Joint Shapley values

Chris Harris[1] Richard Pymar[2] Colin Rowat[3]

[1]Visual Alpha, Tokyo
[2]Birkbeck, University of London
[3]University of Birmingham

23 July 2021
Vector Institute

https://github.com/harris-chris/joint-shapley-values

Motivation: joint feature importance

how do I attribute [importance in the presence of] correlations? (Bhatt et al., 2020)

- FAT (Bhatt et al., 2020)
  - "local explainability ...explain the model's behavior for a specific input"
  - "global explainability ...understand the high-level concepts and reasoning used by a model"
- linear models: e.g. $y = \hat{\beta}\boldsymbol{X} + \varepsilon$ ('local' and 'global')
- non-linear models?
- Shapley value: popular, 'model agnostic' approach
  - game theory: average value added by an individual, independently
  - XAI: average change in model's prediction due individual feature's value
- problem: feature dependence
  - collinear: individual insignificance (t-test), joint significance (F-test)
  - here: joint feature importance
- contributions
  - uniquely extend Shapley's value to joint feature importance
  - first index to do so

Games in characteristic function form

- most studied class of cooperative game (von Neumann and Morgenstern, 1944)
- set of agents, $i \in N = \{1, 2, \ldots n\}$
- value assigned to coalitions, $S \subseteq N$ by value function, $v : 2^N \to \mathbb{R}$
- Shapley (1953) value: what's a 'fair' split of $v(N)$?

$$\phi_i^S(v) \equiv \sum_{S \subseteq N \setminus \{i\}} p^i(S) \left[v(S \cup \{i\}) - v(S) \right]$$

where $p^i(S) = \frac{|S|!(n-|S|-1)!}{n!}$ randomises uniformly over singletons
  - sum of value i that adds to possible coalitions S
  - weighted by the (symmetric, independent) probability that i joins S
- arrival order: S arrives (any order), then i, then $N \setminus (S \cup i)$ (any order)
- ϕ^S uniquely satisfies 1 efficiency, 2 null/dummy, 3 symmetry/anonymity, 4 additivity/linearity axioms

Example (The $n = 3$ glove/market game e.g. Lucas (1971))

Let $i = 1$ have a left glove, and $i \in \{2, 3\}$ each have a right glove, with unit value arising from a pair, so that:

$$\begin{aligned} v(\varnothing) = v(1) = v(2) = v(3) = v(2,3) &= 0 \\ v(N) = v(1,2) = v(1,3) &= 1 \end{aligned}$$

The Shapley values are:

$$\phi_1^S(v) = \tfrac{2}{3} > \phi_2^S(v) = \phi_3^S(v) = \tfrac{1}{6}.$$

Consistent with intuition, the Shapley value privileges agent 1. Less intuitively, it gives no sign that value arises from particular pairs.

Extending Shapley from singletons to sets

- what's the average value a set adds?
- let agents arrive as sets, T (including singletons):

$$\phi_T^J(v) \equiv \sum_{S \subseteq N \setminus T} p^T(S)\left[v(S \cup T) - v(S)\right]$$

- extend $p^i(S)$ to $p^T(S)$ by randomising uniformly over sets
- S arrives (any weak order), then T together, then $N \setminus (S \cup T)$ (any weak order)
- will add an order of explanation, k, to efficiency axiom
  - controls computational costs, $\mathcal{O}(3^n \wedge (2^n n^k))$
  - introduced by Dhamdhere, Agarwal, and Sundararajan (2020)
  - $k = 1$ reduces to original Shapley
- two branches of game theory literature:
  1. fix coalitions *a priori* (Owen, 1977)
  2. decompose sets recursively to singletons (Grabisch and Roubens, 1999)

Extending Shapley's axioms: the easy part

(JEF) joint efficiency: total worth partitions among sets

$$\sum_{\substack{T \subseteq N \\ |T| \leq k}} \phi_T^J(v) = v(N)$$

(JNU) joint null: sets adding no worth get no value

$$v(S \cup T) = v(S) \forall S \subseteq N \setminus T \Rightarrow \phi_T^J(v) = 0$$

