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

- Alexey Miroshnikov
- Data Science Research Group, Discover Financial Services

- King Abdullah University of Science and Technology, Applied Mathematics Seminar, November 2024

1. Motivation

Motivation

Model Complexity

- 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.

Setup

Input

- (X, Y), where $X = (X_1, \dots, X_n)$ are features, $Y \in \mathbb{R}$ a response variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
- $x \to f(x) = \mathbb{E}[Y|X = x]$ or $\mathbb{P}(Y = 1|X = x)$ (regressor or classification score).
- P_X a pushforward probability measure, $P_X(A) = \mathbb{P}(X \in A)$, $\mathcal{B}(\mathbb{R}^n)$.

Setup

Input

- (X, Y), where $X = (X_1, \dots, X_n)$ are features, $Y \in \mathbb{R}$ a response variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
- $x \to f(x) = \mathbb{E}[Y|X = x]$ or $\mathbb{P}(Y = 1|X = x)$ (regressor or classification score).
- P_X a pushforward probability measure, $P_X(A) = \mathbb{P}(X \in A)$, $\mathcal{B}(\mathbb{R}^n)$.

(Individual or local) Model explainer

Quantifies the contribution of an observation $x = (x_1, x_2, \dots x_n) \sim X$ to the value $f(x)$.

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

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

Setup

Input

- (X, Y), where $X = (X_1, \dots, X_n)$ are features, $Y \in \mathbb{R}$ a response variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
- $x \to f(x) = \mathbb{E}[Y|X = x]$ or $\mathbb{P}(Y = 1|X = x)$ (regressor or classification score).
- P_X a pushforward probability measure, $P_X(A) = \mathbb{P}(X \in A)$, $\mathcal{B}(\mathbb{R}^n)$.

(Individual or local) Model explainer

Quantifies the contribution of an observation $x = (x_1, x_2, \dots x_n) \sim X$ to the value $f(x)$.

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

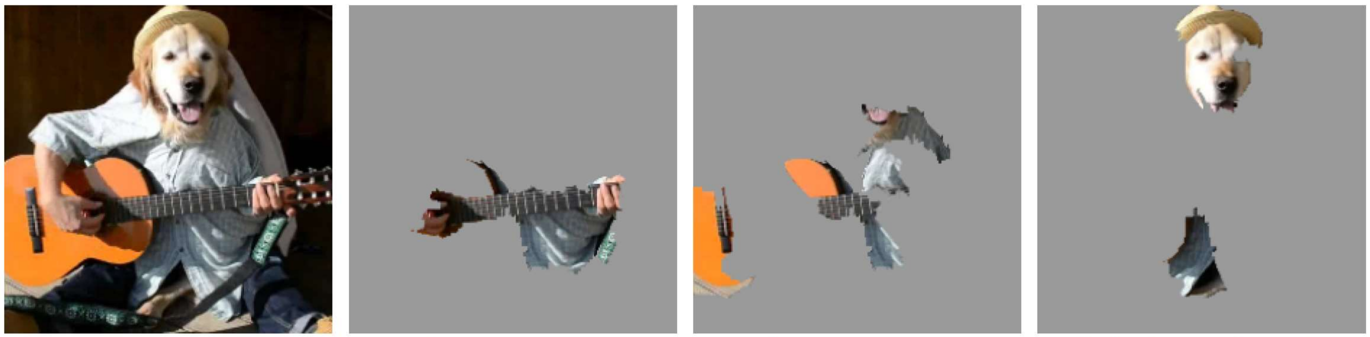
Here the model f, the random vector X and model implementation $\mathcal{I}_f$ serve as parameters.

Example

Linear model: $f(x) = a_1 x_1 + a_2 x_2 \dots + a_n x_n$. Set $E_i(x; f, X) = a_i(x_i - \mathbb{E}[X_i]), i \in N = \{1,2, \dots n\}$.

Example: image classification feature attribution

[Ribeiro et al. “Why should I trust you?”]

