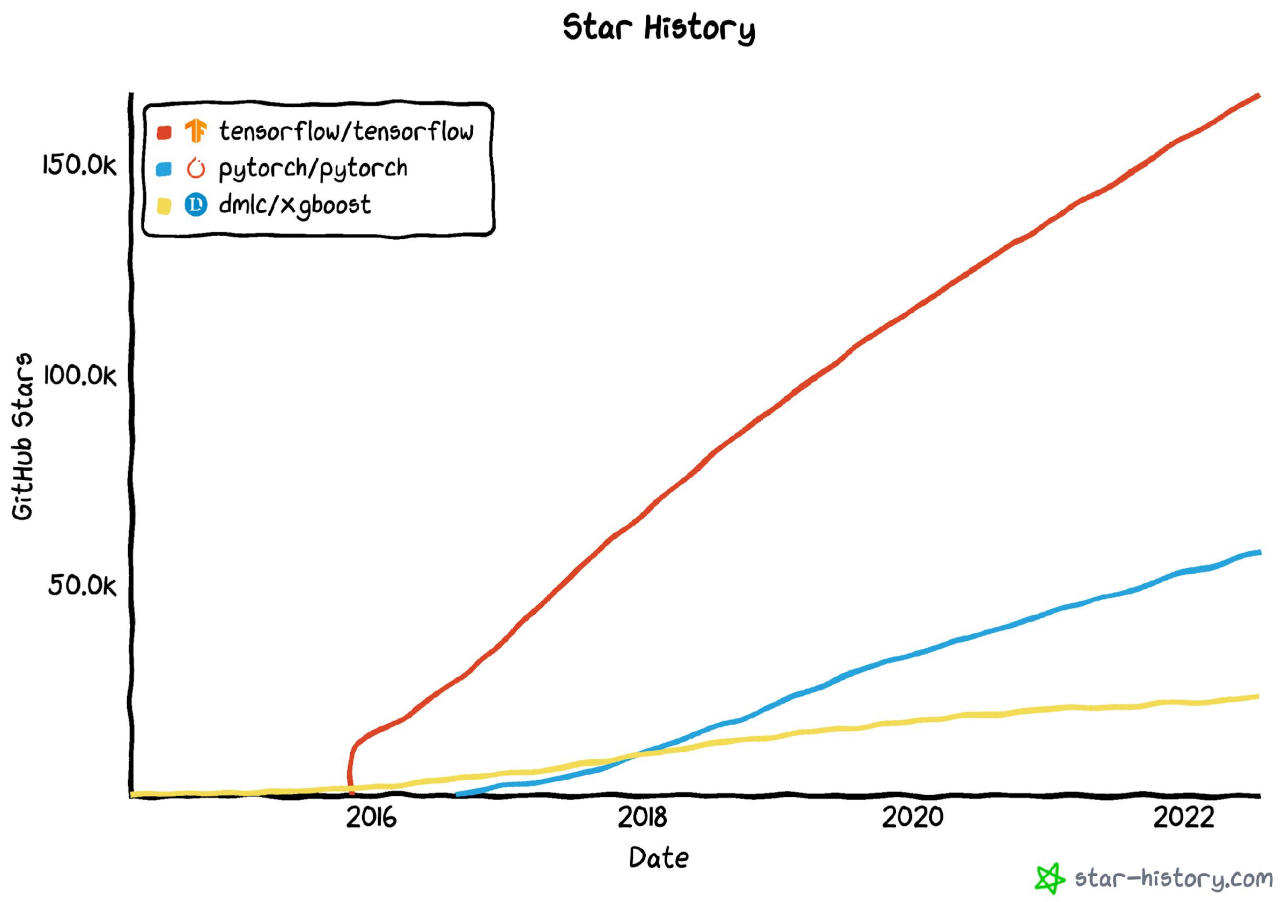
Algorithms to estimate Shapley value feature attributions

Hugh Chen

Topics

- **Why explain models?**
- What are Shapley values?
- What are Shapley value explanations?
- Challenge 1: Feature removal approaches
- Challenge 2: Tractable estimation strategies

Machine learning (ML) is increasingly widespread



Increasing regulatory desire for explanations

General Data Protection Regulation (2018)

"[the data subject should have] the **right … to obtain an explanation** of the decision reached"

