Model Explanation with Shapley Values

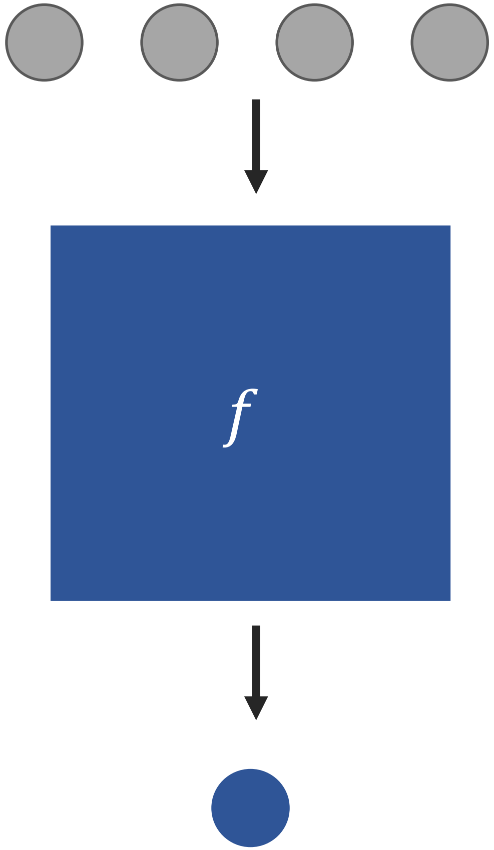
Zijing Ou
ML Group, Tencent AI Lab

2021/12/10

This Talk

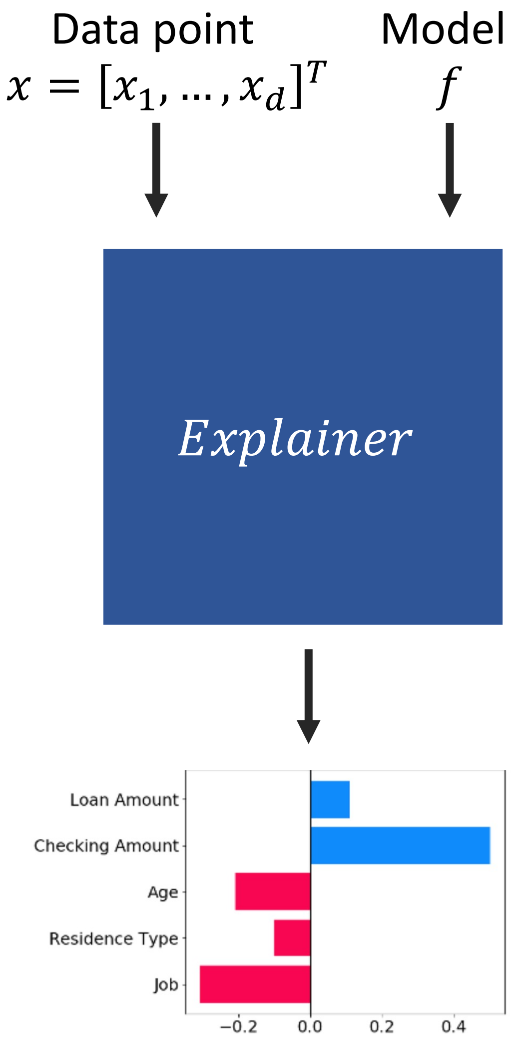
1. A brief intro: Shapley value
2. Model explanation with Shapley value
3. Shapley value estimation
4. Reliable post hoc explanations
5. Looking forward

Modern ML



- ML/AI becoming more widespread
- Black-box model now dominate
  - DNN
- Various concerns about **model transparency**

Model Explanation



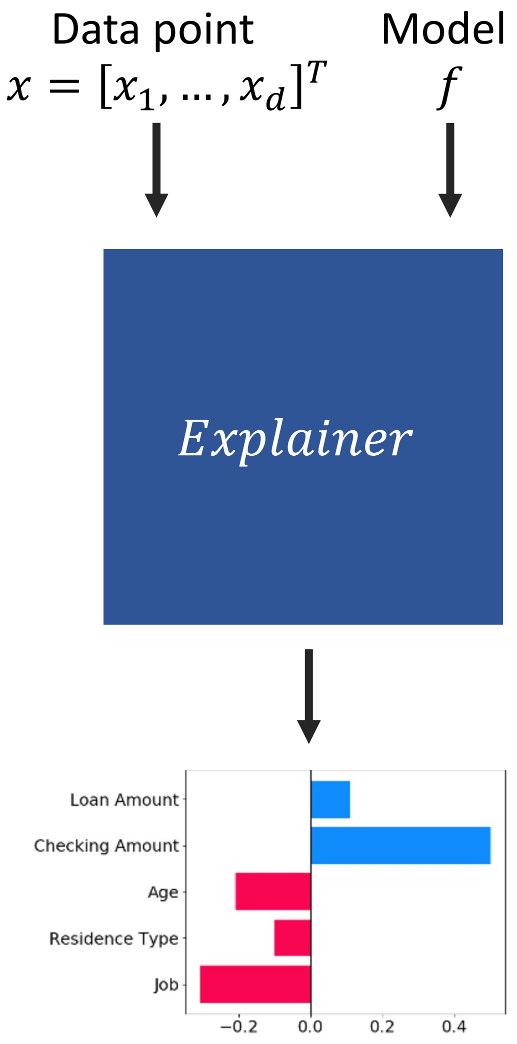
Eligible explainer:

- Model-agnostic (black-box)
- Measure the **contribution** of each feature to the output

  $$Explainer\ \phi_f : X_i \rightarrow \mathbb{R}$$

- Obey some intuitive principles

Model Explanation

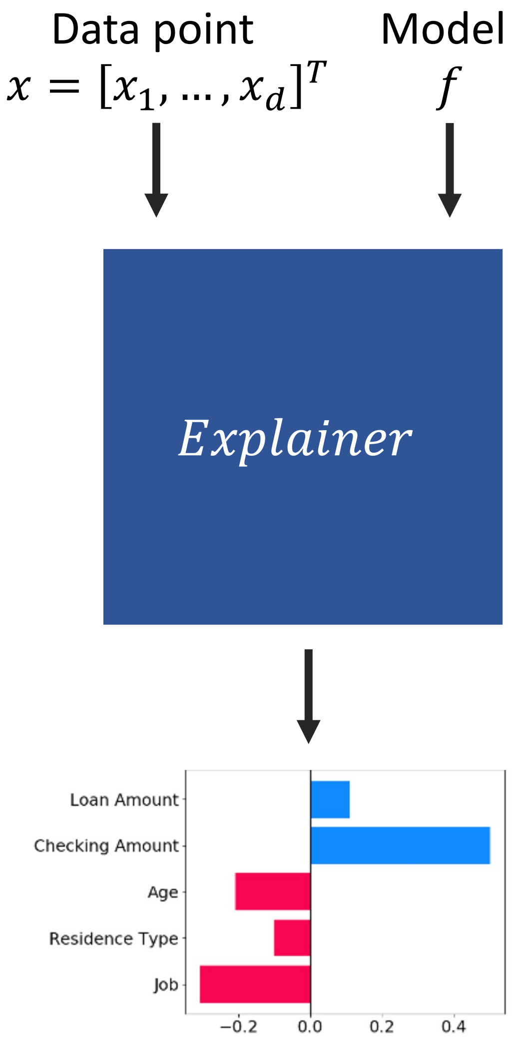


Eligible explainer ϕ_f :

