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

Alexey Miroshnikov
Konstandinos Kotsiopoulos
Khashayar Filom
Arjun Ravi Kannan

Emerging Capabilities & Data Science Research Group, Discover Financial Services

North Carolina State University, Mathematics Department, March 2024

Motivation

- Contemporary predictive ML models are complex:
  - Neural Networks (NN)
  - Gradient Boosting Machines (GBM)
  - Semi-supervised methods
- Interpretability is crucial for business adoption, model documentation, regulatory oversight, and human acceptance and trust:
  - Banking
  - Insurance
  - Healthcare
- Accuracy may come at the expense of interpretability [P. Hall, 2018]:
  - Linear models are easy to interpret, $Y = a_1X_1 + \cdots a_nX_n$
  - Nonlinear models (GBM, NN) are difficult to interpret.

Motivation

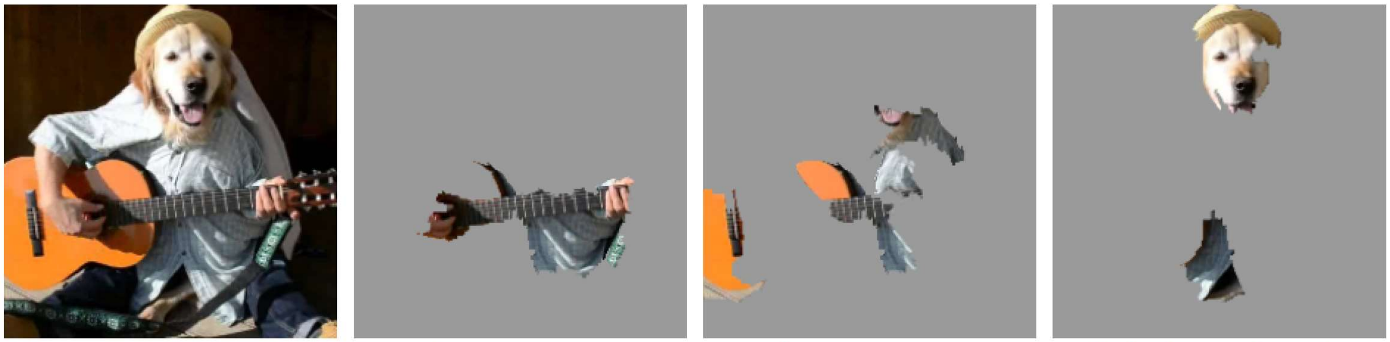
Regulatory requirements

- 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.
- Determining the factor contributing the most to an outcome of a model may be done via individualized feature attributions.

Common approaches:

- Self-interpretable models
- Post-hoc model explanations

Image classification [from 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.

Individualized explanations

Setup

- $(\Omega, \mathcal{F}, \mathbb{P})$ common probability space
- Distribution: random pair (X, Y), where $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ are features, $Y \in \mathbb{R}$ is response variable.
- P_X a pushforward probability measure, $P_X(A) = \mathbb{P}(X \in A)$, $\mathcal{B}(\mathbb{R}^n)$.
- ML model: $f(x) = \widehat{\mathbb{E}}[Y|X = x]$.

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, it can be viewed as a map

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

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

Example

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

Games and game values

A cooperative game (N, v)

- Set of players $N = \{1,2, \dots, n\}$
- Utility v
  - $v(\emptyset) = 0$
  - $v(N)$ is payoff of the game
  - $v(S)$ is the worth of the coalition $S \subseteq N$

Game value

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

- (LN) h is linear if $h[N, v + w] = h[N, v] + h[N, w]$.
- (EF) h is efficient if $\sum_i h_i[N, v] = v(N)$.
- (SM) h is symmetric if it is invariant with respect to player permutations.
- (NP): null-player property: if $i \in N$ is null player (i.e. $v(S \cup i) = v(S), \forall S$) $\Rightarrow$ $h_i[N, v] = 0$.

Linear game value

Shapley value [Shapley, 1953]

$$\varphi_i[v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \big(v(S \cup i) - v(S)\big)$$

- φ is linear, efficient, symmetric, null-player property.
- φ is the unique game value that satisfies (LN), (EF), (SM), [Shapley, 1953].

