Explaining the Uncertain

Stochastic Shapley Values for Gaussian Process Models

Dino Sejdinovic (Adelaide)
joint work with
Siu Lun Chau (CISPA Saarbrücken)
Krikamol Muandet (CISPA Saarbrücken)

NeurIPS 2023, arXiv:2305.15167

Business Analytics Seminar, University of Sydney
16 February 2024


Accurate or Interpretable? Choose One.

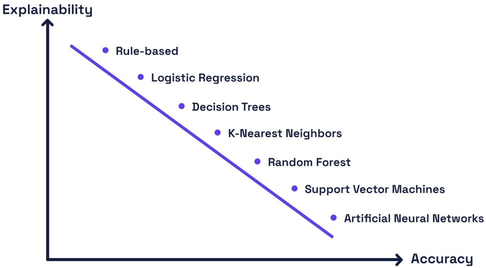
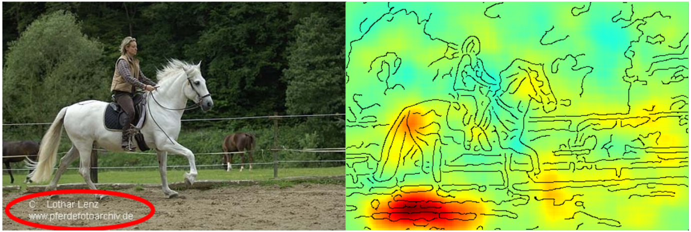


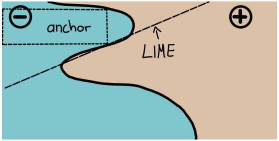
image from holisticai.com/blog/explainable-ai-dimensions

The Need for Explainability

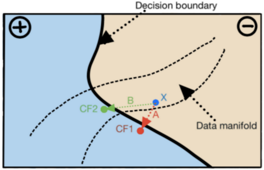


Lapuschkin et al. [2019]: *Unmasking Clever Hans Predictors and Assessing What Machines Really Learn*

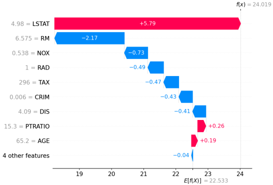
Explainable AI Zoo



(a) Anchor [Ribeiro et al., 2018]



(b) Counterfactual Explanations [Dhurandhar et al., 2018, Verma et al., 2020]



(c) Attribution methods LIME [Ribeiro et al., 2016], Sensitivity Analysis [Saltelli et al., 2008], Integrated Gradients [Qi et al., 2019], Shapley Values [Štrumbelj and Kononenko, 2014], SHAP [Lundberg and Lee, 2017]

Figure: Multitude of explanation methods

Dichotomy of feature attribution

Global Explanations: Understanding features' contribution to the model's overall behaviour (e.g. to the learnt function f over the whole dataset).

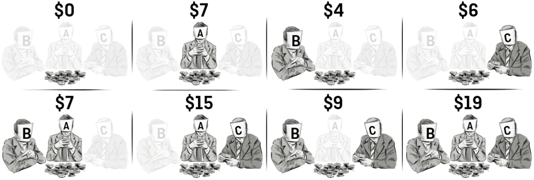
- Examples: linear model weights, global sensitivity analysis, kernel lengthscales in automatic relevance determination Gaussian process.

Local Explanations: Understanding features' contributions to an individual observation x, i.e. how did features contribute to the value of $f(\mathrm{x})$ for this specific x?

- Examples: Integrated Gradients, LIME, SHAP.

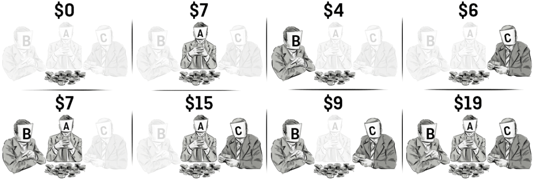
Shapley Values: Fair credit allocation for cooperative games

- Consider a d-player cooperative game where every player agrees to work towards a common goal. Denote $\Omega = \{1, .., d\}$.
- Consider the function $\nu : 2^{\Omega} \rightarrow \mathbb{R}$ that for every subset of players (coalition) returns a corresponding utility score.



Shapley Values: Fair credit allocation for cooperative games

- Consider a d-player cooperative game where every player agrees to work towards a common goal. Denote $\Omega = \{1, .., d\}$.
- Consider the function $\nu : 2^{\Omega} \to \mathbb{R}$ that for every subset of players (coalition) returns a corresponding utility score.