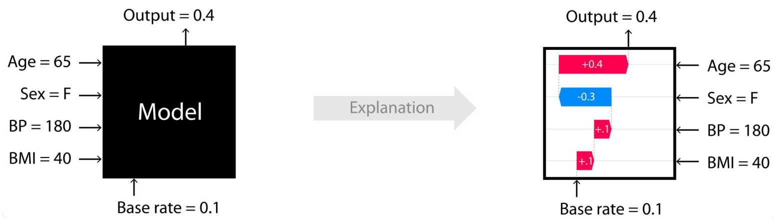
Equal Credit Opportunity Act (1974)

"The statement of reasons for adverse action … must be specific and **indicate the principal reason(s) for the adverse action**"

Many types of explanations

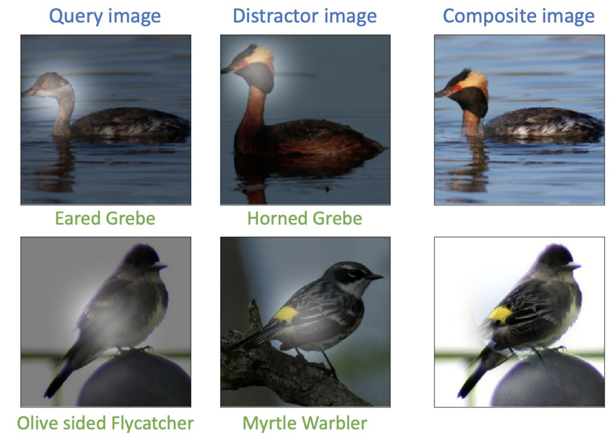
Feature attribution

SHAP



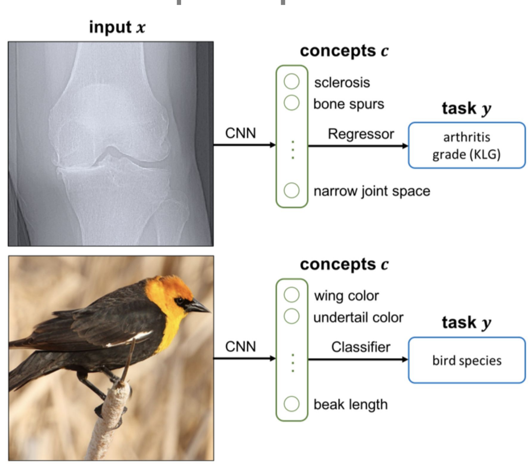
https://github.com/slundberg/shap

Counterfactuals



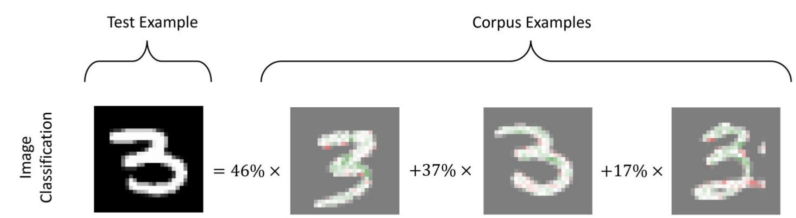
Goyal, Yash, et al. "Counterfactual visual explanations." International Conference on Machine Learning. PMLR, 2019.

Concept explanations



Koh, Pang Wei, et al. "Concept bottleneck models." International Conference on Machine Learning. PMLR, 2020.