- **Null**

$$if\ \forall S \subseteq D_{-i}, f(x_{S \cup i}) = f(x_S), then\ \phi_f(x_i) = 0$$

Model Explanation



Eligible explainer ϕ_f :

- **Null**

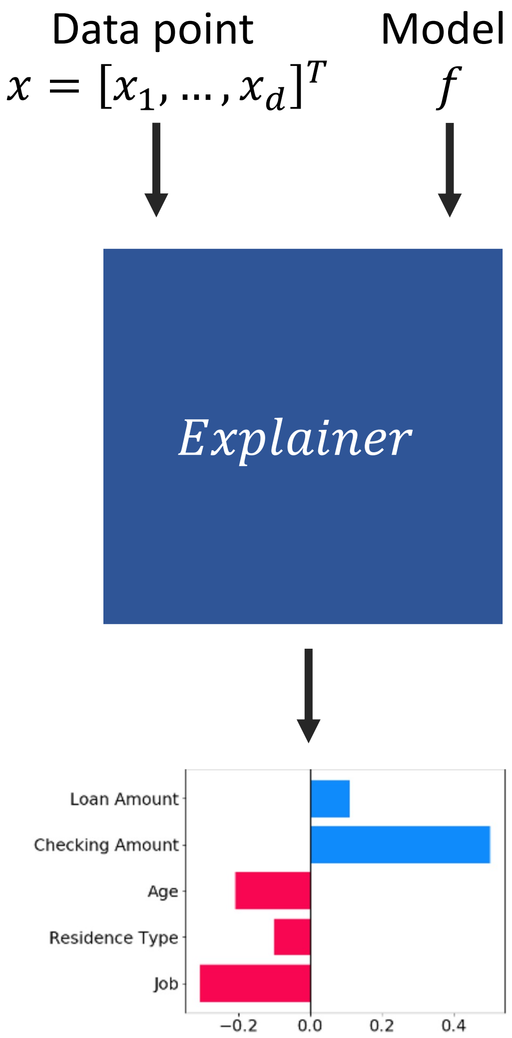
$$if\ \forall S \subseteq D_{-i}, f(x_{S \cup i}) = f(x_S), then\ \phi_f(x_i) = 0$$

- **Symmetry**

$$if\ \forall S \subseteq D_{-i,j}, f(x_{S \cup i}) = f(x_{S \cup j}),$$

$$then\ \phi_f(x_i) = \phi_f(x_j)$$

Model Explanation



Eligible explainer ϕ_f :

- **Null**

$$if\ \forall S \subseteq D_{-i}, f(x_{S \cup i}) = f(x_S), then\ \phi_f(x_i) = 0$$

- **Symmetry**

$$if\ \forall S \subseteq D_{-i,j}, f(x_{S \cup i}) = f(x_{S \cup j}),$$

$$then\ \phi_f(x_i) = \phi_f(x_j)$$

- **Marginalism**

$$if\ \forall S \subseteq D, f(x_{S \cup i}) - f(x_S) = g(x_{S \cup i}) - g(x_S),$$

$$then \phi_f(x_i) = \phi_g(x_i)$$

Shapley Value Equation

Score for feature x_i

Model output with $x_S = \{x_j | j \in S\}$

$$\phi_f(x_i) = \frac{1}{\mathrm{n}} \sum_{S \subseteq D_{-\mathrm{i}}} \binom{n-1}{|S|}^{-1} [f(x_{S \cup i}) - f(x_S)]$$

Weighted average across all subsets where $i \notin S$

Change when incorporating x_i

Lloyd Shapley. A Value for n-Person Games. *Contributions to the Theory of Games, 1953.*

Inducing model behavior $f(x_S)$ for unfixed set of features

Local SHAP

$$\phi(x_i) = \frac{1}{\mathrm{n}} \sum_{S \subseteq D_{-\mathrm{i}}} \binom{n-1}{|S|}^{-1} [v_{f,x}(S+i) - v_{f,x}(S)]$$

Conditional on fixed feature x_S

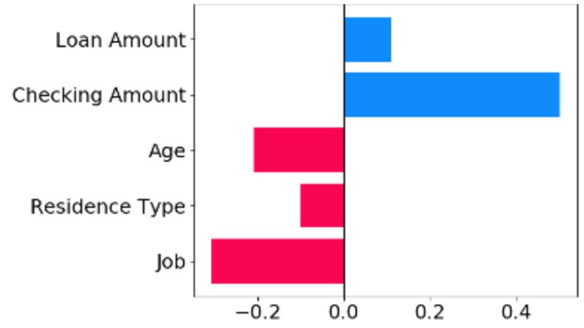
$$v_{f,x}(S) = \mathbb{E}_{p(X_{D \setminus S})}[f(X) | X_S = x_S]$$

Marginalize unfixed feature $x_{D \setminus S}$

Data point x

Loan Amount = $2,500
Checking Amount = $12,000
Age = 23
Residence Type = Apartment
Job = Startup employee

→ Local SHAP →



↑ Model f

Loss SHAP

$$\phi(x_i) = \frac{1}{\mathrm{n}} \sum_{S \subseteq D_{-\mathrm{i}}} \binom{n-1}{|S|}^{-1} [v_{f,x,y}(S+i) - v_{f,x,y}(S)]$$

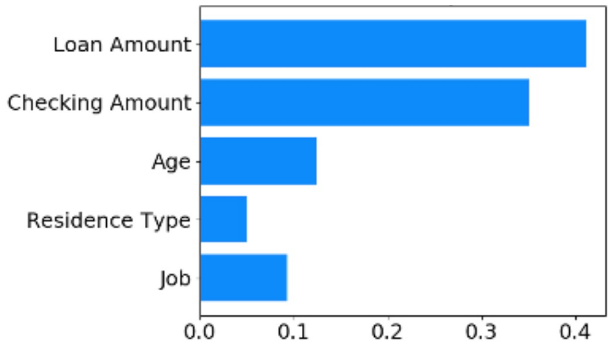
$$v_{f,x,y}(S) = -\ell(\mathbb{E}[f(X)|X_S = x_S], y)$$