Generic linear, symmetric game value in the marginalist form with (NP)

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

h satisfies (NP).

remark: The Shapley value assumes that every player is equally likely to join any coalition of the same size and that all coalitions of a given size are equally likely.

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 (marginal and conditional 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

Definition (marginal and conditional explanations)

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 …

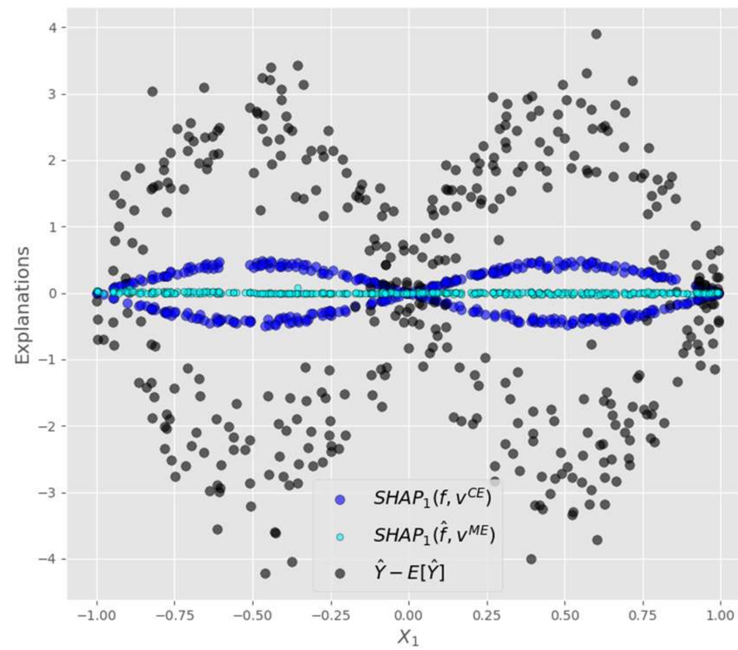
Marginal game

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

Random games

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

Linearity

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

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

Conditional operator

Let $X=(X_1,..X_n)$ be a random vector, $h[N,v]$ be a linear game value.

For $i \in N$ define a map

$$\bar{\mathcal{E}}_i^{CE}: L^2(\mathbb{R}^n, P_X) \mapsto L^2(\Omega, \mathbb{P}) \quad \text{by} \quad \bar{\mathcal{E}}_i^{CE}[f] := h_i[N, v^{CE}(\cdot; X, f)] = h_i^{CE}(X).$$

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

$$\left\|\bar{\mathcal{E}}_i^{CE}[f_1] - \bar{\mathcal{E}}_i^{CE}[f_2]\right\|_{L^2(\mathbb{P})} \le C_i(h)\|f_1 - f_2\|_{L^2(P_X)} = C_i(h)\sqrt{\mathbb{E}\left[\left[(f_1(X) - f_2(X)\right)^2\right]}$$

- If $Y = f(X) + \epsilon$, then

$$h_i[N, \mathbb{E}[Y|X_S]\,] = \bar{\mathcal{E}}_i{}^{CE}[f] + O(\epsilon) \ \text{ in } L^2(\mathbb{P}) \quad \text{(data consistency)}$$

Consequences

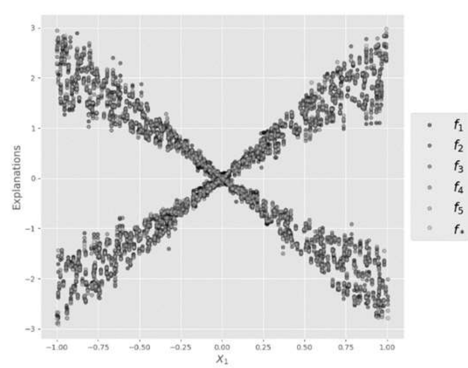
- $f_1(X) \approx f_2(X)$ in $L^2(\mathbb{P}) \Rightarrow h[v^{CE}(f_1)] \approx h[v^{CE}(f_2)]$ in $L^2(\mathbb{P})$.
- Functional representation of f plays no role for explanations, that is, the Rashomon effect does not take place.