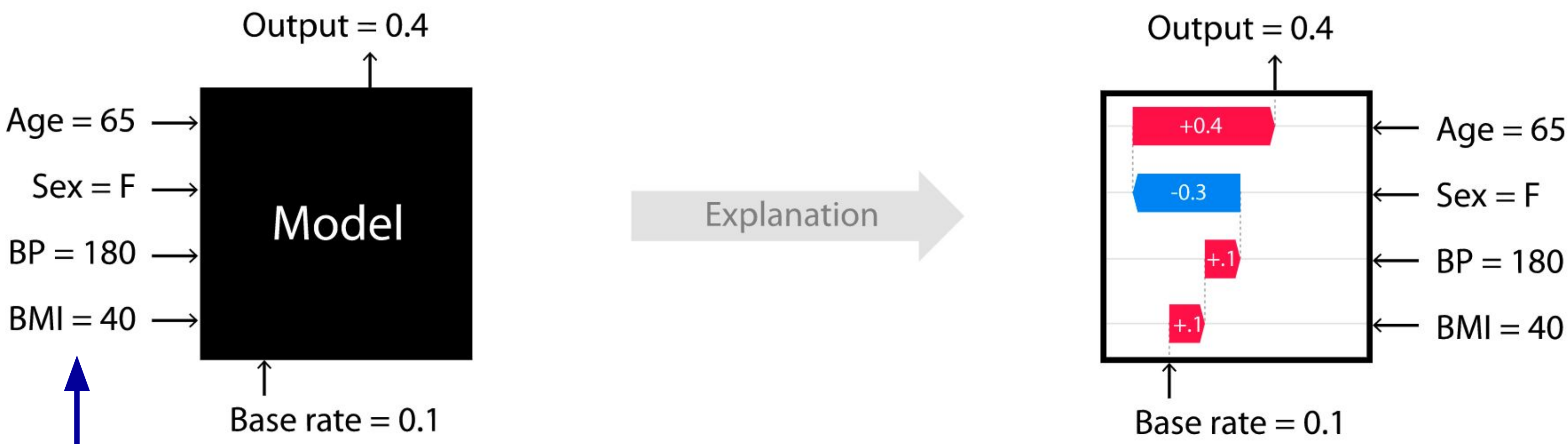
Example attribution



Crabbé, Jonathan, et al. "Explaining Latent Representations with a Corpus of Examples." Advances in Neural Information Processing Systems 34 (2021): 12154-12166.

Local feature attributions

Baseline
$x^b \in \mathbb{R}^d$



Explicand
$x^e \in \mathbb{R}^d$

Model
$f: \mathbb{R}^d \to \mathbb{R}$

Attribution
$\phi \in \mathbb{R}^d$

What defines a good attribution?

https://github.com/slundberg/shap

Topics

- Why explain models?
- **What are Shapley values?**
- What are Shapley value explanations?
- Challenge 1: Feature removal approaches
- Challenge 2: Tractable estimation strategies

The Shapley value

A unique credit allocation of the total surplus of a coalitional game among the game's players.

Players

Coalitional game

a	b	c	v
0	0	0	0
1	0	0	1
0	1	0	1
0	0	1	1
1	0	1	2
1	1	0	2
0	1	1	2
1	1	1	3

Total surplus

Credit allocation

a	1
b	1
c	1

The Shapley value

A unique credit allocation of the total surplus of a coalitional game among the game's players.

Players

Coalitional game

a	b	c	v
0	0	0	0
1	0	0	2
0	1	0	1
0	0	1	1
1	0	1	3
1	1	0	3
0	1	1	2
1	1	1	4

Total surplus

Increasing a's contribution

a	2
b	1
c	1

Credit allocation

Definition of the Shapley value

- Notation:
  - Players are $D = \{1, \cdots, d\}$
  - Coalitional game is $v(S): 2^D \to \mathbb{R}$
- The Shapley value ← Unique solution under a set of axioms

$$\overbrace{\phi_i(v)}^{\text{Shapley value for } i} = \underbrace{\sum_{S \subseteq D \setminus \{i\}}}_{\text{All coalitions excluding } i} \underbrace{W(|S|, |D|)}_{\text{Weight}} \overbrace{\Big(v(S \cup \{i\}) - v(S)\Big)}^{\text{Marginal contribution of } i}$$

Topics

- Why explain models?
- What are Shapley values?
- **What are Shapley value explanations?**
- Challenge 1: Feature removal approaches
- Challenge 2: Tractable estimation strategies

Shapley value feature attributions

- Shapley value explanations

Value is model prediction

Players are features

Feature removal approach

Define $v(S)$ → **Calculate** $\phi(v)$

Tractable estimation strategy

- We will review many techniques and algorithms to estimate Shapley value explanations
  - First, we will define two factors of complexity

Ian Covert

Scott Lundberg

Chen, Hugh* and Covert, Ian* and Lundberg, Scott and Lee, Su-In. "Algorithms to estimate Shapley value feature attributions." *Nature Machine Intelligence* 2023.

Factor of complexity 1
Feature removal approach

Define
$v(S)$

- The original paper on Shapley value explanations proposed SHAP values
- They were shown to be a unique solution in the class of additive feature attribution methods based on a set of axioms
- However, its uniqueness depends on defining a coalitional game based on the model
- This has led to distinct Shapley value explanations that differ in how they remove features

Chen, Hugh* and Covert, Ian* and Lundberg, Scott and Lee, Su-In. "Algorithms to estimate Shapley value feature attributions." Nature Machine Intelligence 2023.

Factor of complexity 2
Tractable estimation strategy

Calculate
$\phi(v)$

- Calculating Shapley values is, in the general case, an NP-hard problem
- The original SHAP paper discussed strategies to estimate Shapley values
  - Model-agnostic – KernelSHAP
  - Model-specific – LinearSHAP, MaxSHAP, DeepSHAP
- Since then, many new algorithms have been proposed

Chen, Hugh* and Covert, Ian* and Lundberg, Scott and Lee, Su-In. "Algorithms to estimate Shapley value feature attributions." Nature Machine Intelligence 2023.