(n.b. interaction indices build on singletons)

(JLI) joint linearity:

$$\phi_T^J(\alpha u + \beta v) = \alpha \phi_T^J(u) + \beta \phi_T^J(v)$$

for any non-negative scalars α and β.
(helps extend proofs from particular games to all games)

Extending Shapley's axioms: the harder part

Shapley's symmetry: any arrival order of agents is equally likely

(ANO) anonymity for all permutations, σ, on N,

$$\psi_i(v) = \psi_{\sigma(i)}(\sigma v)$$

for all $i \in N$, all games v and all $\psi : N \mapsto \mathbb{R}$.

(SYM) symmetry

$$v(S \cup i) = v(S \cup j) \forall S \subseteq N \setminus \{i, j\} \Rightarrow \phi_i^S(v) = \phi_j^S(v)$$

- ANO $\Rightarrow$ SYM (Malawski, 2020)
- Shapley's value uniquely solves EFF, DUM, LIN and either ANO/SYM
- what's the right way to generalise these? (There are lots of wrong ways)

Joint anonymity and symmetry

(JAN) joint anonymity: for all permutations, σ, on N,

$$\psi_T(v) = \psi_{\sigma(T)}(\sigma v)$$

for all $T \subseteq N$, all games v and all $\psi : 2^N \mapsto \mathbb{R}$

(JSY) joint symmetry: if two sets add equal worth to sets that they can both join and add no worth to all other coalitions, they receive an equal value:

$$\begin{aligned}
&v(S \cup T) = v(S \cup T') \text{ for all } S \subseteq N \setminus (T \cup T') \\
&v(S \cup T) = v(S) \text{ for all } S \subseteq N \setminus T \text{ s.t. } S \cap T' \neq \varnothing \\
&v(S \cup T') = v(S) \text{ for all } S \subseteq N \setminus T' \text{ s.t. } S \cap T \neq \varnothing \\
&\Rightarrow \phi_T(v) = \phi_{T'}(v)
\end{aligned}$$

Joint Shapley values are unique for each k

Theorem

For each order of explanation $k \in \{1, \dots, n\}$, there is a unique ϕ^J which satisfies axioms JLI, JNU, JEF, JAN and JSY:

$$\phi^J_T(v) = \sum_{S \subseteq N \setminus T} q_{|S|}[v(S \cup T) - v(S)]$$

for each $\varnothing \neq T \subseteq N$ with $|T| \leq k$, where $(q_0, \dots, q_{n-1})$ uniquely solves

$$q_0 = \frac{1}{\sum_{i=1}^{k} \binom{n}{i}}, \quad q_r = \frac{\sum_{s=(r-k)\vee 0}^{r-1} \binom{r}{s} q_s}{\sum_{s=1}^{k \wedge (n-r)} \binom{n-r}{s}};$$

for all $r \in 1, \dots, n-1$.

Arrival orders, probabilities are independent of coalition size: $p^T(S) = q_{|S|}$

Proof sketch: if ϕ satisfies ...

1. JLI, JNU $\Rightarrow$ there exist constants $\{p^T(S)\}$ such that ...

$$\phi_T(v) = \sum_{S \subseteq N \setminus T} p^T(S)[v(S \cup T) - v(S)].$$

$\therefore$ no discrete derivatives, unlike Grabisch and Roubens (1999) etc.

2. JLI, JNU, JEF $\Leftrightarrow$ for each order of explanation k,

$$\delta_N(S) = \sum_{\substack{\emptyset \neq T \subseteq S: \\ |T| \leq k}} p^T(S \setminus T) - \sum_{\substack{\emptyset \neq T \subseteq N \setminus S: \\ |T| \leq k}} p^T(S),$$

for all $\emptyset \neq S \subseteq N$, where $\delta_N(S)$ equals 1 if $S = N$ and 0 otherwise.