Motivational example for marginal explanations "instabilities" in $L^2(P_X)$-norm

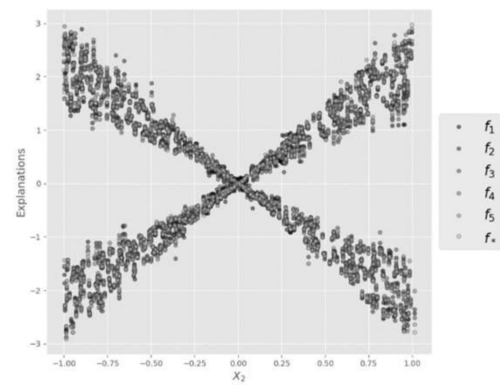
Synthetic model

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

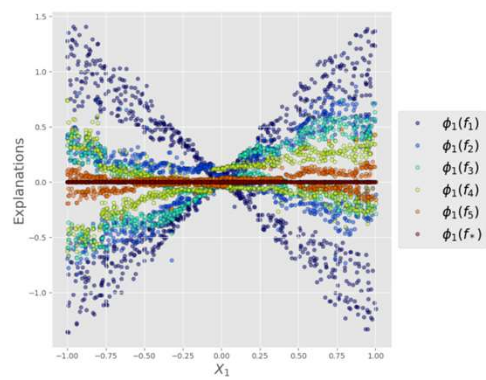
$$
Y = f_*(X_1, X_2, X_3) + \epsilon_3 = 3X_2X_3 + \epsilon_3
$$



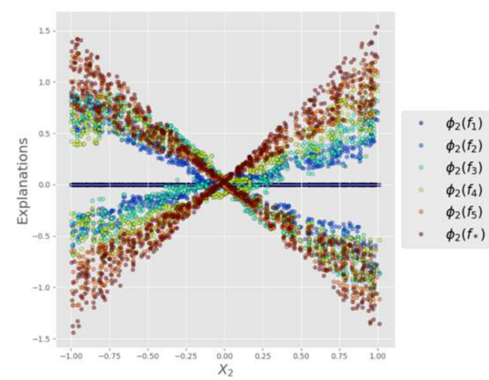
(a) Predictions vs X_1.



(b) Predictions vs X_2.



(a) Explanations φ_1 vs X_1.



(b) Explanations φ_2 vs X_2.

Question: What is a natural domain for marginal explanations to be a well-defined Operator?

Marginal Operator

- $\tilde{P}_X = \frac{1}{2^n} \sum_{S \subseteq N} P_{X_S} \otimes P_{X_{-S}}$
- $\bar{\mathcal{E}}_i^{ME} : L^2(\tilde{P}_X) \to L^2(\mathbb{P})$ defined by

$$\bar{\mathcal{E}}_i^{ME}[f; h, X] := h[v^{ME}(\cdot\,; X, f)]$$

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

- $\bar{\mathcal{E}}_i^{ME} : L^2(\tilde{P}_X) \to L^2(\mathbb{P})$ is well-defined
- $\left\| \bar{\mathcal{E}}_i^{ME}[f_1] - \bar{\mathcal{E}}_i^{ME}[f_2] \right\|_{L^2(\mathbb{P})} \leq C_i(h) \| f_1 - f_2 \|_{L^2(\tilde{P}_X)}$
- $\left(\bar{\mathcal{E}}_i^{ME}, L^2(\tilde{P}_X) \right)$ is bounded and hence continuous

Note: $L^2(\tilde{P}_X)$ in general cannot be embedded in $L^2(P_X)$.

Central questions regarding the marginal operator

- Can the marginal operator be well-defined on a space equipped with $L^2(P_X)$-norm?
- If yes, when is it bounded and when unbounded?

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

Independent features

If $P_X = \bigotimes P_{X_i}$, that is, $X = (X_1, \dots, X_n)$ are independent, then

- $\tilde{P}_X = P_X$
- $v^{ME} = v^{CE}$

$\Rightarrow$ $h[N, v^{CE}] = h[N, v^{ME}] \Rightarrow$ marginal operator is bounded (continuous) in $L^2(P_X)$.

Dependent features

If features are dependent then

- $\tilde{P}_X \neq P_X$ with $P_X \ll \tilde{P}_X$

$\Rightarrow$ Marginal explanations will depend on the representation of $f(x)$.