- How should one allocate the total utility $\nu(\Omega)$ to each player in Ω?

Shapley Values: Axiomatic properties

1. Efficiency
   - Individual credits add up to the grand profit, i.e. $\sum_{i=1}^{d} \phi_i(\nu) = \nu(\Omega)$
2. Null-Player property
   - Free riders get no credit, i.e. if $\nu(S \cup i) = \nu(S)$ for all $S \subseteq \Omega$, $\phi_i(\nu) = 0$
3. Symmetry
   - Indistinguishable players get the same credit, i.e. if $\nu(S \cup i) = \nu(S \cup j)$ for all $S \subseteq \Omega$, then $\phi_i(\nu) = \phi_j(\nu)$
4. Additivity
   - Credits from a sum of games is the sum of credits from each individual game, i.e. $\phi_i(\nu_1 + \nu_2) = \phi_i(\nu_1) + \phi_i(\nu_2)$

Shapley Values: Fair credit allocation for cooperative games

- Player i's contribution depends on the specific coalition. Their marginal contribution with respect to coalition $S \subseteq \Omega \backslash \{i\}$ is given by

$$\nu(S \cup i) - \nu(S)$$

Shapley Values: Fair credit allocation for cooperative games

- Player i's contribution depends on the specific coalition. Their marginal contribution with respect to coalition $S \subseteq \Omega \backslash \{i\}$ is given by

$$\nu(S \cup i) - \nu(S)$$

- Shapley [1953] proved that the following combination of marginal contributions uniquely satisfies all four axioms,

$$\phi_i(\nu) = \frac{1}{d} \sum_{S \subseteq \Omega \backslash \{i\}} \binom{d-1}{|S|}^{-1} \Big(\nu(S \cup i) - \nu(S) \Big).$$

Shapley Values: Fair credit allocation for cooperative games

- Player i's contribution depends on the specific coalition. Their marginal contribution with respect to coalition $S \subseteq \Omega \backslash \{i\}$ is given by

$$\nu(S \cup i) - \nu(S)$$

- Shapley [1953] proved that the following combination of marginal contributions uniquely satisfies all four axioms,

$$\phi_i(\nu) = \frac{1}{\text{\# of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to a coalition}}{\text{\# of coalitions excluding } i \text{ of this size}}.$$

Shapley Values: Fair credit allocation for cooperative games

- Player i's contribution depends on the specific coalition. Their marginal contribution with respect to coalition $S \subseteq \Omega \backslash \{i\}$ is given by

$$\nu(S \cup i) - \nu(S)$$

- Shapley [1953] proved that the following combination of marginal contributions uniquely satisfies all four axioms,

$$\phi_i(\nu) = \frac{1}{\text{\# of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to a coalition}}{\text{\# of coalitions excluding } i \text{ of this size}}.$$

- An alternative interpretation using the order of players:

$$\phi_i(\nu) = \frac{1}{d!} \sum_{\sigma} \left(\nu(P_i^{\sigma} \cup i) - \nu(P_i^{\sigma})\right).$$

where the sum ranges over all $d!$ permutations σ of $\Omega = \{1, \ldots, d\}$ and P_i^{σ} is the set of players which precede i in σ.

Shapley values for explainability?

- **Data**: For concreteness, consider a supervised learning setting, with data $\mathcal{D} = \{\mathrm{x}_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$.

Shapley values for explainability?

- **Data**: For concreteness, consider a supervised learning setting, with data $\mathcal{D} = \{\mathrm{x}_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Fit the model:** Learn some $f : \mathcal{X} \to \mathcal{Y}$ via your favourite ML technique: random forest, kernel ridge regression, deep neural network.... by minimise expected loss.

Shapley values for explainability?

- **Data**: For concreteness, consider a supervised learning setting, with data $\mathcal{D} = \{\mathrm{x}_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Fit the model:** Learn some $f : \mathcal{X} \to \mathcal{Y}$ via your favourite ML technique: random forest, kernel ridge regression, deep neural network.... by minimise expected loss.
- **Explain the model:** How to frame feature attribution as a cooperative game?

Shapley values for explainability?

- **Data**: For concreteness, consider a supervised learning setting, with data $\mathcal{D} = \{\mathrm{x}_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Fit the model:** Learn some $f : \mathcal{X} \to \mathcal{Y}$ via your favourite ML technique: random forest, kernel ridge regression, deep neural network.... by minimise expected loss.
- **Explain the model:** How to frame feature attribution as a cooperative game?
  - Players are features: $\Omega = \{1, \dots, d\}$ (features indices)

Shapley values for explainability?

