

A Shapley Value Proxy for Efficient Data-Set Valuation

Felipe Garrido-Lucero^{1,2}, Maxime Vono³, Benjamin Heymann³,
Patrick Loiseau¹ & Vianney Perchet^{2,3}

1 INRIA FairPlay Team

2 CREST, ENSAE

3 Criteo AI Lab

From Matchings to Markets - CIRM

December 14th 2023



Games and Artificial Intelligence Multidisciplinary Summer School 2024

GAIMSS24

- Université de Lorraine, Metz, France, **24th - 26th June 2024**
- For PhD students and Postdocs
 - (Math) **Roberto Cominetti**: **Non-Expansive Maps and Fixed Point Iterations**
 - (CS) **Panayotis Mertikopoulos**: **Online Optimization and Learning in Games**
 - (Econ) **Nicolas Vieille**: **Social Learning**
- Workshop 27th and 28th June 2024
- www.gaimss24.org or info@gaimss24.org

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

Data-set valuation

Quantify the **contribution** of players when sharing their data-sets towards solving some machine learning task

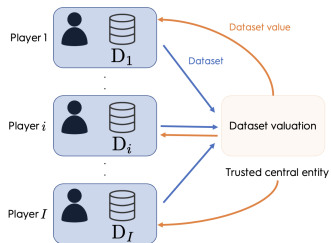


Figure: Data-set valuation problem

Applications

- **Mechanism design**: First step towards incentivizing data sharing
- **Federated learning**: Optimally/fairly agents' revenue

Approach

- Cooperative game theory
- Shapley value

- **Data-set valuation problem**

Agarwal et al. 2019, Sim et al. 2020, Tay et al. 2021

- **Shapley value** has found its place in machine learning e.g. in

- **data valuation**: Jia et al. 2019, Ghorbani et al. 2020, Kwon et al. 2021, Kwon and Zou 2022, Schoch et al. 2022

- **variable selection**: Cohen et al. 2005

- **feature importance**: Lundberg et al. 2020, Lundberg and Lee 2017, Covert et al. 2020

- **model interpretation**: Chen et al. 2019

- **Shapley value approximations**

Castro et al. 2009, Ghorbani and Zou 2019, Mitchell et al. 2022

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

Model

- Set of agents $I = \{1, \dots, I\}$
- Player i has a data-set $D_i \subset X \times Y$ from some distribution F_i
- Denote $n_i = |D_i|$ and $N = \sum_{i \in I} n_i$