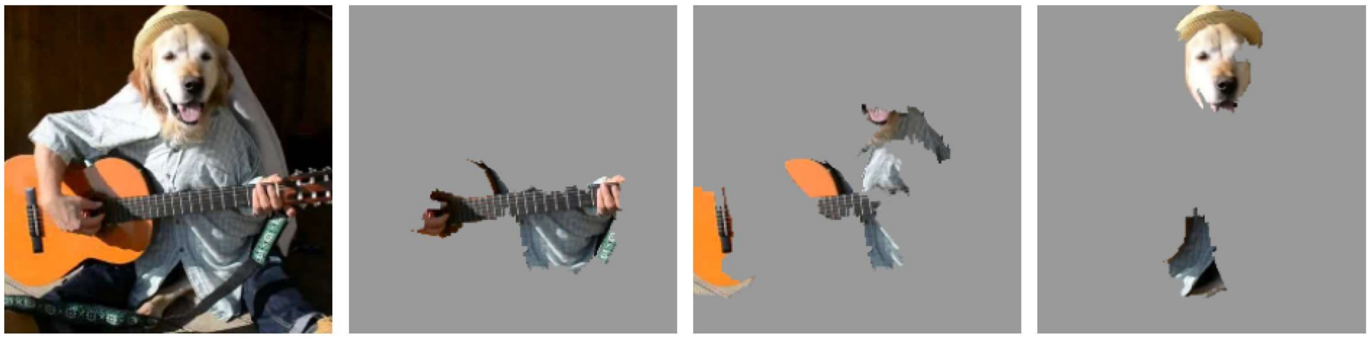


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Example: image classification feature attribution

[Ribeiro et al. “Why should I trust you?”]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google’s Inception neural network. The top 3 classes predicted are “Electric Guitar” ($p = 0.32$), “Acoustic guitar” ($p = 0.24$) and “Labrador” ($p = 0.21$)

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

Games and game values

- 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 $(N, v) \rightarrow h[N, v] = \{h_i[N, v]\}_{i=1}^{n} \in \mathbb{R}^n$.

  Game value 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).

Games and game values

- Cooperative game (N, v).
  - $N = \{1,2,\ldots,n\}$, set of players.
  - v is utility. $v(S)$ is the worth of the coalition $S \subseteq N$.

- Game value. A map $(N, v) \to h[N, v] = \{h_i[N, v]\}_{i=1}^{n} \in \mathbb{R}^n$.

  Game value 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]

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

Game theoretic approach for ML models

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

Marginal and conditional deterministic games

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

Marginal and conditional (local) explanations

Given a game value $h[N, v]$ conditional and marginal explanations of f at x are defined:

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

Marginal vs conditional

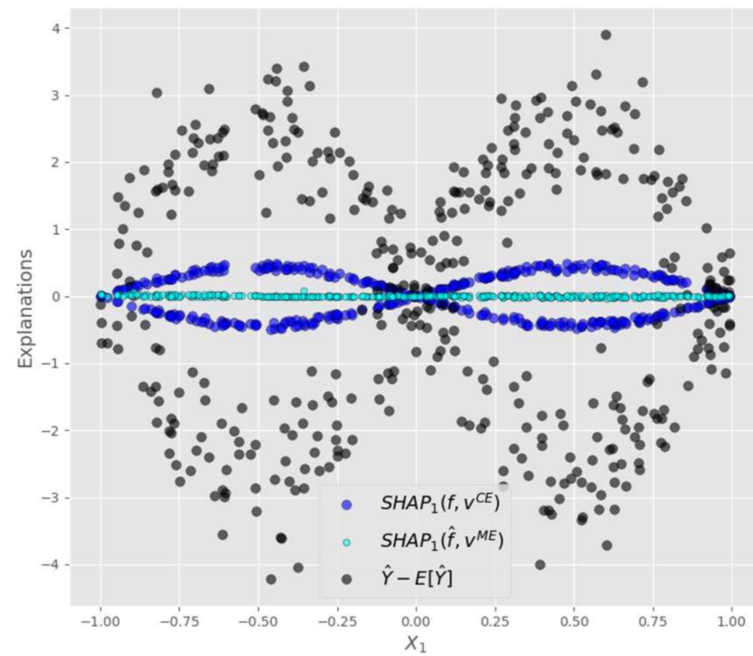
Marginal game

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

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



Random games and linear operator

Random games

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

Linearity

For $v \in \{v^{CE}, v^{ME}\}$ and two models f, g

- $v(S; X, \alpha \cdot f + g) \to \alpha \cdot v(S; X, f) + v(S; X, g),\, S \subseteq N$
- $h_i[N, v(\,\cdot\,; X, \alpha \cdot f + g)] \to \alpha \cdot h_i[N, v(\,\cdot\,; X, f)] + h_i[N, v(\,\cdot\,; X, g)]$

Random games and operators

Given a game value

$$h_i[N,v] = \sum_{S\subseteq N\setminus\{i\}} w(S,n)\cdot\big(v(S\cup i) - v(S)\big), i \in N = \{1,2,\ldots n\}$$

define linear operators

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

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 $L^2(\mathbb{P})$.