Conduct on loss function $\ell(\hat{y}, y)$

Data point x

Loan Amount = \$2,500
Checking Amount = \$12,000
Age = 23
Residence Type = Apartment
Job = Startup employee

→ Loss SHAP →



Model f

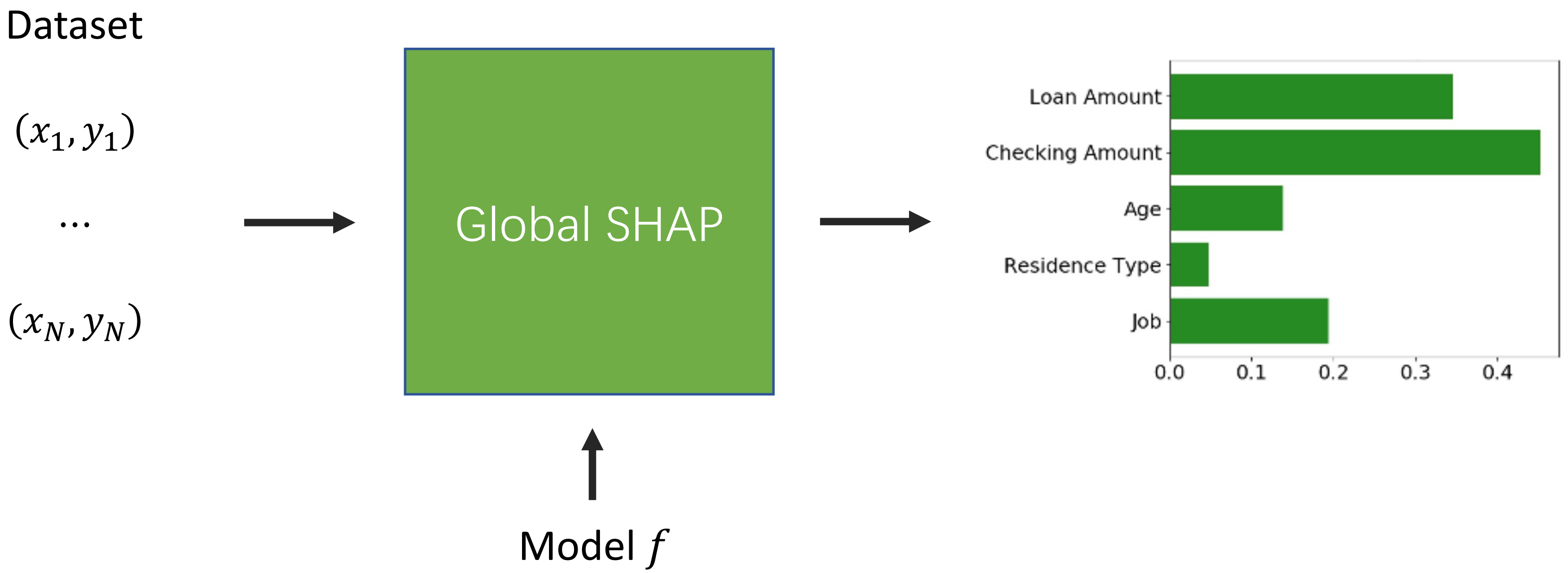
Lundberg et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nature Machine Intelligence, *2020.*

$$\phi_i = \frac{1}{\text{n}} \sum_{S \subseteq D_{-\text{i}}} \binom{n-1}{|S|}^{-1} [v_f(S+i) - v_f(S)]$$

Global SHAP

$$v_f(S) = -\mathbb{E}_{XY}[\ell(\mathbb{E}[f(X)|X_S = x_S], y)]$$

Expectation over dataset $(x, y) \sim \mathbb{p}_{XY}$



Covert et al. Understanding Global Feature Contributions With Additive Importance Measures. NIPS, *2020.*

How to estimate the Shapley value?

Castro Sampling

$$\phi_i = \frac{1}{\mathrm{n}} \sum_{S \subseteq D_{-\mathrm{i}}} \binom{n-1}{|S|}^{-1} [v(S+i) - v(S)]$$

Average (expectation) over all permutations

$$\phi_i^{cs} = \frac{1}{\mathrm{n}!} \sum_{\Pi} [v(\Pi_{:i} + i) - v(\Pi_{:i})]$$

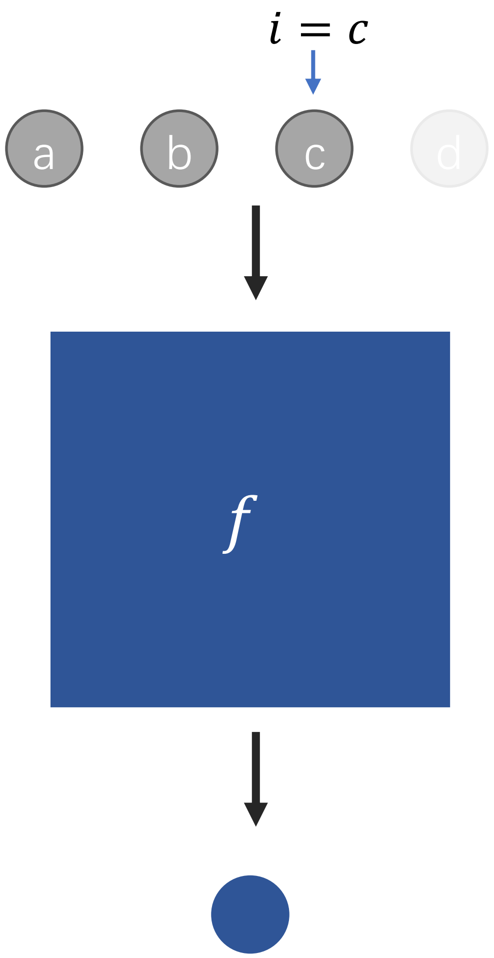
Set of features precede i

n! features permutations Π

Castro et al. Polynomial calculation of the Shapley value based on sampling. Computers & Operations Research, *2009.*