Regression

$$Y_\ell = f(X_\ell) + \varepsilon_\ell,$$

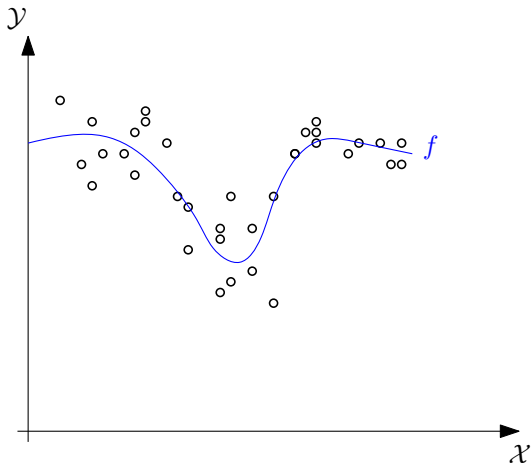
ε_ℓ is white noise and f is unknown

Goal

- To estimate $f(x) = \mathbb{E}[Y \mid X = x]$

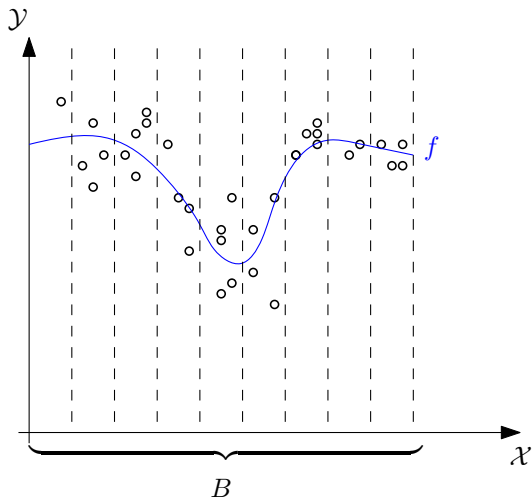
BINNING METHOD - REGRESSOGRAM

Let B N be fixed



BINNING METHOD - REGRESSOGRAM

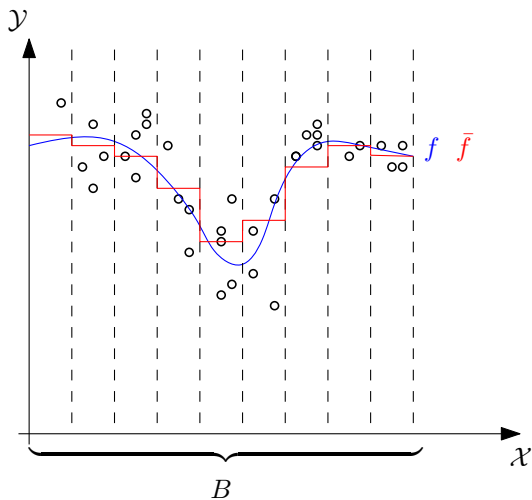
Let $B \subset \mathbb{N}$ be fixed. $D_i = \bigcup_{b \in [B]} D_i^b$, with $D_i^b \cap D_i^{b'} = \emptyset$ for any two bins b and b'



BINNING METHOD - REGRESSOGRAM

Risk - $E_x [(\hat{f}(x) - f(x))^2] = E_x [(\hat{f}(x) - \bar{f}(x))^2] + E_x [(\bar{f}(x) - f(x))^2]$

$$\bar{f} = \sum_{b \in [B]} \bar{f}_b(x), \quad \bar{f}_b(x) = E_x[f(x) / x \in b]$$



- With this in mind,

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] = \underbrace{\mathbb{E}[(\hat{f}(x) - \bar{f}(x))^2]}_{\text{Estimation error}} + \underbrace{\mathbb{E}[(\bar{f}(x) - f(x))^2]}_{\text{Approximation error}}$$

- Estimation error B and N
 - Approximation error B
- } - Trade-off in B
- Approximation error does not depend on $(D_i)_{i=1}^N$ but only on B

Data-set valuation

Quantify the contribution of players when sharing their data-sets towards solving some ML task decreasing the estimation error

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

- Set of players $I = \{1, \dots, I\}$
- Value function $v : 2^I \rightarrow \mathbb{R}, S \subseteq I,$

$$v(S) = -\mathbb{E}_x[(\hat{f}_S(x) - \bar{f}(x))^2] = -\sum_{b \in [B]} \mathbb{E}_x^b[(\hat{f}_{S,b}(x) - \bar{f}_b(x))^2] =: \sum_{b \in [B]} v_b(S)$$

- We measure the agents' contribution in each bin
- For a fixed bin $b \in [B]$ we suppose an homogeneous distribution,

$$v_b(S) = w_b(n_S^b), \text{ where } n_S^b = \sum_{i \in S} n_i^b$$

- In linear regression ($X = \mathbb{R}^d, \sigma_\varepsilon = \mathbb{E}[\varepsilon_\ell^2]$),

$$w_b(n_S^b) = \frac{-d\sigma_\varepsilon^2}{n_S^b - (d+1)}$$

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

- Classical solution concept in cooperative game theory
- **Average marginal contribution** of player i to all subcoalitions $S \subseteq I \setminus \{i\}$
- Given $v : 2^I \rightarrow \mathbb{R}$, the **Shapley value** of player i is

$$\varphi_i(v) = \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v(S \cup \{i\}) - v(S)]$$

- Classical solution concept in cooperative game theory
- **Average marginal contribution** of player i to all subcoalitions $S \subseteq I \setminus \{i\}$
- Given $v : 2^I \rightarrow \mathbb{R}$, the **Shapley value** of player i is

$$\begin{aligned} \varphi_i(v) &= \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v(S \cup \{i\}) - v(S)] \\ &= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v_b(S \cup \{i\}) - v_b(S)] \end{aligned}$$