Rashomon effect on marginal explanations

Synthetic model

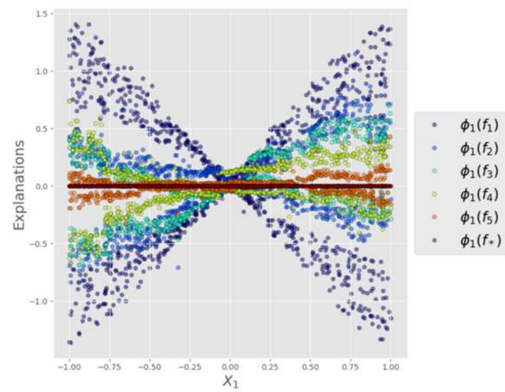
$$Y = f_*(X_1, X_2, X_3) + \epsilon_3 = 3X_2X_3 + \epsilon_3$$

$$Z \sim Unif(-1, 1)$$

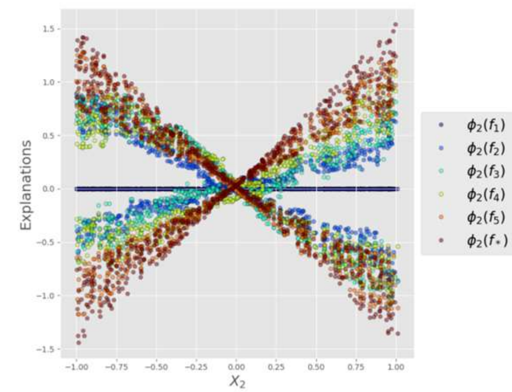
$$X_1 = Z + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0, 0.05),$$

$$X_2 = \sqrt{2}\sin(Z(\pi/4)) + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, 0.05),$$

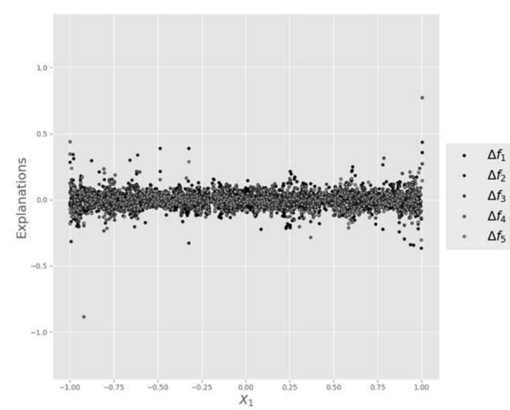
$$X_3 \sim Unif\big([-1, -0.5] \cup [0.5, 1]\big).$$



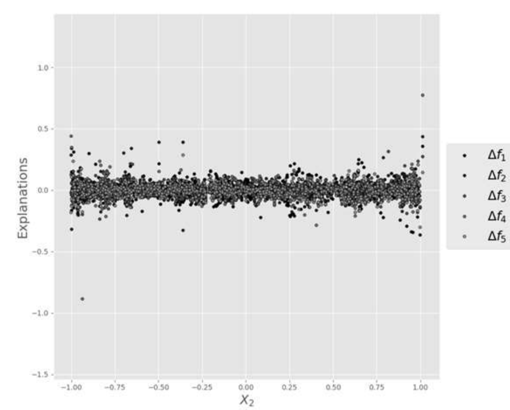
(a) Explanations φ_1 vs X_1.



(b) Explanations φ_2 vs X_2.



(c) Differences of predictions vs X_1.



(d) Differences of predictions vs X_2.

Challenges

- Features are almost never independent.
- Conditional feature attributions (based on v^{CE}) often differ from the marginal ones (based on v^{ME}).

Questions

- When marginal explanations are stable in $L^2(P_X)$? How to mitigate instabilities (if any)?
- Can the two type of explanations be reunited?

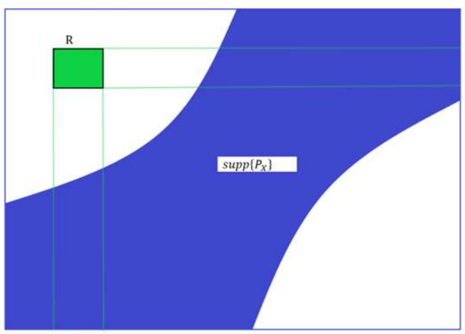
To answer these questions, it is necessary to consider the relationship between

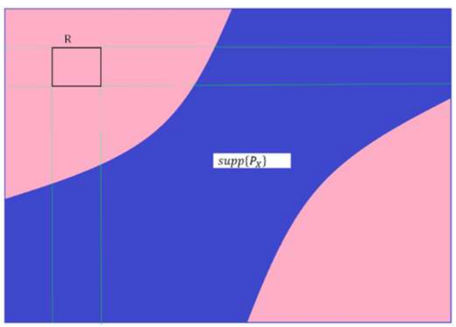
$\tilde{P}_X = \frac{1}{2^n} \sum_{S \subseteq N} P_{X_S} \otimes P_{X_{-S}}$ and P_X. Note: $\tilde{P}_X = P_X$ only when features are independent.

