Mitigation. Grouping features as a stabilization mechanism.

Computing explanations of groups formed by dependencies (e.g. variable clustering tree)

- Unifies marginal and conditional explanations and achieve stability of marginal explanations
- Removes splits of explanations across dependencies



Quotient game explainers

Given $\mathcal{P} = \{S_1, S_2, \dots S_m\}$, treat each group predictor X_{S_j} as a player $j \in \{1,2,\dots,m\}$

Quotient game: $v^{\mathcal{P}}(A) = v\big(\cup_{j\in A} S_j\big),\ A \subset M = \{1,2,\dots m\}$

Quotient game explainers: $f \mapsto h_j[M, v^{\mathcal{P}}(f)],\ v \in \{v^{CE}, v^{ME}\}$

Proposition [AM, Kotsiopoulos, Filom, Ravi Kannan (2023,revised)]

- If groups $\{X_{S_1}, X_{S_2}, \dots, X_{S_m}\}$ are independent, $h[v]$ is linear,

  $h_j[M, v^{CE,\mathcal{P}}(f)] = h_j[M, v^{ME,\mathcal{P}}(f)]$ and hence continuous in $L^2(P_X)$.

- Let $Q_A = \cup_{j\in A} S_j$. If $r_A = \frac{d(P_{X_{Q_A}} \otimes P_{X_{-Q_A}})}{dP_X}$ is bounded for $A \subseteq M$, then

  $H_X \ni f \to h_j[M, v^{ME,\mathcal{P}}(f)]$ is bounded in $L^2(P_X)$ with the bound

  $$\sim C(w) \cdot \max_{A\subseteq M}(r_A - 1)$$

References

- J.F. Banzhaf, Weighted voting doesn't work: a mathematical analysis. Rutgers Law Review 19, 317-343, (1965).
- P. Hall, N. Gill, An Introduction to Machine Learning Interpretability, O'Reilly. (2018).
- S.M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, 31st Conference on Neural Information Processing Systems, (2017).
- A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Mutual information-based group explainers with coalition structure for machine learning model explanations, *arXiv preprint* arXiv:2102.10878v4 (2022). Revised version: A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Stability theory of game-theoretic group feature explanations for machine learning models, *arXiv preprint,* arXiv:2102.10878v6 (2024).
- L. S. Shapley, A value for n-person games, Annals of Mathematics Studies, No. 28, 307-317 (1953).
- G. Owen, Values of games with a priori unions. In: Essays in Mathematical Economics and Game Theory (R. Henn and O. Moeschlin, eds.), Springer, 76 {88 (1977).
- E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst., 41, 3, 647-665, (2014).