$$\phi_i = \frac{1}{n!}\sum_{\Pi}\left[v(\Pi_{:i} + i) - v(\Pi_{:i})\right]$$

Castro Sampling

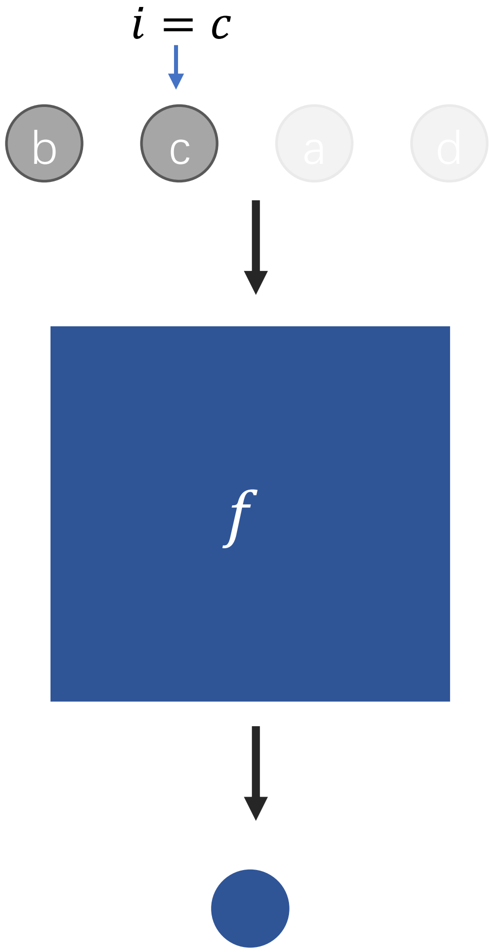


$$\phi^{(1)}_{i=c} = v(\{a, b\} + \{c\}) - v(\{a, b\})$$

Castro et al. Polynomial calculation of the Shapley value based on sampling. Computers & Operations Research, *2009.*

$$\phi_i = \frac{1}{n!}\sum_{\Pi} [v(\Pi_{:i} + i) - v(\Pi_{:i})]$$

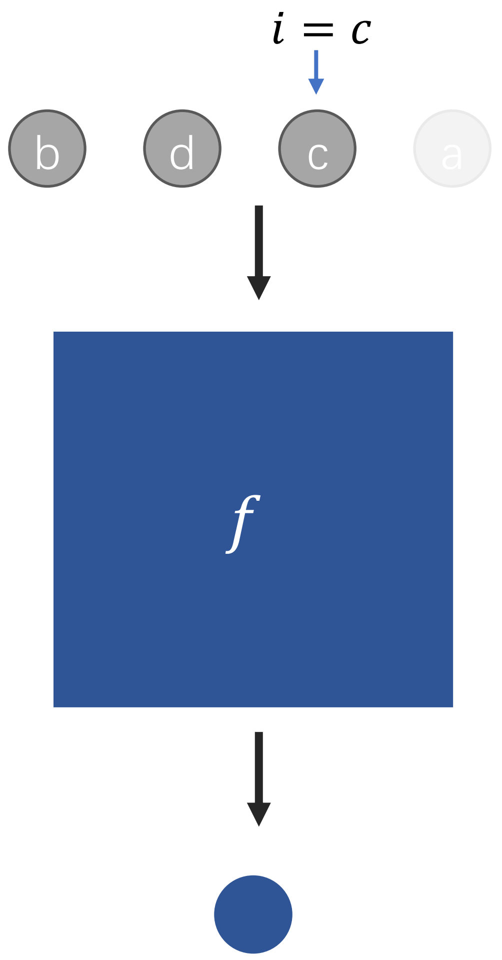
Castro Sampling



$$\phi_{i=c}^{(2)} = v(\{b\} + \{c\}) - v(\{b\})$$

$$\phi_i = \frac{1}{n!} \sum_{\Pi} [v(\Pi_{:i} + i) - v(\Pi_{:i})]$$

Castro Sampling

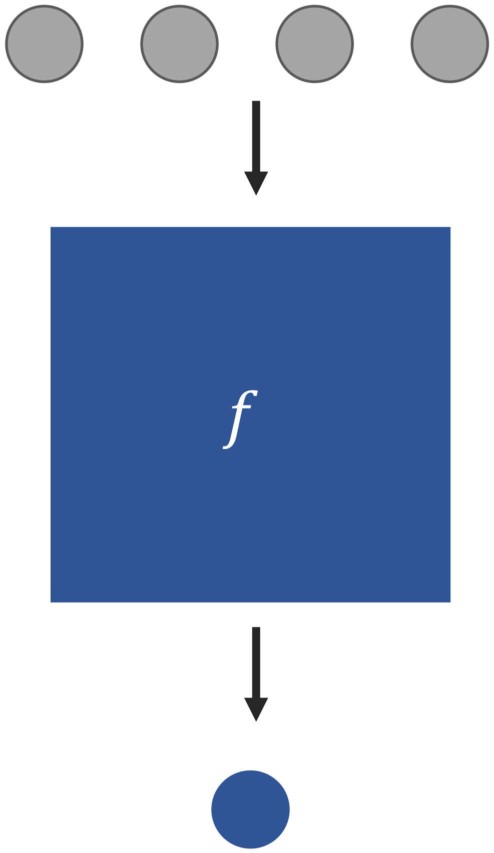


$$\phi_{i=c}^{(3)} = v(\{b, d\} + \{c\}) - v(\{b, d\})$$

Castro et al. Polynomial calculation of the Shapley value based on sampling. Computers & Operations Research, *2009.*

Castro Sampling

$$\phi_i = \frac{1}{n!}\sum_{\Pi}[v(\Pi_{:i} + i) - v(\Pi_{:i})]$$



$$\phi_{i=c} = \frac{1}{3}\left(\phi_{i=c}^{(1)} + \phi_{i=c}^{(2)} + \phi_{i=c}^{(3)}\right)$$

Castro et al. Polynomial calculation of the Shapley value based on sampling. Computers & Operations Research, *2009.*

Owen Sampling

$$\phi_i = \frac{1}{\mathrm{n}} \sum_{S \subseteq D_{-\mathrm{i}}} \binom{n-1}{|S|}^{-1} [v(S+i) - v(S)]$$

Numerical quadrature

$$\phi_i^{os} = \int_0^1 \mathbb{E}_{q(S|x\mathbf{1}, x_i=0)} [v(S+i) - v(S)] dx$$

Monte Carlo sampling

$$q(S|\boldsymbol{x}) \coloneqq \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j)$$

Okhrati et al. A Multilinear Sampling Algorithm to Estimate Shapley Values. ICPR, *2021.*

Owen Sampling

$$\phi_i^{os} = \int_0^1 \underbrace{\mathbb{E}_{q(S|x\mathbf{I},x_i=0)}[v(S+i) - v(S)]}_{g(x;i)} dx$$

Variance reduction with **antithetic sampling**

$$\phi_i^{as} = \int_0^{0.5} g(x;i) + g(1-x;i)\ dx$$

Art B. Owen. Monte Carlo Theory, Methods and Examples, chapter 8 Variance Reduction. 2013.