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

- The marginal game (v^{ME}, H_X) on $H_X = \left(L^2\left(\tilde{P}_X\right)/H_X^0, \|\cdot\|_{L^2(P_X)}\right)$ is well-defined if and only if $\tilde{P}_X \ll P_X$.
- If $\tilde{P}_X \ll P_X$, $H_X = \left(L^2\left(\tilde{P}_X\right), \|\cdot\|_{L^2(P_X)}\right)$
- 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 strength 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 (bounded) [AM, Kotsiopoulos, Filom, Ravi Kannan (2023,revised)]

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

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

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

Theorem (unbounded) [AM, Kotsiopoulos, Filom, Ravi Kannan (2024,revised)]

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

- *Let $\varnothing \neq S \subset N$. Suppose that either*

$$\sup\left\{\frac{[P_{X_S} \otimes P_{X_{-S}}](A \times B)}{P_X(A \times B)} \cdot P_{X_{-S}}(B),\ A \in \mathcal{B}(\mathbb{R}^{|S|}),\ B \in \mathcal{B}(\mathbb{R}^{|-S|}), P_X(A \times B) > 0\right\} = \infty. \tag{UG1}$$

or the non-negative, well-defined Borel function

$$\rho(x_S) := \int r_S^{1/2}(x_S, x_{-S}) P_{X_{-S}}(dx_{-S}) \tag{UG2}$$

with values in $\mathbb{R} \cup \{\infty\}$ is not P_{X_S}-essentially bounded.

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

Theorem (unbounded) [AM, Kotsiopoulos, Filom, Ravi Kannan (2024,revised)]

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

- *Let* $\varnothing \neq S \subset N$. *Suppose that either*

$$\sup\left\{\frac{[P_{X_S} \otimes P_{X_{-S}}](A \times B)}{P_X(A \times B)} \cdot P_{X_{-S}}(B),\ A \in \mathcal{B}(\mathbb{R}^{|S|}),\ B \in \mathcal{B}(\mathbb{R}^{|-S|}), P_X(A \times B) > 0\right\} = \infty. \qquad \text{(UG1)}$$

or the non-negative, well-defined Borel function

$$\rho(x_S) := \int r_S^{1/2}(x_S, x_{-S}) P_{X_{-S}}(dx_{-S}) \qquad \text{(UG2)}$$

with values in $\mathbb{R} \cup \{\infty\}$ *is not* P_{X_S}*-essentially bounded.*

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

- *Suppose there exist two distinct indices* $i, j \in N$ *such that*

$$\sup\left\{\frac{[P_{X_i} \otimes P_{X_j}](A \times B)}{P_{(X_i,X_j)}(A \times B)} \cdot P_{X_j}(B),\ \ A, B \in \mathcal{B}(\mathbb{R}), P_{(X_i,X_j)}(A \times B) > 0\right\} = \infty.$$

Suppose that the weights in (3.7) *satisfy the non-negativity condition (NN) and*

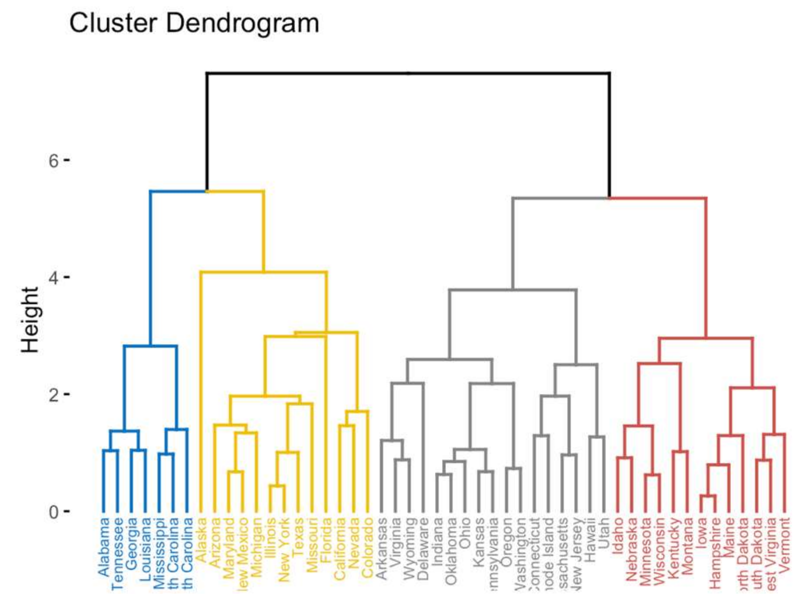
$$\sum_{S \subseteq N \setminus \{i,j\}} w(S, n) > 0.$$