- **Data**: For concreteness, consider a supervised learning setting, with data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Fit the model:** Learn some $f : \mathcal{X} \to \mathcal{Y}$ via your favourite ML technique: random forest, kernel ridge regression, deep neural network.... by minimise expected loss.
- **Explain the model:** How to frame feature attribution as a cooperative game?
  - Players are features: $\Omega = \{1, \dots, d\}$ (features indices)
  - The grand profit is the prediction itself, i.e. $\nu_{x,f}(\Omega) = f(x)$

Shapley values for explainability?

- **Data**: For concreteness, consider a supervised learning setting, with data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Fit the model:** Learn some $f : \mathcal{X} \to \mathcal{Y}$ via your favourite ML technique: random forest, kernel ridge regression, deep neural network.... by minimise expected loss.
- **Explain the model:** How to frame feature attribution as a cooperative game?
  - Players are features: $\Omega = \{1, \dots, d\}$ (features indices)
  - The grand profit is the prediction itself, i.e. $\nu_{x,f}(\Omega) = f(x)$
  - To define the value function on any coalition of features $S \subset \Omega$, average the predictions over the remaining features:

$$\nu_{x,f}(S) := \mathbb{E}_{r(X|X_S = x_S)}[f(X) \mid X_S = x_S],$$

where r is some reference distribution and x_S is the subvector of x corresponding to features in S.

Shapley values for explainability?

$$\phi_{x,i}(\nu) = \frac{1}{d} \sum_{S \subseteq \Omega \setminus \{i\}} \binom{d-1}{|S|}^{-1} \Big(\nu_{x,f}(S \cup i) - \nu_{x,f}(S) \Big).$$

Shapley values for explainability?

$$\phi_{x,i}(\nu) = \frac{1}{d} \sum_{S \subseteq \Omega \setminus \{i\}} \binom{d-1}{|S|}^{-1} \Big(\nu_{x,f}(S \cup i) - \nu_{x,f}(S)\Big).$$

Note the sum over all subsets of the set of features – this is not going to be possible to compute even for a moderate number of features!

Additive feature attribution model

- The best explanation of a simple model is the model itself.
- What to do for a complex model? Build a simpler one: explanation model.
- A simple idea: place a locally linear model $u_{\mathrm{x}} : \{0,1\}^d \mapsto \mathbb{R}$ around the input x as a function of which features are switched on/off:

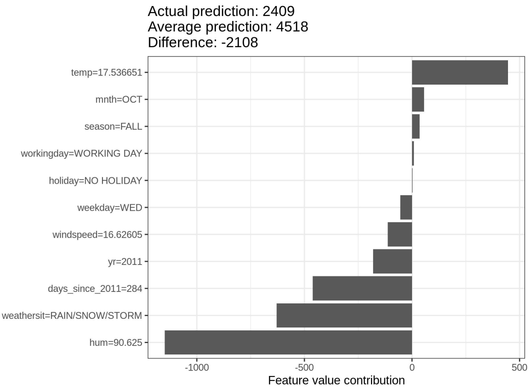
$$u_{\mathrm{x}}(S) := \phi_{\mathrm{x},0} + \sum_{i=1}^{d} \phi_{\mathrm{x},i} \mathrm{z}_i$$

  with $\mathrm{z}_i = 1\{i \in S\}$. Models like LIME [Ribeiro et al., 2016] take this perspective. We want $u_{\mathrm{x}}(S) \approx \nu_{x,f}(S)$.
- Lundberg and Lee [2017] makes a connection to Shapley values: they are solution to the weighted least squares problem

$$\min_{u_{\mathrm{x}}} \sum_{S} w(S) \left(u_{\mathrm{x}}(S) - \nu_{x,f}(S)\right)^2 .$$

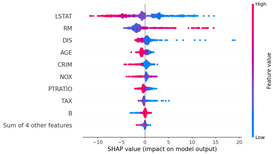
- **SHAP algorithm**: sample as many S as you can afford, compute the value function for those coalitions and simply solve weighted least squares regression.

Example: Bike Rental

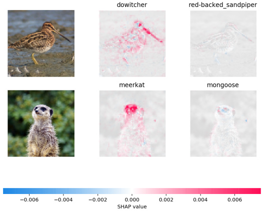


Example from Molnar. The weather situation and humidity had the largest negative contributions. The temperature on this day had a positive contribution. The sum of Shapley values yields the difference of actual and average prediction, i.e. $f(x) - \mathbb{E}_X[f(X)]$.