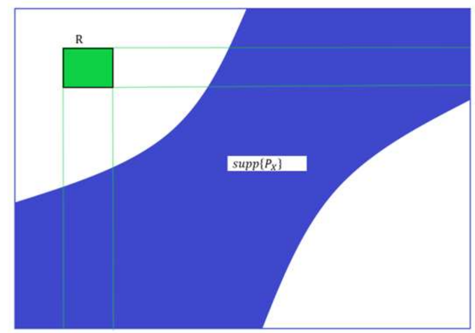
Lemma [AM, Kotsiopoulos, Filom, Ravi Kannan (2022)]

Suppose $\tilde{P}_X$ is not absolutely continuous w.r.t. P_X.

- The identity map $I: L^2(\tilde{P}_X) \to L^2(P_X)$ is not one-to-one.
- The identity map $I:\ L^2(\tilde{P}_X)/H_X^0 \to L^2(P_X)$ is one-to-one.

Then $H_X = \left(L^2(\tilde{P}_X)/H_X^0, \|\cdot\|_{L^2(P_X)}\right)$ we have

$$f \in H_X \to \{v^{ME}(S; X, f)\}_{S\subseteq N} \in \left(L^2(\mathbb{P})\right)^n \text{ is an ill-posed operator.}$$

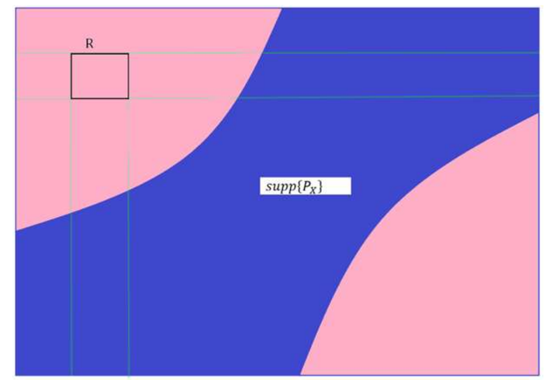


Proof: $\exists R \in \mathcal{B}(\mathbb{R}^n)$ s.t. $[1_R]_{H_X} = [0]_{H_X}$ but $v^{ME}(S; X, 1_R) \neq v^{ME}(S; X, 0)$ for some $S \subseteq N$.

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

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

- The identity map $I: L^2(\tilde{P}_X) \to L^2(P_X)$ is one-to-one.
- $H_X = \left(L^2(\tilde{P}_X), \|\cdot\|_{L^2(P_X)}\right)$ is well-defined.
- $f \in H_X \to \{v^{ME}(S; X, f)\}_{S \subseteq N} \in \left(L^2(\mathbb{P})\right)^n$ is a well-defined operator.
- $f \in H_X \to h_i[N, v^{ME}(\cdot; X, f)], \in L^2(\mathbb{P}), i \in N$ is a well-defined operator.



Question: Is there any relationship between boundedness and dependencies?

- 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:

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

It turns out the Radon-Nikodym derivative can shed light on the boundedness/continuity

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

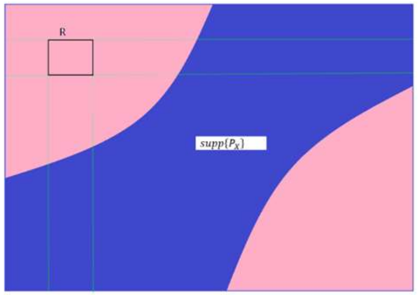
If $\tilde{P}_X \ll P_X$ and $r_X \in L^\infty(P_X)$. Then

$(\bar{\mathcal{E}}^{ME}, H_X)$ is a **well-defined bounded linear** operator satisfying

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

Proof: By definition of RN derivative and v^{ME}.

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



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

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

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

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.*

Grouping features as a stabilization mechanism

1. The choice between the two games is application specific

- In applications where it is crucial to understand the true scientific reason behind observed data v^{CE} might be preferable
- In other applications, where the model is required to be explained, the game v^{ME} should be used

2. Complexity and stability

- Marginal explanations are consistent with the model; unstable with respect to model perturbation in $L^2(P_X)$. Expensive.
- Computing conditional explanations is infeasible; stable with respect to model perturbation in $L^2(P_X)$.