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



Mitigation. Grouping features as a stabilization mechanism.

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

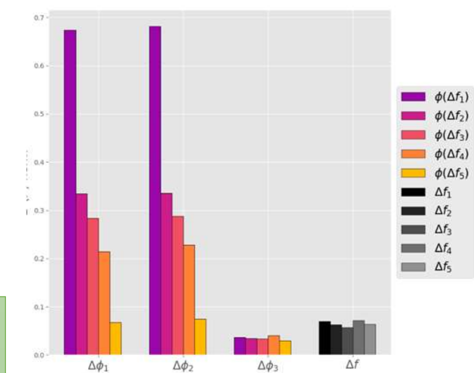
Mitigation. Grouping features as a stabilization mechanism.

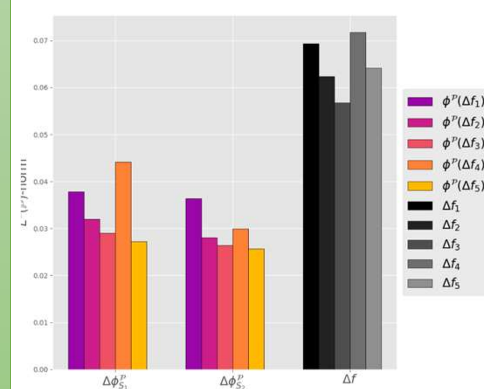
Quotient game explainers

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





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

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

$$h_j\big[M, v^{CE,\mathcal{P}}(f)\big] = h_j\big[M, v^{ME,\mathcal{P}}(f)\big] \text{ and hence continuous.}$$

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

$$H_X \ni f \to h_j\big[M, v^{ME,\mathcal{P}}(f)\big] \in L^2(P_X) \text{ is bounded with the bound} \sim \max_{A\subset M} \|r_A - 1\|.$$

References

- J.F. Banzhaf, Weighted voting doesn't work: a mathematical analysis. Rutgers Law Review 19, 317-343, (1965).
- H. Chen, J. Danizek, S. Lundberg, S.-I. Lee, True to the Model or True to the Data. arXiv preprint arXiv:2006.1623v1, (2020)
- J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of Statistics, Vol. 29, No. 5, 1189-1232,(2001).
- A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics, 24:1, 44-65 (2015).
- P. Hall, B. Cox, S. Dickerson, A. Ravi Kannan, R. Kulkarni, N. Schmidt, A United States Fair Lending Perspective on Machine Learning. Front. Artif. Intell. 4:695301. doi: 10.3389/frai.2021.695301 (2021).
- P. Hall, N. Gill, An Introduction to Machine Learning Interpretability, O'Reilly. (2018).
- T. Hastie, R. Tibshirani and J. Friedman The Elements of Statistical Learning, 2-nd ed., Springer series in Statistics (2016).
- S.M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, 31st Conference on Neural Information Processing Systems, (2017).
- Y. Kamijo, A two-step Shapley value in a cooperative game with a coalition structure. International Game Theory Review 11 (2), 207–214, (2009).
- 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)
- A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Stability theory of game-theoretic group feature explanations for machine learning models, *arXiv preprint,* arXiv:2102.10878v5 (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).
- G. Owen, Modification of the Banzhaf-Coleman index for games with apriory unions. In: Power, Voting and Voting Power (M.J. Holler, ed.), Physica-Verlag, 232-238. and Game Theory (R. Henn and O. Moeschlin, eds.), Springer, 76-88 (1982).
- M.T. Ribeiro, S. Singh and C. Guestrin, “Why should I trust you?” Explaining the predictions of any classifier, 22nd Conference on Knowledge Discovery and Data Mining, (2016).
- E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst., 41, 3, 647-665, (2014).
- J. Wang, J. Wiens, S. Lundberg Shapley Flow: A Graph-based Approach to Interpreting Model Predictions arXiv preprint arXiv:2010.14592, (2020).