Notes on game theory with finitely many players with applications to ML explainability

Alexey Miroshnikov

ML group reading seminar, LANL, November 2021

Introduction

- Contemporary predictive ML models are complex:
  - Neural networks
  - Gradient Boosting Machines
  - Random Forests
  - Semi-supervised methods

- Accuracy versus interpretability [P. Hall, 2018]. Accuracy comes at the expense of interpretability:
  - Linear models is easy to interpret, $Y = a_1 X_1 + \cdots a_n X_n$
  - Nonlinear models (GBM, RF) are difficult to interpret.

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

- Banking [P. Hall et al. 2020]
- Insurance
- Healthcare

Some approaches:

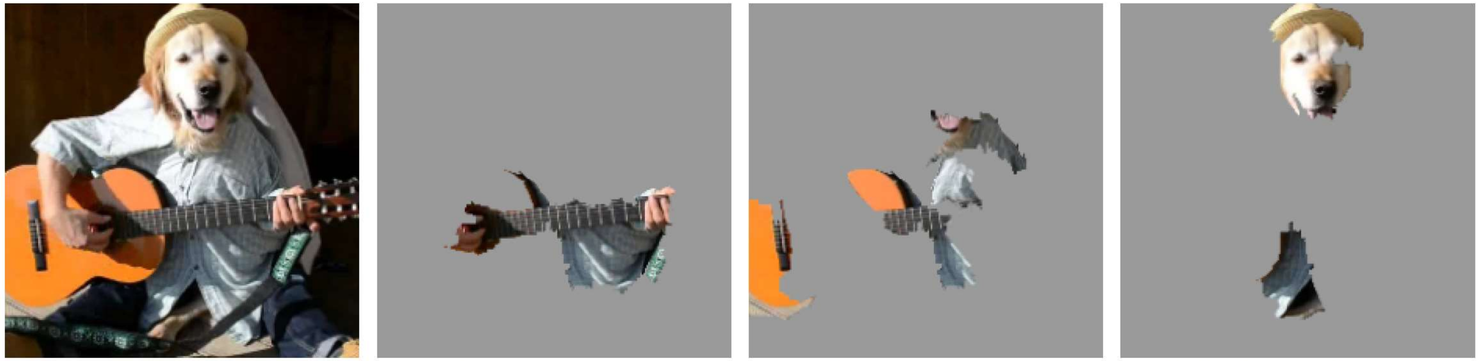
- Self-explainable models
- Post-hoc explanations

Explainers

- Data: (X, Y), predictors $X \in \mathbb{R}^n$, response $Y \in \mathbb{R}$
- Model: $f(x) = \widehat{\mathbb{E}}\,[Y|X = x]$
- Model explainer quantifies the contribution of predictor(s) to the value $f(x)$, $x \in supp(P_X)$,

$$E[x; f] = [E_1(x; f), E_2(x; f), \ldots E_2(x; f)]$$

Motivational examples [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.

Partial Dependence Function (PDP) [Friedman, 2001]

Given a sample $\boldsymbol{x} = (\boldsymbol{x_i}, \boldsymbol{x_{-i}})$, $-i = \{1,2, \ldots, n\} \backslash \{i\}$:

$$x_i \rightarrow PDP_i(x_i; f) = \mathbb{E}[f(x_i, X_{-i})] \approx \frac{1}{N} \sum_{j=1}^{N} f(x_i, X_{-i}^{(j)})$$

Example

$$f(X) = f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n)$$

$$PDP_i(x_i; f) - \mathbb{E}[f(X)] = f_i(x_i) - \mathbb{E}[f_i(X_i)]$$

Partial Dependence Function (PDP) [Friedman, 2001]

Given a sample $\boldsymbol{x} = (\boldsymbol{x_i}, \boldsymbol{x_{-i}})$, $-i = \{1,2,\dots,n\}\backslash\{i\}$:

$$x_i \to PDP_i(x_i; f) = \mathbb{E}[f(x_i, X_{-i})] \approx \frac{1}{N}\sum_{j=1}^{N} f(x_i, X_{-i}^{(j)})$$