3. JLI, JNU, JEF, JAN $\Leftrightarrow$

$$p^T(S) = p^{T'}(S') \, \forall S \subseteq N \setminus T, \, S' \subseteq N \setminus T' \text{ s.t. } |S| = |S'|, |T| = |T'|$$

4. JLI, JNU, JEF, JSY $\Leftrightarrow p^T(S) = p^{T'}(S) \, \forall S \subseteq N \setminus (T \cup T')$

$\therefore$ JAN, JSY not nested; non-existence w/o extra JSY terms

Interaction indices

- interaction indices assess interactions within sets
- joint Shapley measures the value-added of a set of features

1. Shapley interaction index, ϕ^{SI} (Grabisch and Roubens, 1999)

   i and j ...exhibit a positive interaction when the worth of coalition $\{i,j\}$ is more than the sum of individual worths ...

$$v(S \cup \{i,j\}) - v(S \cup \{i\}) - v(S \cup \{j\}) + v(S)$$

2. added-value index, ϕ^{AV} (Alshebli et al., 2019)

   the difference between the outcome of that group and the expected contribution of each member

3. Shapley-Taylor, ϕ^{ST} (Dhamdhere, Agarwal, and Sundararajan, 2020)

   interactions of subsets up to some size k ...analogous to how the truncated Taylor series decomposes the function value

Example ($n = 3$ glove game)

T	ϕ_i	ϕ^{SI}	ϕ^{AV}	$\phi^{ST}(2)$	$\phi^{ST}(3)$	$\phi^{J}(2)$	$\phi^{J}(3)$
1	2/3	2/3	−5/12	0	0	7/18	7/21
2	1/6	1/6	−1/6	0	0	1/18	1/21
1, 2		1/2	5/12	2/3	1	4/18	4/21
2, 3		−1/2	−1/3	−1/3	0	1/18	1/21
N		−1	1/4		−1		3/21

Example ($n = 3$ majority game: $v(i) = 0 < v(j, k) = v(N) = 1$)

T	ϕ_i	ϕ^{SI}	ϕ^{AV}	$\phi^{ST}(2)$	$\phi^{ST}(3)$	$\phi^{J}(2)$	$\phi^{J}(3)$
i	1/3	1/3	−1/3	0	0	1/9	2/21
i, j		0	1/3	1/3	1	2/9	4/21
N		−2	0		−2		3/21

Example ($n = 2$ collinearity game: $v(1) = v(2) = v(N) = 1$)

T	ϕ_i	ϕ^{SI}	ϕ^{AV}	$\phi^{ST}(2)$	$\phi^{J}(2)$
i	1/2	1/2	1/4	1	1/3
N		−1	−1/2	−1	1/3

Shapley & explainable AI

- in large, highly non-linear models, how explain a feature's importance?
- Lipovetsky and Conklin (2001), Štrumbelj and Kononenko (2014): use Shapley's value
  - agents $\Rightarrow$ features
  - characteristic function $v \Rightarrow$ prediction function f
  - $i \in S \Rightarrow$ evaluated feature i at specific x_i
  - $i \in N \backslash S \Rightarrow$ evaluate feature at mean $\bar{x}_i$
  - ϕ_i: how much does a specific value of x_i change the prediction?
- implementational decisions include: condition means on observational (q.v. Lundberg and S.-I. Lee, 2017; Frye, Mijolla, et al., 2021) or interventional (q.v. Datta, Sen, and Zick, 2016; Janzing, Minorics, and Blöbaum, 2020; Sundararajan and Najmi, 2020) data?
- for simplicity, and to ease replicability, we call the popular SHAP package (Lundberg and S.-I. Lee, 2017; Lundberg, Erion, et al., 2020)

Boston housing data (Harrison and Rubinfeld, 1978)

- 12 numerical features, and one binary feature; 506 data points
- Dhamdhere, Agarwal, and Sundararajan: random forest regression