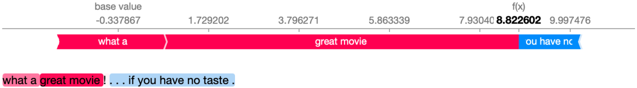
Attribution examples



(a) tabular data



(b) image

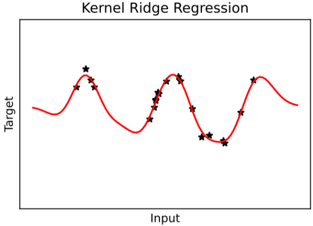


(c) text

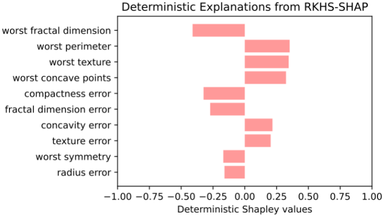
Figure: SHAP on different data types

Motivation: Feature attribution as explanation

Deterministic model...

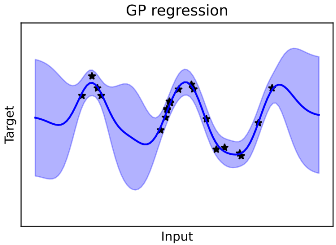


gives deterministic explanations..

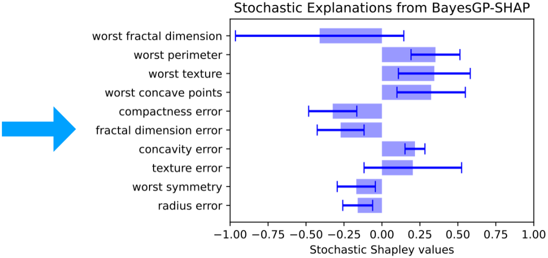


Motivation: Feature attribution as explanation

GP gives predictive uncertainty



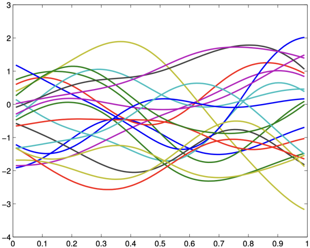
shouldn't its explanations also?

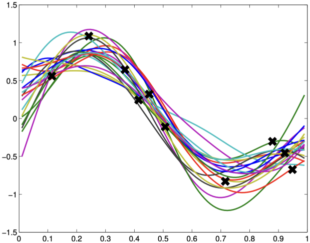


Recap on Gaussian process

Consider function values $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$ at a set of inputs $\mathbf{X}$, and observations $\mathbf{y} = [y_1, ..., y_n]$, with prior and likelihood as,

$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K}), \qquad \mathbf{y} \mid \mathbf{f} \sim p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} p(y_i \mid f(\mathbf{x}_i))$$





Recap on Gaussian process

GP Regression

- Given data $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^n$ and a GP prior $f \sim \mathcal{GP}(0, k)$, assuming likelihood:

$$y_i = f(\mathrm{x}_i) + \epsilon_i, \ \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

then the posterior $f \mid \mathcal{D}$ is also a GP with,

$$\tilde{m}(\mathrm{x}) = k(\mathrm{x}, \mathrm{X})(\mathrm{K_{XX}} + \sigma^2 I)^{-1}\mathrm{y}$$
$$\tilde{k}(\mathrm{x}, \mathrm{x}') = k(\mathrm{x}, \mathrm{x}') - k(\mathrm{x}, \mathrm{X})(\mathrm{K_{XX}} + \sigma^2 I)^{-1} k(\mathrm{X}, \mathrm{x}')$$

Other likelihoods

- Variational framework for computational scalability and other likelihood models (classification, Poisson regression etc) [Titsias, 2009]

What's useful about GPs?

Probabilistic

- Instead of giving a point estimate, a GP model returns a predictive distribution and quantifies uncertainty.

Nonparametric

- GPs do not assume a fixed parametric form for the underlying function being modelled.

Prior knowledge

- The choice of covariance function can incorporate structural assumptions about functions being modelled.

Versatile

- Can be applied to supervised or unsupervised learning, spatiotemporal models, probabilistic integration, Bayesian optimization...

Now let's explain GP?...