Example

$$f(X) = f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n)$$

$$PDP_i(x_i; f) - \mathbb{E}[f(X)] = f_i(x_i) - \mathbb{E}[f_i(X_i)]$$

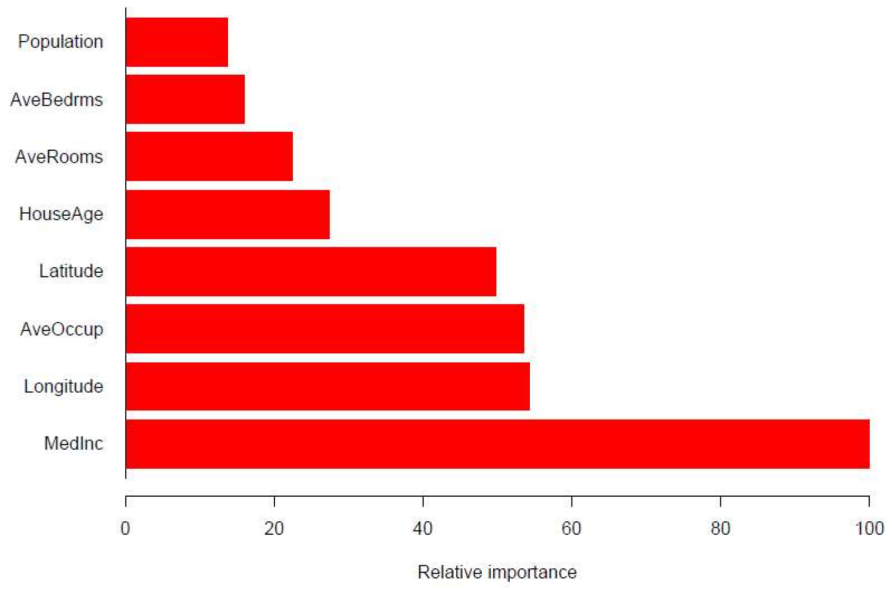
why not use conditional expectation?

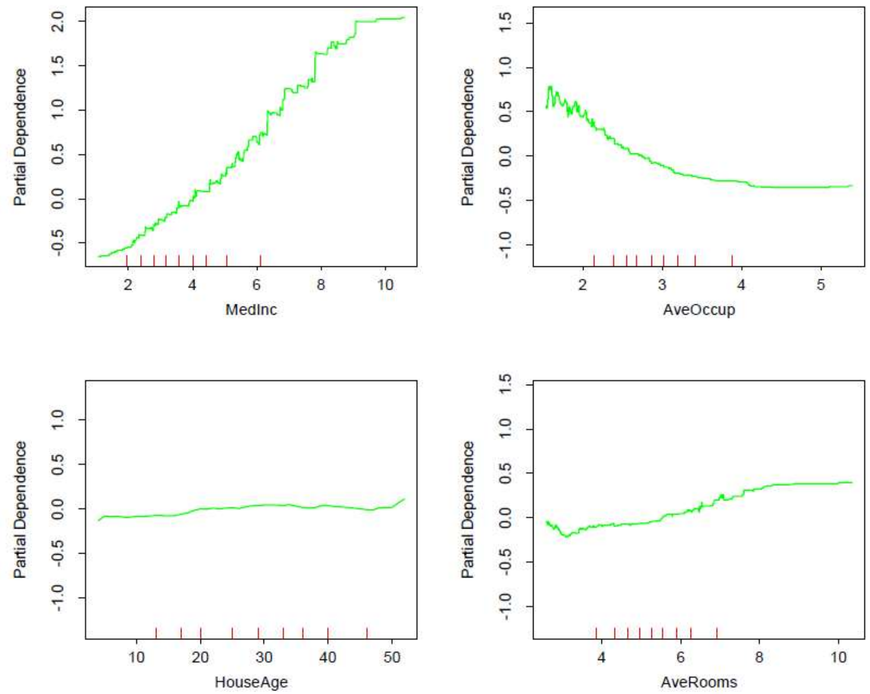
$\mathbb{E}[f(X)|X = x_i] \approx \mathbb{E}[Y|X = x_i]$ explains the data (response) not the model:

$$f(x) = x_1 + x_2, Y = f(X) + \epsilon, \ X_i = Z + \epsilon_i, \mathbb{E}[X_i] = 0 \quad \Rightarrow \ \mathbb{E}[f(x_i, X_{-i})] = x_i, \mathbb{E}[f(X)|X = x_i] \approx 2x_i$$

Example from Hastie et al [Elements of Machine Learning, p. 373,374]

Analysis of the house value versus other predictors using GBM:

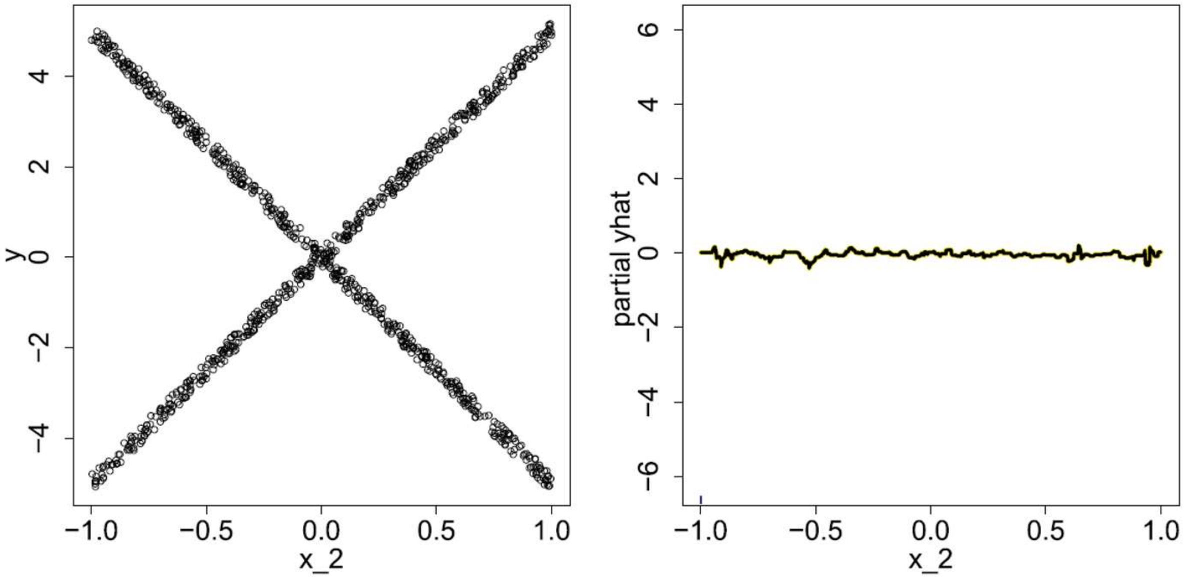




Interactions issues of PDPs [example from Goldstein et al, 2015]

$X_1, X_2, X_3 \sim Unif[-1,1]$

$Y = f(X) + \epsilon = 0.2X_1 - 5X_2 + 10X_2 1_{\{X_3 \geq 0\}}$



(a) Scatterplot of Y versus X_2 (b) PDP

Games and game values

Game

n players, $N = \{1,2, \dots, n\}$

- Game is a super-additive function $v(S)$, $S \subset N$, $v(\emptyset) = 0$
- $v(N)$ is payoff of the game (think of profit)
- $v(S)$ is the worth of the coalition S

Game value

Map $v \rightarrow \mathrm{h}[N, v] = (h_1[N, v], h_2[N, v], \dots, h_n[N, v])$

Example 1: Shapley value

$\varphi_i[v] = \sum_{S \subset N} \frac{(s-1)!(n-s)!}{n!} [v(S) - v(S \backslash \{i\})]$, [Shapley, 1953]

- (LN) φ is linear: $\varphi[v + w] = \varphi[v] + \varphi[w]$
- (EF) φ is efficient: $\sum_i \varphi_i[v] = v(N)$
- (SM) φ is symmetric (abstract games)

$\Rightarrow$ (NP) null player property: null player $i \in N$ of v $\Rightarrow \varphi_i[v] = 0$.

remark: φ is a unique game value that satisfies (LN), (EF), (SM), [Shapley, 1953].