Why review this literature?

- These two factors of complexity have led to an abundance of papers and algorithms
- Coupled with the complexity of the topic the literature has become difficult to navigate

Method	Factors of complexity			Properties		
	Estimation strategy	Removal approach	Removal variant	Model-agnostic	Bias-free	Variance-free
ApproSemivalue [30]	SV	None	Exact	Yes	Yes	No
L-Shapley [26]	SV	Marginal	Empirical	Yes	No	No♣
C-Shapley [26]	SV	Marginal	Empirical	Yes	No	No♣
ApproShapley [30]	RO	None	Exact	Yes	Yes	No
IME [27]	RO	Marginal	Empirical	Yes	Yes	No
CES [22]	RO	Conditional	Empirical	Yes	No	No
Shapley cohort refinement [53]	RO	Conditional	Empirical*	Yes	No	No
Generative model [50]	RO	Conditional	Generative	Yes	No	No
Surrogate model [50]	RO	Conditional	Surrogate	Yes	No	No
Multilinear extension sampling [31]	ME	Marginal	Empirical	Yes	Yes♢	No
SGD-Shapley [54]	WLS	Baseline	Exact	Yes	No♡	No
KernelSHAP [15, 52]	WLS	Marginal	Empirical	Yes	Yes♠	No
Parametric KernelSHAP [49]	WLS	Conditional	Parametric	Yes	No	No
Nonparameteric KernelSHAP [49]	WLS	Conditional	Empirical*	Yes	No	No
FastSHAP [32]	WLS	Conditional	Surrogate	Yes	No	No
LinearSHAP [28]	Linear	Marginal	Empirical	No	Yes	Yes
Correlated LinearSHAP [28]	Linear	Conditional	Parametric	No	No	No
Interventional TreeSHAP [16]	Tree	Marginal	Empirical	No	Yes	Yes
Path dependent TreeSHAP [16]	Tree	Conditional	Empirical*	No	No	Yes
DeepLIFT [17]	Deep	Baseline	Exact	No	No	Yes
DeepSHAP [15]	Deep	Marginal	Empirical	No	No	Yes
DASP [33]	Deep	Baseline	Exact	No	No	No♣
Shallow ShapNet [34]	Deep	Baseline	Exact	No	Yes	Yes
Deep ShapNet [34]	Deep	Baseline	Exact	No	No	Yes

Topics

- Why explain models?
- What are Shapley values?
- What are Shapley value explanations?
- **Challenge 1: Feature removal approaches**
- Challenge 2: Tractable estimation strategies

Feature removal approaches

Define $v(S)$

- To use Shapley values, we first need a coalitional game
  - But ML models are not coalitional games!
    - Models take vector inputs ($\mathbb{R}^d$)
    - Games take set inputs (2^D)
- Define a coalitional game based on the model
  - If a feature is in S, it is present
  - If a feature is not in S, it is absent