Owen Sampling

$$\phi_i^{os} = \int_0^1 \underbrace{\mathbb{E}_{q(S|x\mathbf{I},x_i=0)}[v(S+i) - v(S)]}_{g(x;i)} dx$$

Variance reduction with **antithetic sampling**

$$\phi_i^{as} = \int_0^{0.5} g(x;i) + g(1-x;i)\ dx$$

$$Var(\phi_i^{as}) = Var(\phi_i^{os})(1+\rho)$$

$$\rho = Corr(g(x), g(1-x))$$

Art B. Owen. Monte Carlo Theory, Methods and Examples, chapter 8 Variance Reduction. 2013.

Kernel SHAP

$$\phi_i = \frac{1}{\mathrm{n}} \sum_{S \subseteq D_{-\mathrm{i}}} \binom{n-1}{|S|}^{-1} [v(S+i) - v(S)]$$

$$p(\boldsymbol{s}) \propto \frac{d-1}{\binom{d}{\mathbf{1}^T \boldsymbol{s}} \cdot \mathbf{1}^T \boldsymbol{s} \cdot (d - \mathbf{1}^T \boldsymbol{s})}$$

$$\boldsymbol{\phi} := [\phi_1, \ldots, \phi_d]^T$$

$$\operatorname*{argmin}_{\boldsymbol{\phi}} \mathbb{E}_{p(\boldsymbol{s})} [v(\boldsymbol{s}) - v(\mathbf{0}) - \boldsymbol{s}^T \boldsymbol{\phi}]^2$$

$$s.t. \mathbf{1}^T \boldsymbol{\phi} = v(\mathbf{1}) - v(\mathbf{0})$$

$$\boldsymbol{s} := \{0,1\}^d$$

Lundberg and Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017

Kernel SHAP

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}}\ \mathbb{E}_{p(\boldsymbol{s})}[v(\boldsymbol{s}) - v(\boldsymbol{0}) - \boldsymbol{s}^T\boldsymbol{\phi}]^2$$

$$s.t.\ \boldsymbol{1}^T\boldsymbol{\phi} = v(\boldsymbol{1}) - v(\boldsymbol{0})$$

$$\widehat{\boldsymbol{\phi}}_n = \hat{A}_n^{-1}\left(\widehat{\boldsymbol{b}}_n - \boldsymbol{1}\,\frac{\boldsymbol{1}^T\hat{A}_n^{-1}\widehat{\boldsymbol{b}}_n - v(\boldsymbol{1}) + v(\boldsymbol{0})}{\boldsymbol{1}^T\hat{A}_n^{-1}\boldsymbol{1}}\right)$$

$$\hat{A}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i\boldsymbol{s}_i^T \qquad \widehat{\boldsymbol{b}}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i(v(\boldsymbol{1}) - v(\boldsymbol{0}))$$

Lundberg and Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017

Fast SHAP

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathbb{E}_{p(\boldsymbol{s})}[v(\boldsymbol{s}) - v(\boldsymbol{0}) - \boldsymbol{s}^T\boldsymbol{\phi}]^2$$

$$s.t.\, \boldsymbol{1}^T\boldsymbol{\phi} = v(\boldsymbol{1}) - v(\boldsymbol{0})$$

Neural network $\boldsymbol{\phi}_{\boldsymbol{\theta}}: \boldsymbol{X} \to \mathbb{R}^{\boldsymbol{d}}$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{x})} \mathbb{E}_{p(\boldsymbol{s})}[v(\boldsymbol{s}) - v(\boldsymbol{0}) - \boldsymbol{s}^T\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x})]^2$$

Jethani et al. FastSHAP: Real-Time Shapley Value Estimation. ICLR 2022 (under review)

Fast SHAP

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}}\ \mathbb{E}_{p(\boldsymbol{s})}[v(\boldsymbol{s}) - v(\mathbf{0}) - \boldsymbol{s}^T\boldsymbol{\phi}]^2$$

$$s.t.\ \mathbf{1}^T\boldsymbol{\phi} = v(\mathbf{1}) - v(\mathbf{0})$$

Neural network $\boldsymbol{\phi}_{\boldsymbol{\theta}}: \boldsymbol{X} \rightarrow \mathbb{R}^{\boldsymbol{d}}$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{x})} \mathbb{E}_{p(\boldsymbol{s})}[v(\boldsymbol{s}) - v(\mathbf{0}) - \boldsymbol{s}^T\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x})]^2$$

Is the estimation of Shapley value **reliable**?

Kernel SHAP Recap

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s})[v(\boldsymbol{s}) - \boldsymbol{s}^{\boldsymbol{T}}\boldsymbol{\phi}]^2$$

$$\pi(\boldsymbol{s}) \propto \frac{d-1}{\binom{d}{\mathbf{1}^T\boldsymbol{s}} \cdot \mathbf{1}^T\boldsymbol{s} \cdot (d - \mathbf{1}^T\boldsymbol{s})}$$

Kernel SHAP Recap

Data $\boldsymbol{s}$

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s})[v(\boldsymbol{s}) - \boldsymbol{s}^{\boldsymbol{T}}\boldsymbol{\phi}]^2$$

Target v

Weight $\boldsymbol{\phi}$

Kernel SHAP Recap

$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s})[v(\boldsymbol{s}) - \boldsymbol{s}^{\boldsymbol{T}}\boldsymbol{\phi}]^2$$

Kernel SHAP as linear regression models

- Dataset: $\mathcal{D} = \{\boldsymbol{v}, \boldsymbol{S}\}$
  $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \dots]^T \in \{0,1\}^{2^d \times d}$
  $\boldsymbol{v} = [v_1, v_2, \dots]^T \in \mathbb{R}^{2^d \times 1}$

- Goal: find $\boldsymbol{\phi} \in \mathbb{R}^d$ such that

$$||\boldsymbol{v} - \boldsymbol{S\phi}||^2 \approx 0$$ ← neglect $\pi(\boldsymbol{s})$ for brevity

Is the SHAP Reliable?

$$\underset{\boldsymbol{\phi}}{\text{argmin}} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s})[v(\boldsymbol{s}) - \boldsymbol{s}^T\boldsymbol{\phi}]^2$$

Kernel SHAP as linear regression models

- Dataset: $\mathcal{D} = \{\boldsymbol{v}, \boldsymbol{S}\}$
  $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \dots]^T \in \{0,1\}^{2^d \times d}$
  $\boldsymbol{v} = [v_1, v_2, \dots]^T \in \mathbb{R}^{2^d \times 1}$

- Goal: find $\boldsymbol{\phi} \in \mathbb{R}^d$ such that

$$||\boldsymbol{v} - \boldsymbol{S}\boldsymbol{\phi}||^2 \approx 0$$

Why not apply **Bayesian regression**?

lower variance ⇔ more reliable

Is the SHAP Reliable?