1 & 2 motivate us to design methods that employ *grouping by dependencies*

- Unify two types of explanations (to achieve stability of marginal explanations)
- Reduce complexity of computations
- Explanations are split under dependencies (grouping allows to compute an "explanation of information")

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

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

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

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

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

- 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

$$H_X \ni f \to h_j[M, v^{ME,\mathcal{P}}(f)] \text{ is bounded.}$$

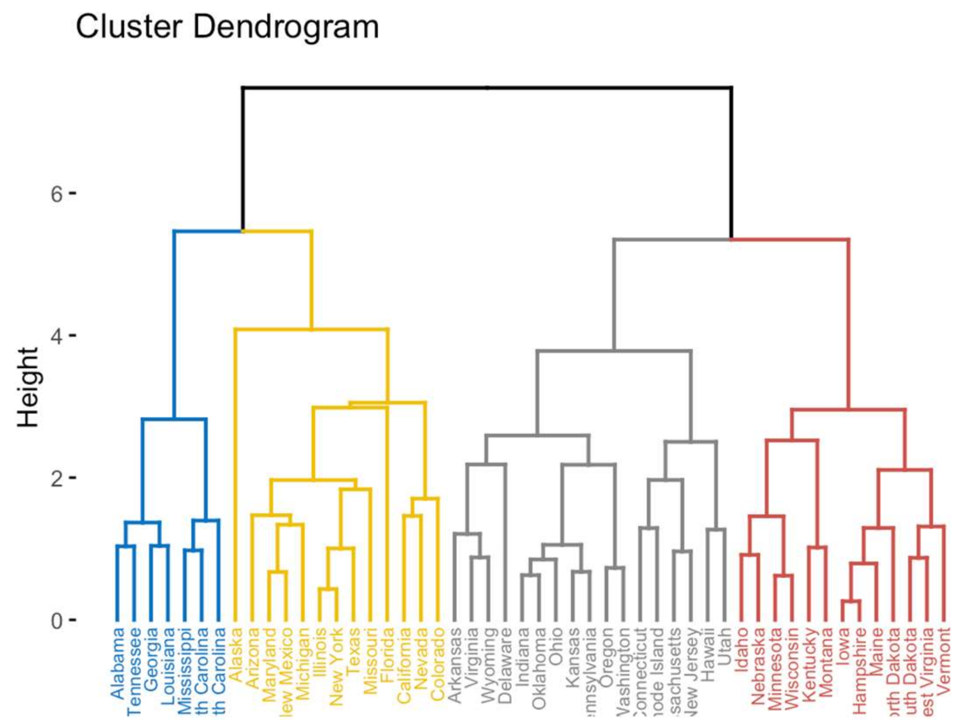
Variable hierarchical clustering

Inputs

- features $X_1, X_2, \dots, X_n$
- variable dissimilarity $d_{var}(X_i, X_j)$
- intergroup dissimilarity $\mathrm{d}_{\mathrm{group}}(S_k, S_m)$
- energy functional for minimization W

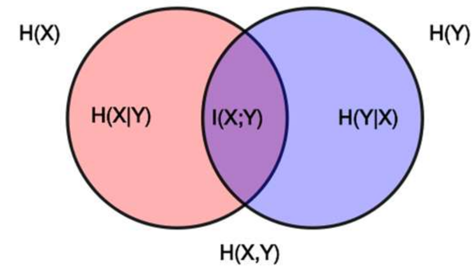
Output

- dendrogram
- height of each node reflects the level of dissimilarity



Clustering based on MIC

1. Mutual information [Shannon 1948]



- Measure of the mutual dependence between two variables:

$$I(X,Y) = D_{KL}\left(P_{(X,Y)}|P_X \otimes P_Y\right) \in [0,\infty]$$

2. Maximal Information coefficient, MIC_* [Reshef et al, 2011, 2016]

- $MIC_*(X,Y)$= Regularized mutual information $\in [0,1]$
- Equitable: $MIC_*(X,Y) = MIC_*\left(g(X), g(Y)\right)$
- Transitive: $MIC_*(X,Y) \approx\ MIC_*(Z,W)\ \Rightarrow\ MIC_*(X+\epsilon_1\,, Y+\epsilon_2) \approx\ MIC_*(Z+\epsilon_1, W+\epsilon_2)$
- Fast algorithm $O(\#samples)$

3. For variable clustering $d_{var}\left(X_i, X_j\right) = 1 - MIC_*(X_i, X_j)$. Group dissimilarity information theoretic.

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).