Feature removal approaches

- Baseline Shapley values

$$v(S) = f\left(x_S^e, x_{\bar{S}}^b\right)$$

- Marginal Shapley values

$$v(S) = \mathbb{E}_{p(x_{\bar{S}})}[f(x_S^e, x_{\bar{S}})]$$

- Conditional Shapley values

$$v(S) = \mathbb{E}_{p(x_{\bar{S}} | x_S)}[f(x_S^e, x_{\bar{S}})]$$

Too dependent on a single baseline

…but actually estimated this

The original SHAP paper proposed this…

Simulated example

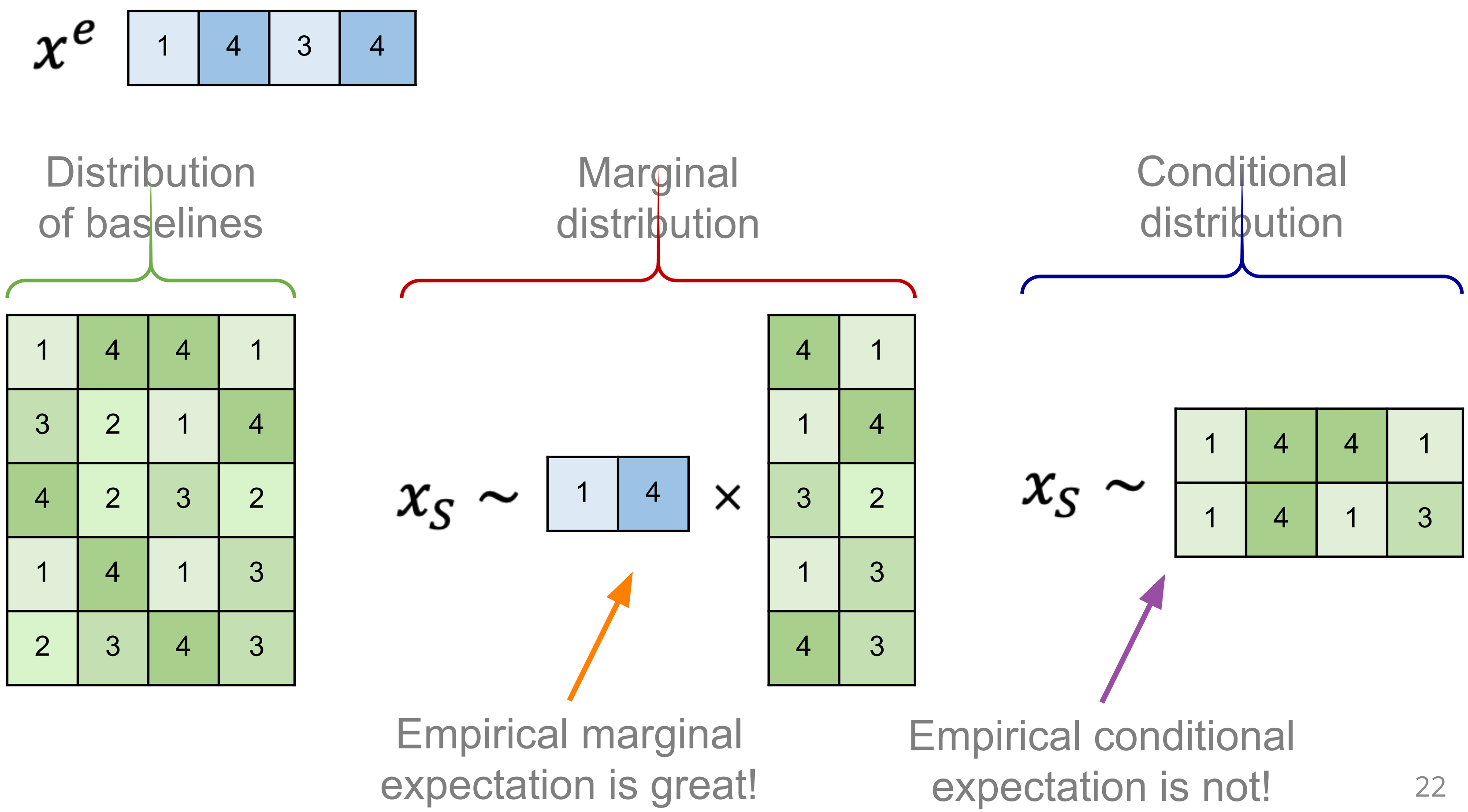
	Linear model coefficients β	Covariance Σ			Marginal Shapley value ϕ^m	Conditional Shapley value ϕ^c
Independent full model	1	1	0	0	1	1
	2	0	1	0	2	2
	3	0	0	1	3	3
Dependent full model	1	1	0	0	1	1
	2	0	1	0.99	2	2.495
	3	0	0.99	1	3	2.505
Independent partial model	1	1	0	0	1	1
	2	0	1	0	2	2
	0	0	0	1	0	0
Dependent partial model	1	1	0	0	1	1
	2	0	1	0.99	2	1.01
	0	0	0.99	1	0	0.99

Tradeoffs

- Tradeoffs (marginal vs. conditional):
  - Intuitive: off-manifold vs. on-manifold
  - True to: model vs. data
  - Computation: easy vs. hard ← How to estimate?
- Some cite the multiple Shapley value explanations as a weakness
  - Fundamental tradeoff in the presence of correlated features

Feature removal algorithms (empirical)

$S = \{1,2\}$ $\quad\quad$ $v(S) = \mathbb{E}[f(x_S)]$



Feature removal algorithms (conditional)

Poor estimates → Empirical

Curse of dimensionality → Empirical (similarity)

Strong assumptions → Parametric assumptions

Requires a deep model → Generative models

Distribution $p(x_{\bar{S}} | x_S)$

Conditional Shapley values

Expectation $\mathbb{E}_{p(x_{\bar{S}} | x_S)}[f(x_S^e, x_{\bar{S}})]$

Separate models ← Model independent

Missingness during training ← Not post-hoc

Surrogate model ← Requires a deep model

Takeaways

- Marginal Shapley values are estimated empirically
  - Can have unbiased estimates
- Conditional Shapley values can be estimated in numerous ways
  - Generally, cannot have unbiased estimates
  - The most promising approaches require training a deep model, which can be a hurdle in an explanation pipeline