- Consider a standard SHAP procedure for GP: for a GP f, $f(\mathrm{x})$ is a (Gaussian) random variable, and hence the value function $\nu_{\mathrm{x},f} : S \mapsto \mathbb{E}[f(X) \mid X_S = \mathrm{x}_S]$ is also random.
- We can proceed two ways:
  - Sample multiple realisations of f from $p(f \mid D)$ and apply SHAP to each of them individually [Marx et al., 2023].
  - Model value function and Shapley values themselves as stochastic processes.

GP explainability through a stochastic game

Build stochastic game out of GP:

- Stochastic games : $\nu_{\mathrm{x},f} : 2^{\Omega} \to \mathcal{L}_2(\mathbb{R})$ given by

$$\nu_{\mathrm{x},f}(S) := \mathbb{E}_X[f(X) \mid X_S = \mathrm{x}_S].$$

Recall: this quantity is random because f is random.

GP explainability through a stochastic game

Build stochastic game out of GP:

- Stochastic games : $\nu_{x,f} : 2^{\Omega} \to \mathcal{L}_2(\mathbb{R})$ given by

$$\nu_{x,f}(S) := \mathbb{E}_X[f(X) \mid X_S = x_S].$$

  Recall: this quantity is random because f is random.

- In Chau et al. [2021], we studied ways to model **conditional expectations of GPs - which are themselves GPs** by linearity.

Let $f \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ with integrable sample paths, i.e. $\int_{\mathcal{X}} |f| dp_X < \infty$ a.s. The stochastic payoff function $\nu_{x,f}$ induced by f is a GP (on $\mathbb{R}^d \times 2^{\Omega}$) with the following mean and covariance functions:

$$m_\nu(x, S) := \mathbb{E}_X[\tilde{m}(X) \mid X_S = x_S],$$
$$k_\nu((x, S), (x', S')) := \mathbb{E}_{X,X'}\left[\tilde{k}(X, X') \mid X_S = x_S, X'_{S'} = x'_{S'}\right].$$

We can estimate these using standard tricks from RKHS mean embeddings.

GP explainability through a stochastic game

Build stochastic game out of GP:

- Stochastic games : $\nu_{\mathrm{x},f} : 2^{\Omega} \to \mathcal{L}_2(\mathbb{R})$ given by

$$\nu_{\mathrm{x},f}(S) := \mathbb{E}_X[f(X) \mid X_S = \mathrm{x}_S].$$

  Recall: this quantity is random because f is random.

- In Chau et al. [2021], we studied ways to model **conditional expectations of GPs - which are themselves GPs** by linearity.

Let $f \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ with integrable sample paths, i.e. $\int_{\mathcal{X}} |f| dp_X < \infty$ a.s. The stochastic payoff function $\nu_{\mathrm{x},f}$ induced by f is a GP (on $\mathbb{R}^d \times 2^{\Omega}$) with the following mean and covariance functions:

$$m_\nu(\mathrm{x}, S) := \mathbb{E}_X[\tilde{m}(X) \mid X_S = \mathrm{x}_S],$$

$$k_\nu\left((\mathrm{x}, S), (\mathrm{x}', S')\right) := \mathbb{E}_{X,X'}\left[\tilde{k}(X, X') \mid X_S = \mathrm{x}_S, X'_{S'} = \mathrm{x}'_{S'}\right].$$

We can estimate these using standard tricks from RKHS mean embeddings.

GP explainability through a stochastic game

Build stochastic game out of GP:

- Stochastic games : $\nu_{\mathrm{x},f} : 2^{\Omega} \to \mathcal{L}_2(\mathbb{R})$ given by

$$\nu_{\mathrm{x},f}(S) := \mathbb{E}_X[f(X) \mid X_S = \mathrm{x}_S].$$

  Recall: this quantity is random because f is random.

- In Chau et al. [2021], we studied ways to model **conditional expectations of GPs - which are themselves GPs** by linearity.

Let $f \sim \mathcal{GP}(\tilde{m}, \tilde{k})$ with integrable sample paths, i.e. $\int_{\mathcal{X}} |f| dp_X < \infty$ a.s. The stochastic payoff function $\nu_{\mathrm{x},f}$ induced by f is a GP (on $\mathbb{R}^d \times 2^{\Omega}$) with the following mean and covariance functions:

$$m_\nu(\mathrm{x}, S) := \mathbb{E}_X[\tilde{m}(X) \mid X_S = \mathrm{x}_S],$$
$$k_\nu\left((\mathrm{x}, S), (\mathrm{x}', S')\right) := \mathbb{E}_{X,X'}\left[\tilde{k}(X, X') \mid X_S = \mathrm{x}_S, X'_{S'} = \mathrm{x}'_{S'}\right].$$

We can estimate these using standard tricks from RKHS mean embeddings.