Kernel SHAP as Bayesian regression models

- $\boldsymbol{s} \in \{0,1\}^d$: input feature; $v \in \mathbb{R}$: output value
- Model assumes as noisy output with Gaussian noise

$$v = \boldsymbol{s}^T \boldsymbol{\phi} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \beta^{-1})$$
$$\Rightarrow \quad p(v|\boldsymbol{s}, \boldsymbol{\phi}) = \mathcal{N}(v; \boldsymbol{s}^T \boldsymbol{\phi}, \beta^{-1})$$

- Prior distribution $p(\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\phi}; \mathbf{0}, \lambda^{-1}\mathbb{I})$

Goal: find posterior $p(\boldsymbol{\phi}|\boldsymbol{S}, \boldsymbol{v})$

Is the SHAP Reliable?

$$v = \boldsymbol{s}^T \boldsymbol{\phi} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \beta^{-1})$$
$$p(\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\phi}; \boldsymbol{0}, \lambda^{-1}\mathbb{I})$$

Apply normal normal-mean conjugacy

$$p(\boldsymbol{\phi} | \boldsymbol{S}, \boldsymbol{v}) = \mathcal{N}(\boldsymbol{\phi}; \mu_n, \Sigma_n)$$

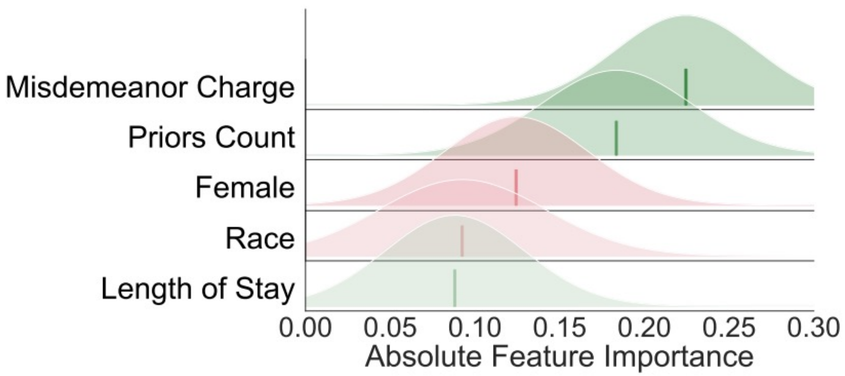
$$\mu_n = \beta \Sigma_n^{-1} \boldsymbol{S}^T \boldsymbol{v}$$ ← Mean of Shapley value

$$\Sigma_n = \lambda \mathbb{I} + \beta \boldsymbol{S}^T \boldsymbol{S}$$ ← Variance of Shapley value

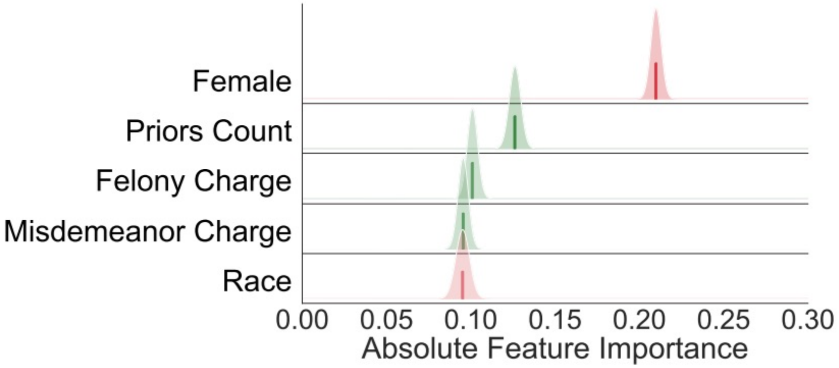
Set $\beta = \lambda = 1$, μ_n recovers the original shapley value.

Bayes SHAP

A Bayesian framework for Shapley value estimation: measure the uncertainty and reliability.



more samples



Applications:

- How many perturbations to sample (Hypothesis testing)
- How to sample for fast convergence (active learning)

Slack et al. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. NIPS, 2021.

Bayes SHAP

$$v = \boldsymbol{s}^T \boldsymbol{\phi} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbb{I}) \quad \sigma^2 \sim \text{Inv}\chi^2(n_0, \sigma_0^2)$$

Apply normal normal-inverse-chi-square conjugacy

$$p(\boldsymbol{\phi} | \boldsymbol{S}, \boldsymbol{v}, \sigma^2) = \mathcal{N}(\widehat{\boldsymbol{\phi}}; \boldsymbol{V}^{-1}\, \sigma^2)$$ ⟵ Uncertainty: estimate via sampling

$$\widehat{\boldsymbol{\phi}} = \boldsymbol{V}^{-1} \boldsymbol{S}^T \boldsymbol{S} \boldsymbol{v}$$ ⟵ Mean: recover the Shapley value

$$\boldsymbol{V} = \boldsymbol{S}^T \boldsymbol{S} + \mathbb{I}$$

Slack et al. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. NIPS, 2021.
Ou. Conjugate Bayesian analysis of common distributions. Technique report.