Example 1: Shapley value

$$\varphi_i[v] = \sum_{S \subset N} \frac{(s-1)!(n-s)!}{n!} [v(S) - v(S \setminus \{i\})],$$ [Shapley, 1953]

- (L) φ is linear: $\varphi[v + w] = \varphi[v] + \varphi[w]$
- (E) φ is efficient: $\sum_i \varphi_i[v] = v(N)$
- (S) φ is symmetric (abstract games)

$\Rightarrow$ (N) null player property: null player $i \in N$ of v $\Rightarrow \varphi_i[v] = 0$.

remark: φ is a unique game value that satisfies (L), (E), (S), [Shapley, 1953].

Example 2: Banzhaf value

$$Bz_i[v] = \sum_{S \subset N} \frac{1}{2^n} [v(S) - v(S \setminus \{i\})],$$ [Banzhaf, 1965]

- BZ satisfies (L), (S), total power property.

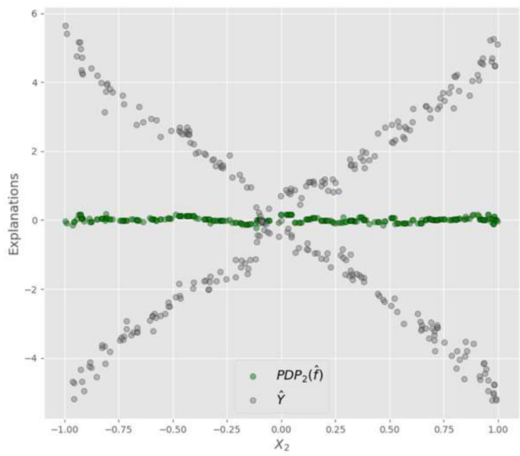
remark: The Shapley value that assumes that every player is equally likely to join any coalition of the same size and that all coalitions of a given size are equally likely. The Banzhaf value assumes that every player is equally likely to enter any coalition.

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

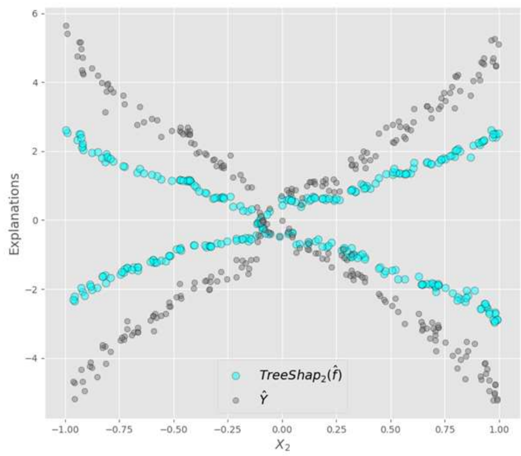
Given (X, f) and $S \subset N = \{1,2, \ldots n\}$

- $v^{CE}(S; x, f) = \mathbb{E}[f(X_S, X_{-S}) | X_S = x_s]$, conditional game
- $v^{ME}(S; x, f) = \mathbb{E}[f(x_S, X_{-S})]$, marginal game
- $\varphi[v^{CE}]$, $\varphi[v^{ME}] \in \mathbb{R}^n$ are conditional and marginal explanations, respectively

Interactions are handled better with Shapley values (application to the example in [Goldstein et al, 2015])



PDP



marginal Shapley

Remarks: Some facts for $v \in \{v^{ME}, v^{CE}\}$

- v^{ME}, v^{CE} are not cooperative games because $v^{ME}(\emptyset) = v^{CE}(\emptyset) = \mathbb{E}[f(X)]$

$$\varphi_0 = \mathbb{E}[f(X)] \;\Rightarrow\; \sum_{i=0}^{n} \varphi_i[v^{ME}(x; f)] = \sum_{i=0}^{n} \varphi_i[v^{CE}(x; f)] = f(x)$$

- $v^{ME}(S; X, f), v^{CE}(S; X, f)$ are random games where randomness comes from predictors.