- **TL;DR: the stochastic game is also a GP that can be characterised nicely.**

Now the (stochastic) game is defined. Let's Shapley.

- Given value function evaluations $\mathrm{v}_\mathrm{x} := [\nu_f(\mathrm{x}, S_1), \dots \nu_f(\mathrm{x}, S_m)]^\top$ for m coalitions, SHAP algorithm gives vector $\phi_\mathrm{x}(\nu) = \mathrm{A}\mathrm{v}_\mathrm{x}$ with $\mathrm{A} = (\mathrm{Z}^\top \mathrm{W}\mathrm{Z})^{-1}\mathrm{Z}^\top \mathrm{W}$ where Z is the binary matrix representing sampled coalitions, and W is the corresponding weight matrix.
  - WLS solution of additive feature attribution model

Now the (stochastic) game is defined. Let's Shapley.

- Given value function evaluations $\mathrm{v_x} := [\nu_f(\mathrm{x}, S_1), \dots \nu_f(\mathrm{x}, S_m)]^\top$ for m coalitions, SHAP algorithm gives vector $\phi_x(\nu) = \mathrm{A}\mathrm{v_x}$ with $\mathrm{A} = (\mathrm{Z}^\top \mathrm{W}\mathrm{Z})^{-1}\mathrm{Z}^\top \mathrm{W}$ where Z is the binary matrix representing sampled coalitions, and W is the corresponding weight matrix.
  - WLS solution of additive feature attribution model
- If $\nu_{\mathrm{x},f}$ is a stochastic game, the corresponding stochastic vector of Shapley values $\phi_x(\nu)$ follows a d-dimensional multivariate Gaussian distribution

$$\phi_x(\nu) \sim \mathcal{N}(\mathrm{A}\mathbb{E}[\mathrm{v_x}], \mathrm{A}\mathbb{V}[\mathrm{v_x}]\mathrm{A}^\top)$$

Now the (stochastic) game is defined. Let's Shapley.

- Given value function evaluations $\mathrm{v_x} := [\nu_f(\mathrm{x}, S_1), \dots \nu_f(\mathrm{x}, S_m)]^\top$ for m coalitions, SHAP algorithm gives vector $\phi_x(\nu) = \mathrm{A}\mathrm{v_x}$ with $\mathrm{A} = (\mathrm{Z}^\top \mathrm{W} \mathrm{Z})^{-1}\mathrm{Z}^\top \mathrm{W}$ where Z is the binary matrix representing sampled coalitions, and W is the corresponding weight matrix.
  - WLS solution of additive feature attribution model
- If $\nu_{\mathrm{x},f}$ is a stochastic game, the corresponding stochastic vector of Shapley values $\phi_x(\nu)$ follows a d-dimensional multivariate Gaussian distribution

$$\phi_x(\nu) \sim \mathcal{N}(\mathrm{A}\mathbb{E}[\mathrm{v_x}], \mathrm{A}\mathbb{V}[\mathrm{v_x}]\mathrm{A}^\top)$$

- Moreover, this is a (multi-output) Gaussian process in x with tractable covariance function – we can easily "amortize": fit Shapley values as smooth functions of x.

Short summary

- Stochastic game built for GPs are themselves GPs that can be fully characterised.
- Stochastic Shapley values for this stochastic game are also GPs.
- Estimation is straightforward utilising RKHS tools (conditional mean embeddings).

Bonus: BayesGP-SHAP

Integrating BayesSHAP [Slack et al., 2021] with GP-SHAP to tackle more uncertainty.

- Besides predictive uncertainty from the GP, there is additional epistemic uncertainty arising due to *estimation* of Shapley values through the WLS approach.
- Slack et al. [2021] captures this uncertainty by turning the WLS into a Bayesian WLS.
- We incorporate their approach into GP-SHAP seamlessly thanks to Gaussian conjugacy.