- Classical solution concept in cooperative game theory
- **Average marginal contribution** of player i to all subcoalitions $S \subseteq I \setminus \{i\}$
- Given $v : 2^I \rightarrow \mathbb{R}$, the **Shapley value** of player i is

$$\begin{aligned}
 \varphi_i(v) &= \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v(S \cup \{i\}) - v(S)] \\
 &= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v_b(S \cup \{i\}) - v_b(S)] \\
 &= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [w_b(n_S^b + n_i^b) - w_b(n_S^b)]
 \end{aligned}$$

- Classical solution concept in cooperative game theory
- **Average marginal contribution** of player i to all subcoalitions $S \subseteq I \setminus \{i\}$
- Given $v : 2^I \rightarrow \mathbb{R}$, the **Shapley value** of player i is

$$\begin{aligned}
 \varphi_i(v) &= \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v(S \cup \{i\}) - v(S)] \\
 &= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [v_b(S \cup \{i\}) - v_b(S)] \\
 &= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq I \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} [w_b(n_S^b + n_i^b) - w_b(n_S^b)] \\
 &=: \sum_{b \in [B]} \varphi_i^b(w_b)
 \end{aligned}$$

- We compute each local Shapley value $\varphi_i^b(w_b)$ (**intractable**)

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

- Take $B = 1$. The (local) Shapley value can be rewritten as,

$$\varphi_i(w) = E_{K \sim U(\{0,1,\dots,I-1\})} \left[E_{S \sim U(2_K^I \setminus \{i\})} [w(n_S + n_i) - w(n_S)] \right]$$

- What is the distribution of $(n_S)_{S \sim U(2_K^I \setminus \{i\})}$?

- Take $B = 1$. The (local) Shapley value can be rewritten as,

$$\varphi_i(w) = \mathbb{E}_{K \sim U(\{0,1,\dots,I-1\})} \left[\mathbb{E}_{S \sim U(2^{I \setminus \{i\}})} [w(n_S + n_i) - w(n_S)] \right]$$

- What is the distribution of $(n_S)_{S \sim U(2^{I \setminus \{i\}})}$?

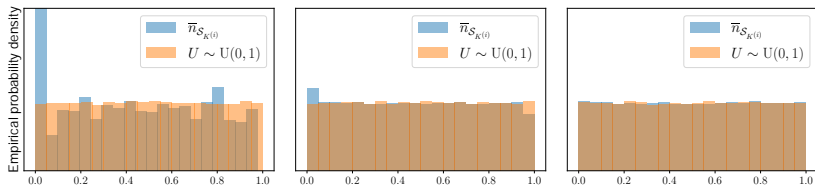


Figure: (left) $I = 10$, (middle) $I = 50$, (right) $I = 500$. 10^5 samples for each random variable and a number of data points per player drawn from $U([100])$. \bar{n}_{S_K} stands for n_{S_K} normalised.

Theorem Let $n_{S_K} := \sum_{j \in S_K} n_j$, where $S_K \in \mathcal{U}(2^{I \setminus \{i\}})$ and $K \in \mathcal{U}(\{0, \dots, I-1\})$. Then,

$$\frac{n_{S_K}}{\sum_{j \in I \setminus \{i\}} n_j} \stackrel{I}{\sim} \mathcal{U}([0, 1])$$

Theorem Let $n_{S_K} := \sum_{j \in S_K} n_j$, where $S_K \in \mathcal{U}(2^{I \setminus \{i\}})$ and $K \in \mathcal{U}(\{0, \dots, I-1\})$. Then,

$$\frac{n_{S_K}}{\sum_{j \in I \setminus \{i\}} n_j} \stackrel{I}{\sim} \mathcal{U}([0, 1])$$

Definition Discrete uniform Shapley

$$\psi_i := \frac{1}{I} \sum_{k=0}^{I-1} [w(k\mu_{-i} + n_i) - w(k\mu_{-i})], \quad \mu_{-i} := \frac{1}{I-1} \sum_{j \in I \setminus \{i\}} n_j$$

Theorem Let $n_{S_K} := \sum_{j \in S_K} n_j$, where $S_K \in \mathcal{U}(2^{I \setminus \{i\}})$ and $K \in \mathcal{U}(\{0, \dots, I-1\})$. Then,

$$\frac{n_{S_K}}{\sum_{j \in I \setminus \{i\}} n_j} \stackrel{I}{\sim} \mathcal{U}([0, 1])$$