- Random games are linear with respect to models. The maps

$$f \to \varphi[v^{ME}(X; f)],\; f \to \varphi[v^{CE}(X; f)]$$

  are linear operators on appropriate domains [Miroshnikov et al 2021b].

- For additive models $f = \sum f_i(X_i)$

$$\varphi_i[v^{ME}(X; f)] = f_i(X_i) - \mathbb{E}[f_i(X)]$$

but in general PDPs and marginal Shapley value differ.

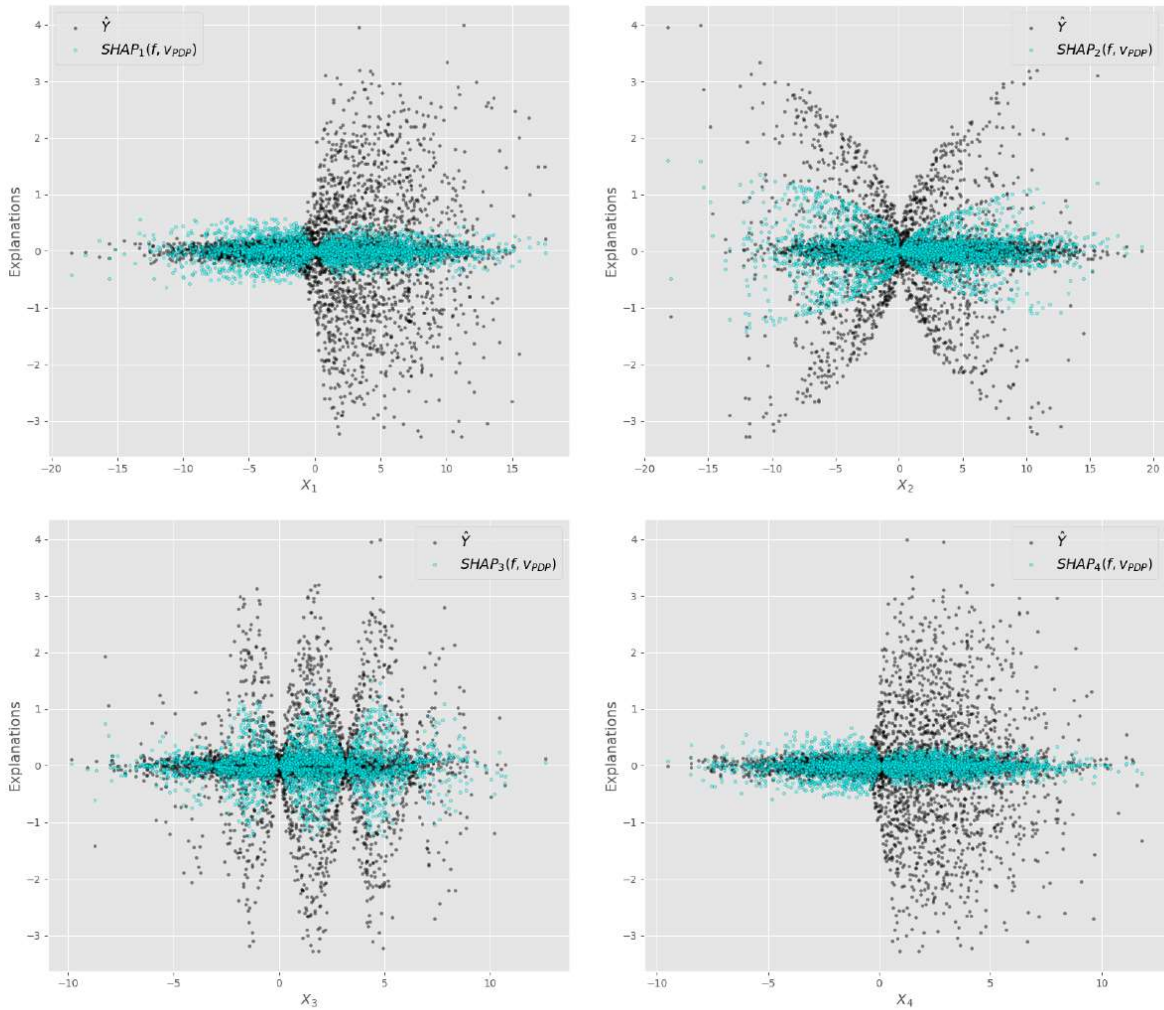
Example [Miroshnikov et al. 2021b]