Topics

- Why explain models?
- What are Shapley values?
- What are Shapley value explanations?
- Challenge 1: Feature removal approaches
- **Challenge 2: Tractable estimation strategies**

Tractable estimation strategies

- Computing Shapley values is NP-hard in general

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} W(|S|, |D|)\big(v(S \cup \{i\}) - v(S)\big)$$

Game theory

Machine learning

Unbiased, stochastic → Approximations ⟷ Model-agnostic

Exact, faster → Assumptions ⟷ Model-specific

Roughly analogous

Tractable estimation strategies

- Computing Shapley values is NP-hard in general

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} W(|S|, |D|)\big(v(S \cup \{i\}) - v(S)\big)$$

Game theory

Machine learning

Approximations ⟷ Model-agnostic

Assumptions ⟷ $\boxed{\text{Model-specific}}$ ← Assumes model type and feature removal approach

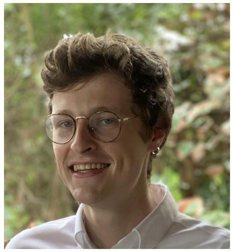
Roughly analogous

Shapley value explanations Linear models

Scott Lundberg

- Linear model $f(x) = \beta x$
- Baseline/marginal Shapley values

$$\phi_i^m(x^e) = \beta_i(x_i^e - \mu_i)$$



Joseph D. Janizek

- Conditional Shapley values

Additional assumption of normality

$$\phi_i^c(x^e) = \beta A_i \mu + \beta B_i x^e$$

  - A_i and B_i are summations over an exponential number of coalitions, which we can estimate

True to the Model or True to the Data? **Hugh Chen***, Joseph D. Janizek*, Scott Lundberg, and Su-In Lee. ICML Workshop on Human Interpretability (2020)

Shapley value explanations Tree models

- Interventional TreeSHAP
  - Exactly computes baseline and marginal Shapley values
- Path Dependent TreeSHAP
  - Approximates conditional Shapley values

Empirical (similarity)
Similarity defined by tree leafs

Scott Lundberg

Gabriel Erion

Alex DeGrave

S. Lundberg, G. Erion, **H. Chen**, A. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S. Lee. Nature Machine Intelligence (2020)

Shapley value explanations Deep models

- Approximates baseline Shapley values
  - DeepLIFT, DeepSHAP (baseline, marginal)
  - DASP (Deep Approximate Shapley Propagation)
  - ShapNets (Shapley Explanation Networks)

Requires first and second-order central moment matching

Requires using their architecture for training models which is restrictive

Tractable estimation strategies

- Computing Shapley values is NP-hard in general

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} W(|S|, |D|)\big(v(S \cup \{i\}) - v(S)\big)$$

Game theory

Machine learning

Approximations ⟷ $\boxed{\text{Model-agnostic}}$

Assumptions ⟷ Model-specific

Roughly analogous

Many estimation strategies, surprisingly they can be tied to characterizations of the Shapley value

Characterizations of the Shapley value

$$\phi(v) = \arg\min_{\beta} \sum_{S \subseteq D} W(S)\big(u(S) - v(S)\big)^2$$

$u(S) = \beta_0 + \sum_{i \in S} \beta_i$ and $W(S) = \dfrac{|D| - 1}{\binom{|D|}{|S|} |S|(|D| - |S|)}$

Least squares value
(Charnes et al 1988)

Draw subsets: $q \rightarrow E_i$

$$\phi_i(v) = \int_0^1 e_i(q)dq \,, e_i(q) = \mathbb{E}[v(E_i \cup \{i\}) - v(E_i)]$$

E_i is a random subset of $\mathrm{D} \setminus \{i\}$ with each player having probability q

Multilinear extension
(Owen 1972)

The Shapley value

Semivalue
(Dubey et al 1981)

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} P(S)\big(v(S \cup \{i\}) - v(S)\big)$$

$$P(S) = \frac{|S|!\,(|D| - |S| - 1)!}{|D|!}$$

Unbiased estimator: draw subsets from $P(S)$ and average marginal contributions

Random order value
(Shapley 1953)

$$\phi_i(v) = \frac{1}{|D|!} \sum_{\pi \in \Pi(\mathrm{D})} v\big(Pre^i(\pi) \cup \{i\}\big) - v\left(Pre^i(\pi)\right)$$

$\pi: \{1, \dots, d\} \rightarrow \{1, \dots, d\}$denotes a permutation mapping from position j to player $\pi(j)$