Ablation study on the captured uncertainties

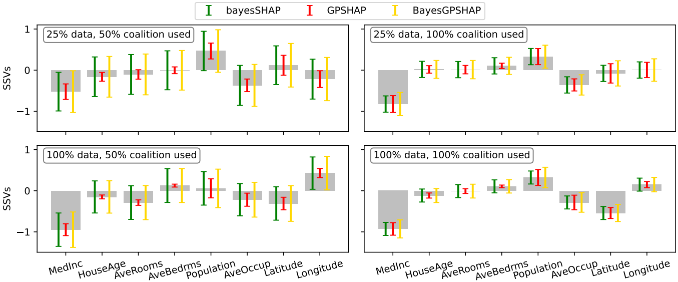


Figure: Ablation study on different uncertainties captured by GP-SHAP, BayesSHAP, and BayesGP-SHAP when computing local explanations (SSVs) using the California housing dataset [Pace and Barry, 1997]. 95% credible intervals around explanations are shown.

Exploring stochastic local explanations

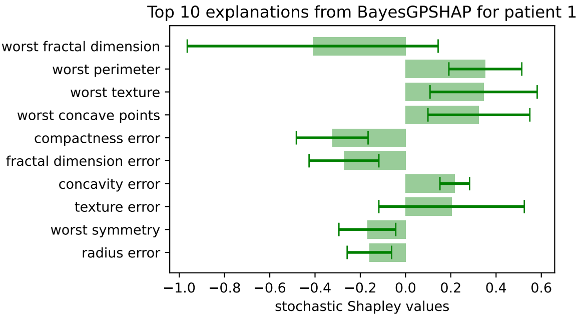
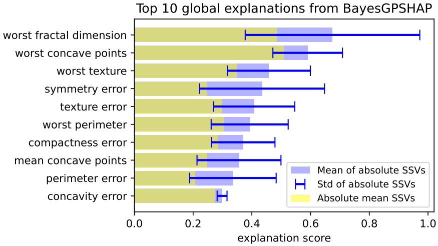


Figure: Besides the usual (mean) contribution, we can quantify the uncertainty around this explanation, and calibrate our belief from this model.

Exploring stochastic global explanations

- Global explanations are often taken as averages (over input distribution) of absolute (deterministic) Shapley values. (Absolute mean SSVs)
- However, this does not take into account the explanation uncertainty.
- Instead, we can look into the distribution of absolute SSVs (folded Gaussian) for each input and then average.
- Global importance ranking changes!



Exploring stochastic explanations: Explanation correlation

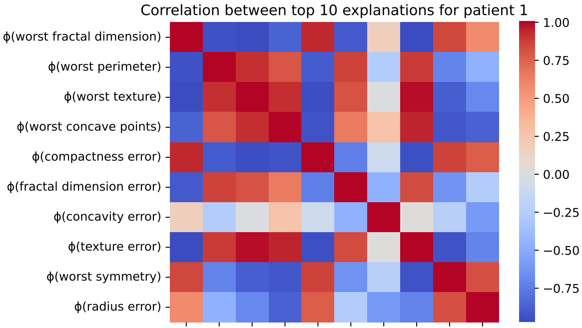


Figure: Tractable covariance structure across explanations allows studying dependencies between feature attributions.

Summary

- Explaining machine learning model through feature attribution can be framed as a cooperative game.
- When the model is probabilistic, the cooperative game and the corresponding attributions become stochastic as well.
- GP-SHAP captures uncertainty in a predictive model with a tractable covariance structure and can be combined with Shapley value estimation uncertainty.

Future work

- Explaining uncertainty in other probabilistic models such as Bayesian Neural Networks.
- Can we use the uncertainty in Shapley values for downstream tasks such as Bayesian optimisation?

(a) Paper

(b) Code

References I

Siu Lun Chau, Shahine Bouabid, and Dino Sejdinovic. Deconditional Downscaling with Gaussian Processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:17813–17825, 2021.

Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. RKHS-SHAP: Shapley values for kernel methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:13050–13063, 2022.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the Uncertain: Stochastic Shapley Values for Gaussian Process Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations (ICLR)*, 2021.

References II

Robert Hu, Siu Lun Chau, Jaime Ferrando Huertas, and Dino Sejdinovic. Explaining preferences with shapley values. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27664–27677, 2022.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2907–2916. PMLR, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, 2017.

References III

Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 7375–7391. PMLR, 2023.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients. In *CVPR Workshops*, volume 2, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 32, 2018.

References IV

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:9391–9404, 2021.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41 (3):647–665, 2014.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 567–574. PMLR, 2009.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Examples of value functions ν

- **Interventional Value functions** [Janzing et al., 2020]

$$\nu^{(I)}_{x,S}(f) = \mathbb{E}_{p_I(X_{S^c})}\left[f\left(\{x_S, X_{S^c}\}\right)\right]$$

where $p_I(X_{S^c}) = \prod_{j \in S^c} p(X^{(j)})$ assumes feature independence.