Looking forward

Bayes SHAP

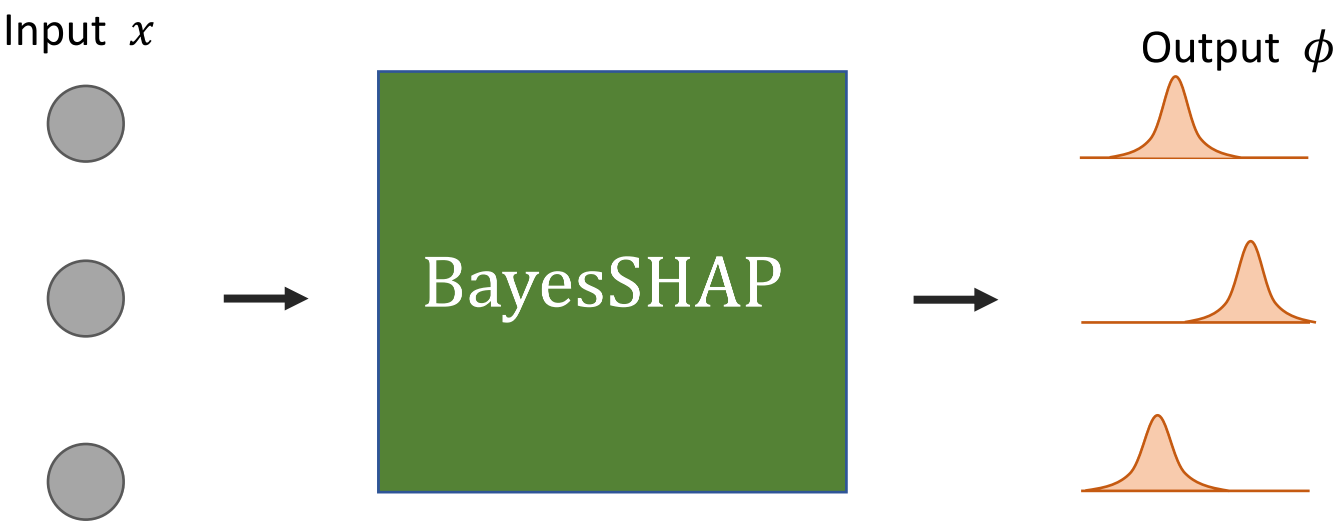
$$\underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s})[v(\boldsymbol{s}) - \boldsymbol{s}^{\boldsymbol{T}}\boldsymbol{\phi}]^2$$

Data: $(v, s) \in \mathcal{V} \times \{0,1\}^d$

Likelihood: $p_\theta(v) \coloneqq Normal(v; s^T\phi_\theta; \sigma^2)$

Prior: $p(\phi, \sigma^2)$

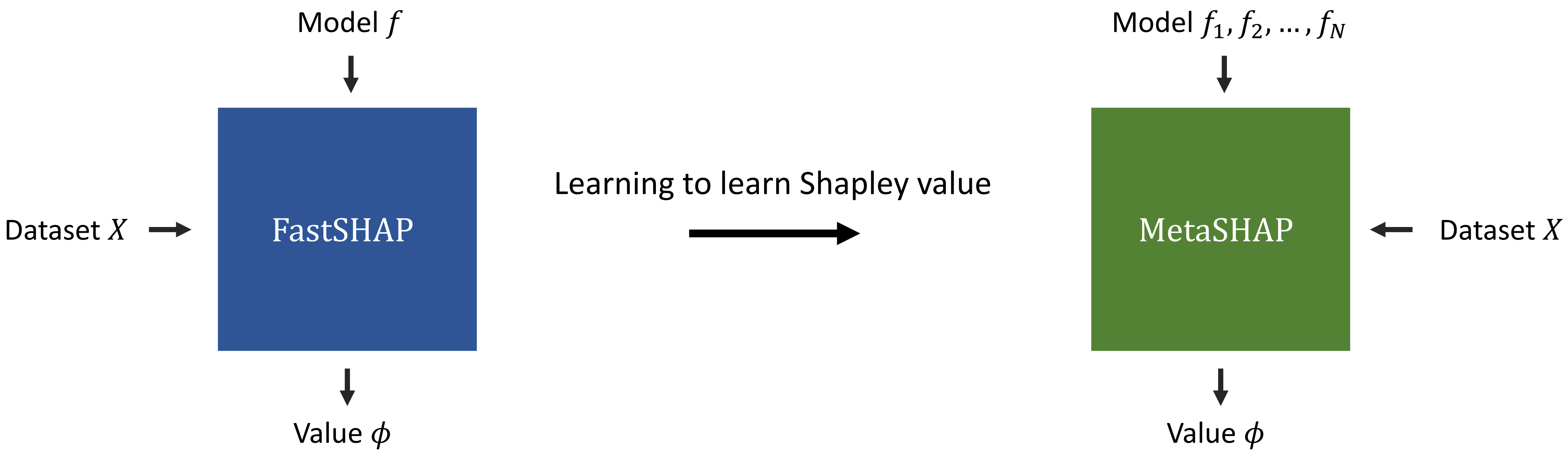
Let's try to do **Bayesian inference** for the Shapley value estimation!



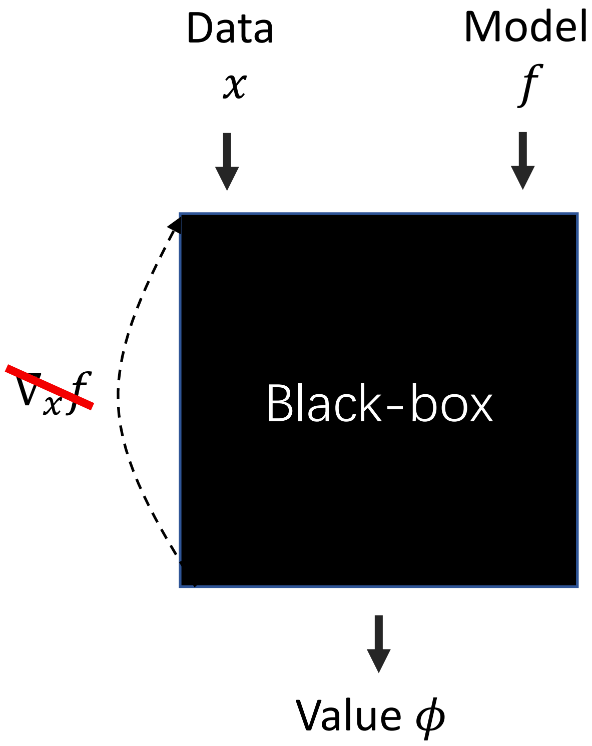
Meta SHAP

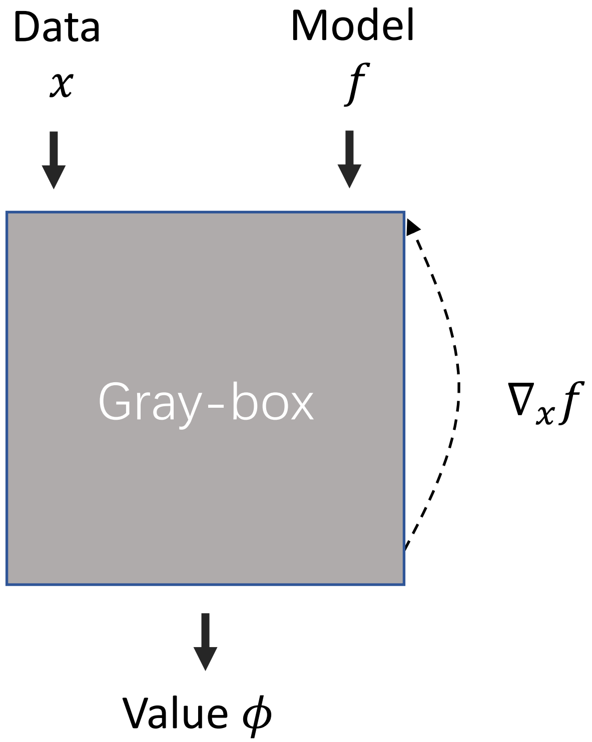
FastSHAP:
  train for each model separately