Draw subsets: $\pi \rightarrow Pre^i(\pi)$

Model-agnostic estimators

SGD-Shapley
(Simon & Vincent 2020)

KernelSHAP
(Lundberg & Lee 2017)

Least squares value
(Charnes et al 1988)

ME Sampling
(Okhrati and Lipani 2020)

Multilinear extension
(Owen 1972)

The Shapley value

Semivalue
(Dubey et al 1981)

ApproSemivalue
(Castro et al 2009)

Random order value
(Shapley 1953)

IME
(Strumbelj & Kononenko 2010)

ApproShapley
(Castro et al 2009)

Two main approaches

Can re-use game evaluations

Joint estimation

Efficient sampling

Feature-wise

Adaptive sampling

```
phi = np.zeros(len(D))
for _ in n_subsets:
  S = draw_subset()
  for i in D:
    phi[i] += v(S+[i])-v(S)
return phi/n_subsets
```

```
phi = np.zeros(len(D))
for i in D:
  for _ in n_subsets:
    S = draw_subset()
    phi[i] += v(S+[i])-v(S)
return phi/n_subsets
```

ApproShapley (RO)
&
ME Sampling (ME)

IME (RO)
&
ME Sampling (ME) New

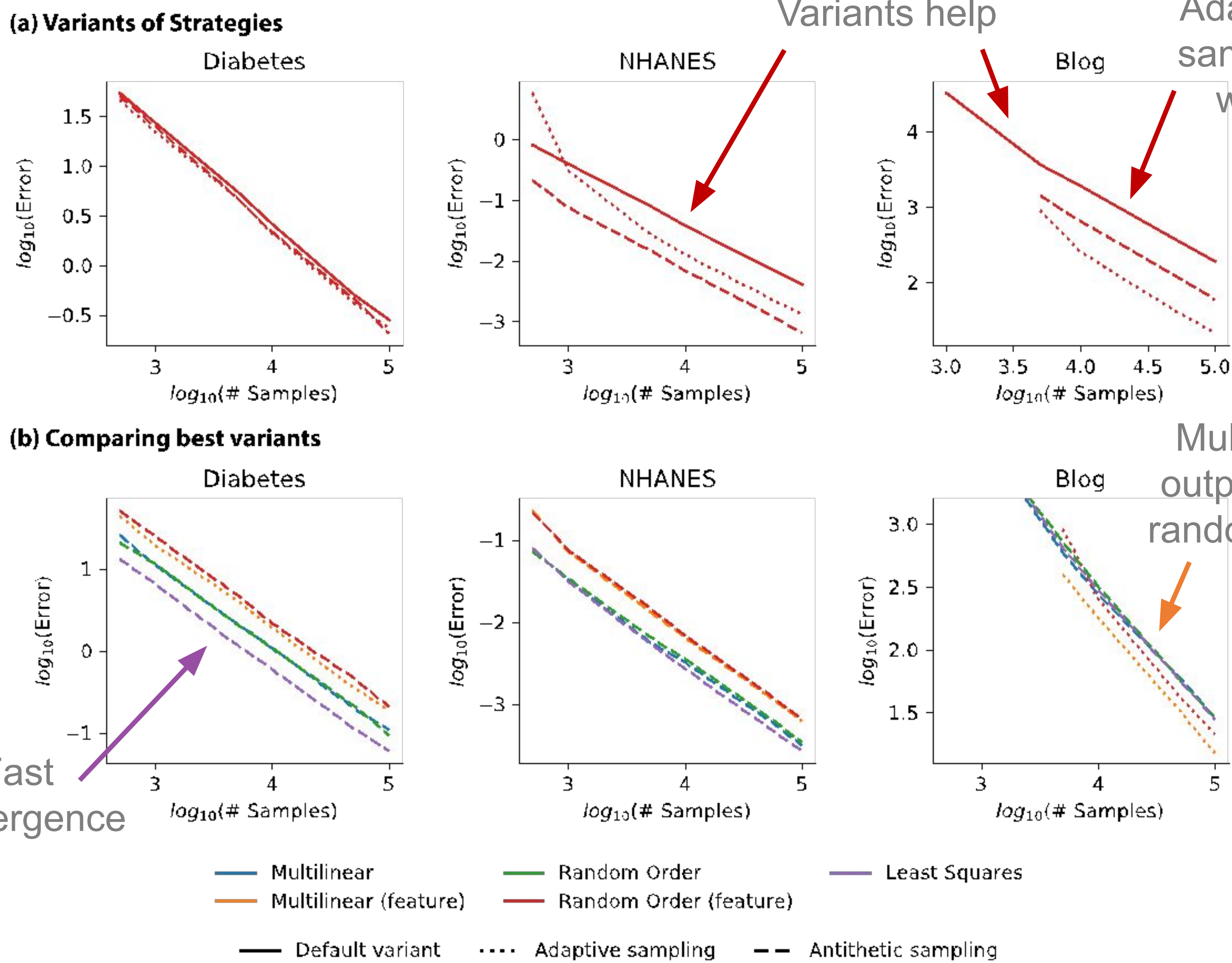
KernelSHAP (fit WLS instead)