$$Y = \prod_{i=1}^{4} f_i(X_i) + \epsilon = f(X) + \epsilon$$

$$f_1(X_1) = logistic(2X_1), \quad f_2(X) = \mathrm{sgn}(X_2)\sqrt{|X_2|},$$
$$f_3(X_3) = \sin(X_3), \quad f_4(X_4) = logistic(5X_4).$$

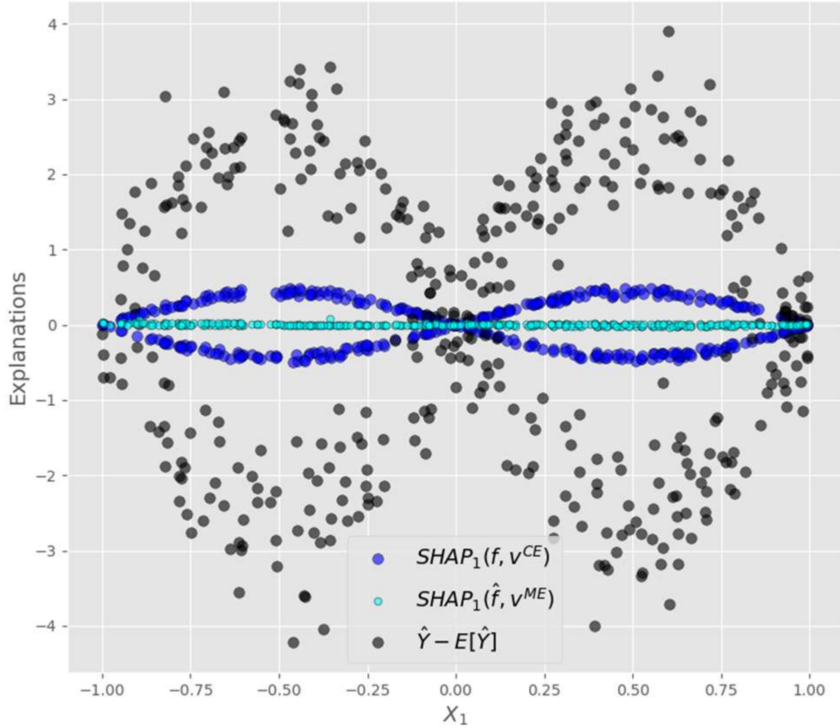
$$(X_1, X_2) \sim \mathcal{N}((1,1), \Sigma_1), \quad \Sigma_1 = \begin{bmatrix} 26 & -10 \\ -10 & 26 \end{bmatrix}$$

$$(X_3, X_4) \sim \mathcal{N}((1,1), \Sigma_2), \quad \Sigma_2 = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$



Question: What is the difference between marginal and conditional games?

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



Example from [Miroshnikov et al 2021b]

Conditional game

- v^{CE} *explores* the joint (X, Y)
- $\varphi[v^{CE}]$ are *consistent* with the data and $f(X)$
- Infeasible due to the curse of dimensionality [Hastie et al].