Shapley ($k = 1$)	$\phi^{ST}(f;k)$ $k = 2$	$\phi^{ST}(f;k)$ $k = 3$	$\phi^{J}(f;k)$ $k = 2$	$\phi^{J}(f;k)$ $k = 3$
RM: 2.57	RM: 3.12	RM: 3.12	LSTAT: 0.63	LSTAT: 0.42
LSTAT: 2.47	LSTAT: 2.04	LSTAT: 2.04	RM: 0.56	RM: 0.34
AGE: 0.81	DIS: 1.55	DIS: 1.55	LSTAT, RM: 0.30	AGE: 0.11
DIS: 0.63	CRIM: 1.37	CRIM: 1.37	AGE, LSTAT: 0.21	LSTAT, RM: 0.11
CRIM: 0.46	B: 1.31	DIS, LSTAT: 1.33	AGE, RM: 0.20	NOX: 0.10
NOX: 0.44	NOX: 1.15	B: 1.31	DIS, RM: 0.19	DIS: 0.08
PTRATIO: 0.27	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B: 0.24	CHAS, RM: 0.15	AGE, CRIM, RM: 0.21	CHAS, TAX: 0.02	RAD, TAX, ZN: 0.00
TAX: 0.22	LSTAT, TAX: 0.15	AGE, DIS, PTRATIO: 0.21	RAD, ZN: 0.01	CHAS: 0.00
INDUS: 0.19	INDUS, RAD: 0.15	DIS, LSTAT, PTRATIO: 0.21	CHAS, RAD: 0.01	CHAS, RAD, TAX: 0.00
RAD: 0.13	NOX, TAX: 0.14	AGE, LSTAT, PTRATIO: 0.20	ZN: 0.01	CHAS, RAD, ZN: 0.00
ZN: 0.07	DIS, INDUS: 0.13	AGE, NOX, RM: 0.20	CHAS: 0.01	CHAS, TAX, ZN: 0.00
CHAS: 0.07	DIS, PTRATIO: 0.12	B, CRIM, LSTAT: 0.19	CHAS, ZN: 0.01	CHAS, ZN: 0.00

- $k = 2$: ϕ^J singletons, pairs mix evenly: JNU doesn't favour singletons
- $k = 3$: ϕ^J triples small, but shift ranking of singletons, pairs

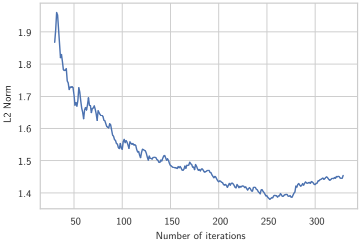
Movie reviews (Pang and L. Lee, 2005)

- 10,600 binary reviews (100 test); encoded as 1,004-vector (BoW)
- fully connected NN (2 hidden layers, 16 units/layer, ReLU activations)

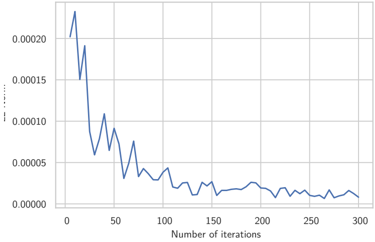
Review	joint Shapleys
1. negation: aficionados of the whodunit won't be disappointed	{disappointed}: -2×10^{-5} {won't}: 6×10^{-5} {be, disappointed}: -9×10^{-8} {won't, disappointed}: $+6 \times 10^{-8}$ {won't, be, disappointed}: 5×10^{-9}
2. enhancement: both inspiring and pure joy	{both}: 2×10^{-4} {and}: 6×10^{-5} {and, both}: 1×10^{-6}
3. context: you wish Jacquot had left well enough alone	{you, well}: $+9 \times 10^{-7}$ {left, well}: -3×10^{-7}
4. lost potential: fascinating little thriller that would have been perfect	{would}: -1×10^{-4} {fascinating}: 2×10^{-4} {would, fascinating}: $+3 \times 10^{-7}$ {would, been, fascinating}: -1×10^{-8}
5. qualifying adjective: director ...award-winning ...make a terrific effort	{effort}: -1×10^{-5} {director}: -9×10^{-6} {terrific, effort}: $+8 \times 10^{-7}$ {winning, director}: $+5 \times 10^{-7}$

- joint Shapleys have a direct interpretation
- BoW understates pairs, triples: co-occurrence rather than *k*-gram