Empirical comparisons

- How do the unbiased stochastic estimators compare in terms of convergence?
- Three datasets:
  - Diabetes (10 features)
  - NHANES I (79 features)
  - Blog (280 features)
- MSE to true baseline Shapley values for a XGB with a single explicand and baseline

Interventiona
l TreeSHAP

Comparing best variants



Variants help

Adaptive sampling wins

Multilinear outperforms random order

Fast convergence

Takeaways

- Model-specific approaches are all unique to the model type and the feature removal strategy
  - The best algorithms are for marginal Shapley values in linear and tree models
  - The others are tractable and can be useful, but more for scientific discovery/model debugging
- Model-agnostic approaches are flexible and independent of the model and removal strategy
  - They will have variance, because we typically have a fixed computational budget
  - We can estimate the variance and it is important to be aware of this

Method	Factors of complexity			Properties		
	Estimation strategy	Removal approach	Removal variant	Model-agnostic	Bias-free	Variance-free
ApproSemivalue [30]	SV	None	Exact	Yes	Yes	No
L-Shapley [26]	SV	Marginal	Empirical	Yes	No	No♣
C-Shapley [26]	SV	Marginal	Empirical	Yes	No	No♣
ApproShapley [30]	RO	None	Exact	Yes	Yes	No
IME [27]	RO	Marginal	Empirical	Yes	Yes	No
CES [22]	RO	Conditional	Empirical	Yes	No	No
Shapley cohort refinement [53]	RO	Conditional	Empirical*	Yes	No	No
Generative model [50]	RO	Conditional	Generative	Yes	No	No
Surrogate model [50]	RO	Conditional	Surrogate	Yes	No	No
Multilinear extension sampling [31]	ME	Marginal	Empirical	Yes	Yes♢	No
SGD-Shapley [54]	WLS	Baseline	Exact	Yes	No♡	No
KernelSHAP [15, 52]	WLS	Marginal	Empirical	Yes	Yes♠	No
Parametric KernelSHAP [49]	WLS	Conditional	Parametric	Yes	No	No
Nonparameteric KernelSHAP [49]	WLS	Conditional	Empirical*	Yes	No	No
FastSHAP [32]	WLS	Conditional	Surrogate	Yes	No	No
LinearSHAP [28]	Linear	Marginal	Empirical	No	Yes	Yes
Correlated LinearSHAP [28]	Linear	Conditional	Parametric	No	No	No
Interventional TreeSHAP [16]	Tree	Marginal	Empirical	No	Yes	Yes
Path dependent TreeSHAP [16]	Tree	Conditional	Empirical*	No	No	Yes
DeepLIFT [17]	Deep	Baseline	Exact	No	No	Yes
DeepSHAP [15]	Deep	Marginal	Empirical	No	No	Yes
DASP [33]	Deep	Baseline	Exact	No	No	No♣
Shallow ShapNet [34]	Deep	Baseline	Exact	No	Yes	Yes
Deep ShapNet [34]	Deep	Baseline	Exact	No	No	Yes

Practical recommendations

- For tabular data, tree models dominate
  - TreeSHAP is a mature solution
  - Often used in finance
- For structured data, off-manifold issues can be worse
  - Conditional expectation with a surrogate model
  - FastSHAP or model-agnostic estimator

Additional recommendations

- Large number of features
  - Increases computational cost
  - Feature selection may be beneficial
- Large number of samples
  - More important for model fitting
  - Conditional expectations requires many samples
- Feature correlation
  - Makes it harder to understand features
  - Larger differences between marginal and conditional
  - Causal, group, or concept explanations

Conclusion

- Aimed to make Shapley value explanation literature more accessible
  - Introduce feature attributions and Shapley values
  - Identify factors of complexity through which we can summarize and understand the literature
    - Helps contextualize existing model-specific algorithms (many of which we have developed)
    - Suggests new algorithms based on connections between existing approaches
    - Identifies future research directions and fundamental limitations