Marginal game

- v^{ME} *explores* the model $f(x)$
- $\varphi[v^{ME}]$ are *consistent* with the model $f(x)$
- Complexity $O(2^n)$

Game values with coalitional structure

In cooperative game theory with coalition structure, the objective is to compute the payoffs of players in a game where players form unions acting in agreement within the union: $\mathcal{P} = \{S_1, S_2, S_3, \dots, S_m\}, \cup S_j = \{1,2,\dots,n\}$

Coalitional value

Given a coalition structure $(N, v, \mathcal{P})$ where $\mathcal{P} = \{S_1, S_2, S_3, \dots, S_n\}$ the coalitional game is a map

$$g[N, v, \mathcal{P}] \in \mathbb{R}^n$$

- Owen values [Owen, 1977]
- Banzhaf-Owen values [Owen, 1982]
- Two-step Shapley [Kamijo, 2009]

remark: Shapley value is a trivial coalitional value

Game values with coalitional structure

In cooperative game theory with coalition structure, the objective is to compute the payoffs of players in a game where players form unions acting in agreement within the union: $\mathcal{P} = \{S_1, S_2, S_3, \dots, S_m\}, \cup S_j = \{1, 2, \dots, n\}$

Coalitional value

Given a coalition structure $(N, v, \mathcal{P})$, $\mathcal{P} = \{S_1, S_2, S_3, \dots, S_m\}$, the coalitional game is a map

$$g[N, v, \mathcal{P}] \in \mathbb{R}^n$$

- Owen values [Owen, 1977]
- Banzhaf-Owen values [Owen, 1982]
- Two-step Shapley [Kamijo, 2009]

remark: Shapley value is a trivial coalitional value

Game values with coalitional structure

In cooperative game theory with coalition structure, the objective is to compute the payoffs of players in a game where players form unions acting in agreement within the union: $\mathcal{P} = \{S_1, S_2, S_3, \ldots, S_m\}, \cup S_j = \{1,2,\ldots,n\}$

Coalitional value

Given a coalition structure $(N, v, \mathcal{P}),\ \mathcal{P} = \{S_1, S_2, S_3, \ldots, S_m\},$ the coalitional game is a map

$$g[N, v, \mathcal{P}] \in \mathbb{R}^n$$

- Owen values [Owen, 1977]
- Banzhaf-Owen values [Owen, 1982]
- Two-step Shapley [Kamijo, 2009]

$$Ow_i[N,v,\mathcal{P}] = \sum_{R\subset M\setminus\{j\}} \sum_{T\subset S_j\setminus\{i\}} \frac{r!(m-r-1)!}{m!} \frac{t!(s_j-t-1)!}{s_j!} (v(Q\cup T\cup\{i\}) - v(Q\cup T))$$

$$BzOw_i[N,v,\mathcal{P}] = \sum_{R\subset M\setminus\{j\}} \sum_{T\subset S_j\setminus\{i\}} \frac{1}{2^{m-1}} \frac{1}{2^{s_j-1}} (v(Q\cup T\cup\{i\}) - v(Q\cup T))$$

where $t = |T|$, $s_j = |S_j|$, $r = |R|$, and $S_j \in \mathcal{P}$, $Q = \cup_{r\in R} S_r$

$$TSh_i[N,v,\mathcal{P}] = \varphi_i[S_j, v] + \frac{1}{|S_j|}(\varphi_j[M, v^{\mathcal{P}}] - v(S_j)), \quad i \in S_j.$$

Quotient game: $v^{\mathcal{P}}(A) = v(\cup_{j \in A} S_j)$

Useful properties for ML explainability

- (2SF) 2-step formulation

Coalitional value can be obtained playing a quotient-like game and then a game inside the union.