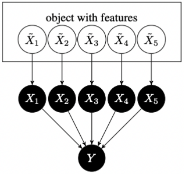
- **Observational value function** [Frye et al., 2021]

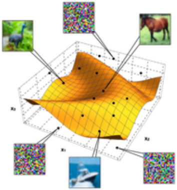
$$\nu^{(O)}_{x,S}(f) = \mathbb{E}_{p(X_{S^c} | X_S = x_S)}\left[f\left(\{x_S, X_{S^c}\}\right)\right]$$

where p is the observed data distribution.

Choice of value functions: A long-standing debate

1. Janzing et al. [2020] argued from a causal perspective that $\nu^{(I)}_{x,S}$ is the correct notion to capture feature relevance, as it treats features as direct causes to model predictions.
2. Frye et al. [2021] argued otherwise, saying that marginal expectations will evaluate value functions at unseen region of the data manifold, thus producing unrealistic explanations. Moreover, it ignores feature correlations.





Something extra: the Shapley prior over explanations

Predicting explanations using a Shapley GP model

- Treat explanation as a vector-valued mapping $\phi : \mathcal{X} \to \mathbb{R}^d$. Starting with a GP prior over f, we have an induced GP prior over ϕ, the explanation function.
  The prior $f \sim \mathcal{GP}(0, k)$ and the corresponding stochastic game $\nu_f(\mathrm{x}, S) = \mathbb{E}[f(X) \mid X_S = \mathrm{x}_S]$ induce a vector-valued GP prior over the explanation functions $\phi \sim \mathcal{GP}(0, \kappa)$ where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$ is the matrix-valued covariance kernel

$$\kappa(\mathrm{x}, \mathrm{x}') = \mathcal{A}(\mathrm{x})^\top \mathcal{A}(\mathrm{x}'), \quad \mathcal{A}(\mathrm{x}) = \Psi(\mathrm{x}) \mathrm{A}^\top$$

  where $\Psi(\mathrm{x}) = \left[\mathbb{E}[k(\cdot, X) \mid X_{S_1} = x_{S_1}], \ldots, \mathbb{E}[k(\cdot, X) \mid X_{S_{2^d}} = x_{S_{2^d}}]\right]$.
- Can now do vector-valued regression on old explanations and predict new ones.
- These explanations do not need to come from a GP model!

Something extra: the Shapley prior over explanations

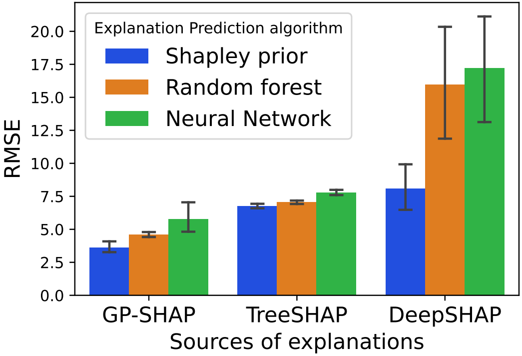
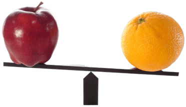
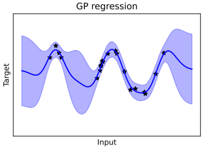
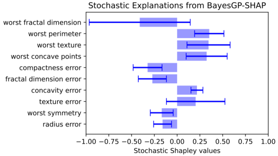


Figure: Predictive performance of using Shapley prior to predict explanations generated from different explanation algorithms on the diabetes dataset.

Shapley Values, Preferences and Uncertainty







- When using a preferential model, should we be explaining the preferences among the two items or the utilities of the individual items?
  Hu et al. [2022]: R. Hu, S. L. Chau, J. F. Huertas, and DS, *Explaining Preferences with Shapley Values*, in NeurIPS, 2022.
- Efficient computation of Shapley values for kernel methods + a method to control particular feature attribution, e.g. fairness constraints.
  Chau et al. [2022]: S. L. Chau, R. Hu, J. Gonzalez, and DS, *RKHS-SHAP: Shapley Values for Kernel Methods*, in NeurIPS, 2022.
- Explain not just point predictions, but also uncertainty in those predictions – *which features are most responsible for the model uncertainty*?
  Chau et al. [2023]: S. L. Chau, K. Muandet, and DS, *Explaining the Uncertain: Stochastic Shapley Values for Gaussian Process Models*, in NeurIPS, 2023.