MetaSHAP:
  train once, plug and play

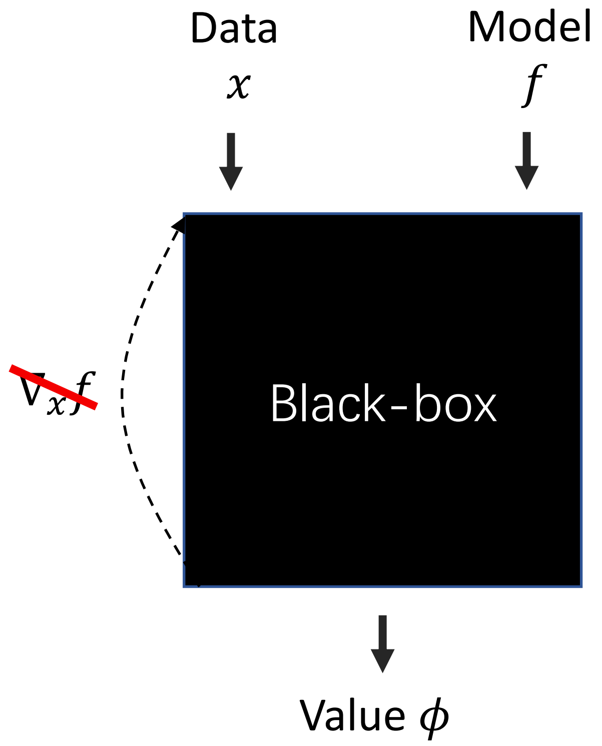


Gray SHAP


Data
x
Model
f
$\nabla_x f$
Black-box
Value ϕ


Data
x
Model
f
Gray-box
$\nabla_x f$
Value ϕ

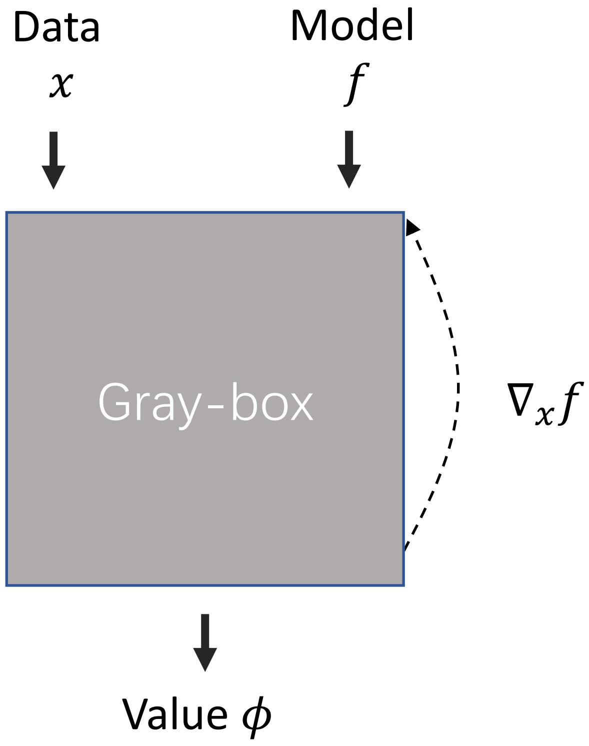
Gray SHAP



Antithetic sampling

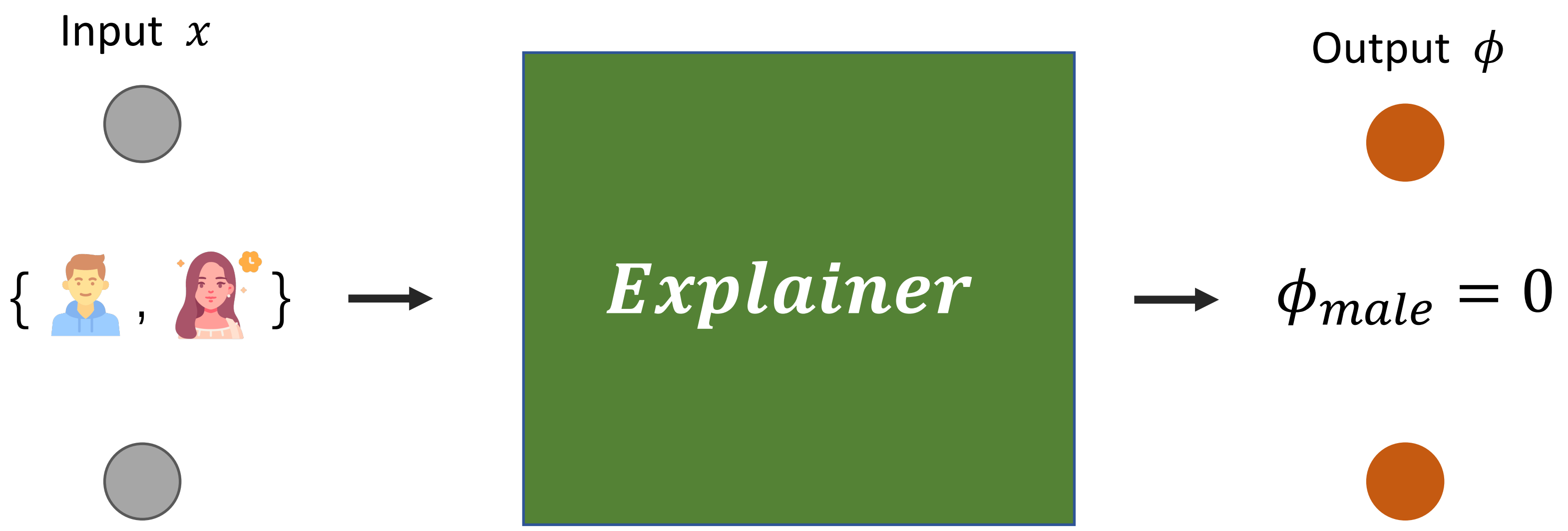
$$\phi_i^{os} = \int_0^1 \mathbb{E}_{q(S|x\mathbf{1},x_i=0)}[v(S+i) - v(S)]dx$$

Gradient guided sampling



Intuition: reducing variance by gradient!

Fair SHAP



Intuition: the value of sensitive feature is zero!

Thank you!