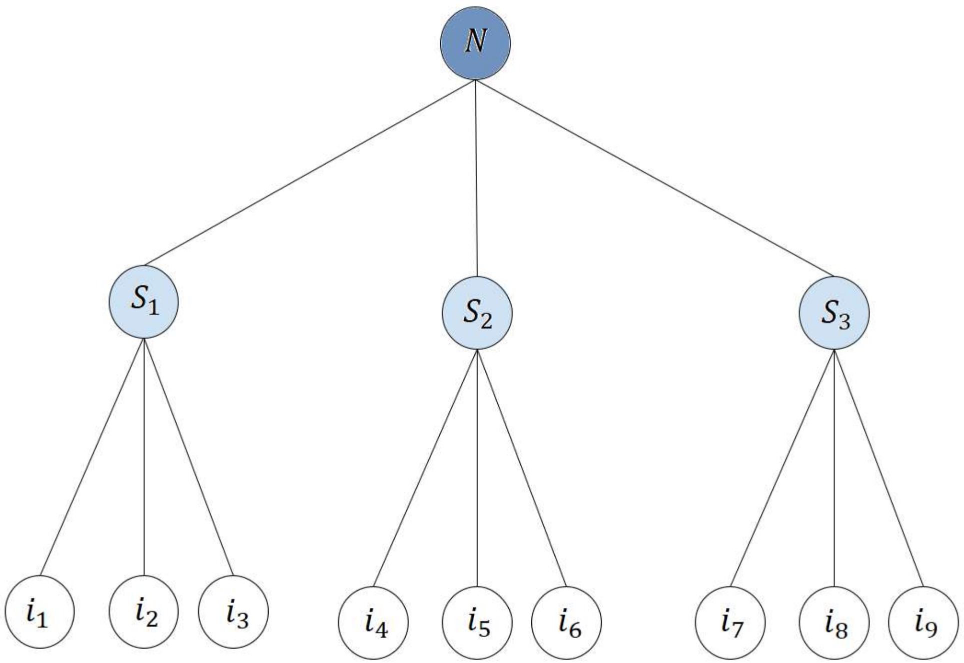
- (QP) Quotient game property

The sum of the payoffs of the players in a union equal to the payoff of the union in the quotient game.

Examples

- (EF) Shapley values
- (QP,2SF,EF) Owen values, Two-step Shapley
- (2SF) Banzhaf-Owen
- (QP,2SF) Symmetrical Banzhaf-Owen

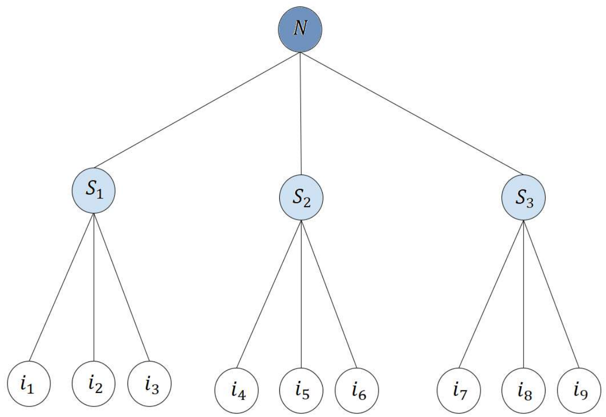
$\mathcal{P} = \{S_1, S_2, S_3\}, N = \{1,2, \ldots, 9\}$



Use of coalitional values in ML explainability

$$\mathcal{P} = \{S_1, S_2, S_3\},\ N = \{1,2,\dots,9\}$$

- Forming partitions by business/scientific/independence reasons
- Reduction in complexity $O\left(2^{|S_j|+|\mathcal{P}|}\right)$
- Generalization to partition trees and graphs [Wang et al. 2020]
- Fairness explainability [Miroshnikov et al. 2021a]
- Unifying marginal and conditional approaches [Aas et al. 2020, Miroshnikov et al. 2021b]



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, R. Franks, A. Ravi Kannan, Wasserstein-based fairness interpretability framework for machine learning models, *arXiv preprint* (2021a), arXiv:2011.03156.
- A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Mutual information-based group explainers with coalition structure for machine learning model explanations, *arXiv preprint* (2021b), arXiv:2102.10878.
- 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).