Sampling joint Shapley values



Sampled ϕ^J converges to exact ϕ^J; $k = 2$ Boston



Difference between consecutive ϕ^J samples averages converges to zero; $k = 2$ movie review #2

Conclusion, discussion

- direct extension of Shapley value, from singletons to sets
  - how much value does a set of agents add?
  - how much does a set of feature change predictions?
  - intuitive, direct interpretation
  - existing interaction indices assess value-added within sets
- arrival order interpretation allows incorporation of causal knowledge (Frye, Rowat, and Feige, 2020)

Example ($n = 3$ majority game, glove game)

Let $i = 1$ arrive first (causal ancestor).

T	ϕ_i	ϕ^{SI}	ϕ^{AV}	$\phi^{ST}(2)$	$\phi^J(1)$	$\phi^J(2)$
1	?	?	?	?	0	0
$2;3$		?	?	?	$1/2$	$1/3$
$2,3$		?	?	?		$1/3$

$\phi^J(2)$ indicates that value only accrues to any agent arriving second.

References I

Alshebli, B. K., T. P. Michalak, O. Skibski, M. Wooldridge, and T. Rahwan (2019). "A measure of added value in groups". *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 13.4, pp. 1–46.

Bhatt, U., A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley (2020). "Explainable machine learning in deployment". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657.

Datta, A., S. Sen, and Y. Zick (2016). "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems". In: *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617.

Dhamdhere, K., A. Agarwal, and M. Sundararajan (2020). "The Shapley Taylor Interaction Index". In: *International Conference on Machine Learning*. PMLR, pp. 9259–9268.

Frye, C., D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige (2021). "Shapley explainability on the data manifold". In: *International Conference on Learning Representations*.

Frye, C., C. Rowat, and I. Feige (2020). "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability". *Advances in Neural Information Processing Systems* 33.

Grabisch, M. and M. Roubens (1999). "An axiomatic approach to the concept of interaction among players in cooperative games". *International Journal of Game Theory* 28.4, pp. 547–565.

Harrison, D. and D. L. Rubinfeld (1978). "Hedonic housing prices and the demand for clean air". *Journal of environmental economics and management* 5.1, pp. 81–102.

Janzing, D., L. Minorics, and P. Blöbaum (26–28 Aug 2020). "Feature relevance quantification in explainable AI: A causal problem". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 2907–2916.

References II

Lipovetsky, S. and M. Conklin (2001). "Analysis of regression in game theory approach". *Applied Stochastic Models in Business and Industry* 17.4, pp. 319–330.

Lucas, W. F. (Oct. 1971). "Some Recent Developments in *n*-person game theory". *SIAM Review* 13.4, pp. 491–523.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee (2020). "From local explanations to global understanding with explainable AI for trees". *Nature machine intelligence* 2.1, pp. 2522–5839.

Lundberg, S. M. and S.-I. Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Malawski, M. (2020). "A note on equal treatment and symmetry of values". In: *Transactions on Computational Collective Intelligence XXXV*. Springer, pp. 76–84.

Owen, G. (1977). "Values of games with *a priori* unions". In: *Mathematical economics and game theory. Essays in honor of Oskar Morgenstern*. Ed. by R. Henn and O. Moeschlin. Lecture Notes in Economics and Mathematical Sciences 141. Springer, pp. 76–88.

Pang, B. and L. Lee (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". In: *Proceedings of the ACL*.

Shapley, L. S. (1953). "A value for *n*-person games". In: *Contributions to the theory of games*. Ed. by H. W. Kuhn and A. W. Tucker. Vol. II. Annals of Mathematical Studies 28. Princeton: Princeton University Press. Chap. 17, pp. 307–317.

Štrumbelj, E. and I. Kononenko (2014). "Explaining prediction models and individual predictions with feature contributions". *Knowledge and Information Systems* 41.3, pp. 647–665.

References III

Sundararajan, M. and A. Najmi (13–18 Jul 2020). "The Many Shapley Values for Model Explanation". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 9269–9278.

von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. 1st edition. Princeton University Press.