Definition Discrete uniform Shapley

$$\psi_i := \frac{1}{I} \sum_{k=0}^{I-1} [w(k\mu_{-i} + n_i) - w(k\mu_{-i})], \quad \mu_{-i} := \frac{1}{I-1} \sum_{j \in I \setminus \{i\}} n_j$$

Theorem Whenever $w \in C^2$ is increasing and $|w''(x)|/x^2 \leq w$,

$$|\varphi_i - \psi_i| \leq \frac{w}{2I\mu_{-i}} (9(1 + \ln(I))\sigma_{-i}^2 + 2R_{-i}\tau_{-i})$$

$$R_{-i} = \max_{j \in I \setminus \{i\}} |n_j - \mu_{-i}|, \quad \tau_{-i} = \max_{j \in I \setminus \{i\}} n_j / \min_{j \in I \setminus \{i\}} n_j$$

- Linear regression satisfies assumptions

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

- **DU-Shapley** ψ_i needs I function valuations
- **Monte Carlo** $\hat{\varphi}_i$, to achieve $P(|\varphi_i(w) - \hat{\varphi}_i(w)| \leq \varepsilon) \geq 1 - \delta$, needs

$$T_{\text{perm}}(\varepsilon, \delta) = \frac{2r_v^2 I}{\varepsilon^2} \log\left(\frac{2I}{\delta}\right), \quad r_v := \max_{S_1, S_2} \{v(S_1) - v(S_2)\}$$

DU-SHAPLEY VS MONTE CARLO BASED METHODS

- DU-Shapley ψ_i needs I function valuations
- Monte Carlo $\hat{\varphi}_i$, to achieve $P(|\varphi_i(w) - \hat{\varphi}_i(w)| \leq \varepsilon) \geq 1 - \delta$, needs

$$T_{\text{perm}}(\varepsilon, \delta) = \frac{2r_v^2 I}{\varepsilon^2} \log\left(\frac{2I}{\delta}\right), \quad r_v := \max_{S_1, S_2} \{v(S_1) - v(S_2)\}$$

- Fixing $T_{\text{perm}} = I$,

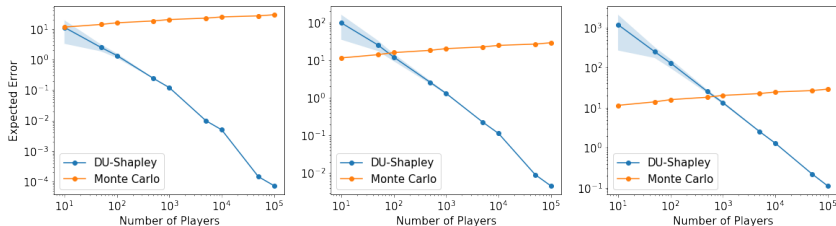
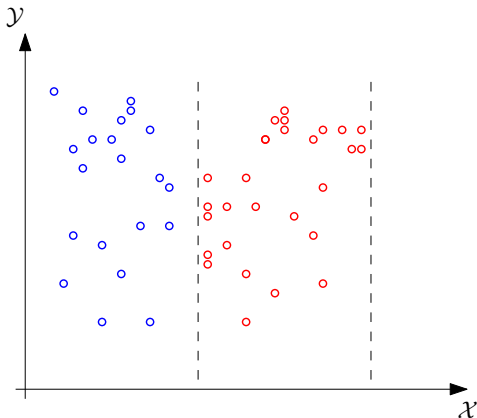


Figure: For each value of I , we drew 100 times the data points of each player from $U([n_{\max}])$, with (left) $n_{\max} = 10^2$, (center) $n_{\max} = 10^3$, and (right) $n_{\max} = 10^4$.

DU-SHAPLEY VS MONTE CARLO BASED METHODS (2)



Linear regression - $w_b(n_S^b) = \frac{d\sigma_\varepsilon^2}{d+1-n_S^b}$ $w_r(n_S^r) = \frac{d\sigma_\varepsilon^2}{d+1-n_S^r}$

DU-SHAPLEY VS MONTE CARLO BASED METHODS (3)

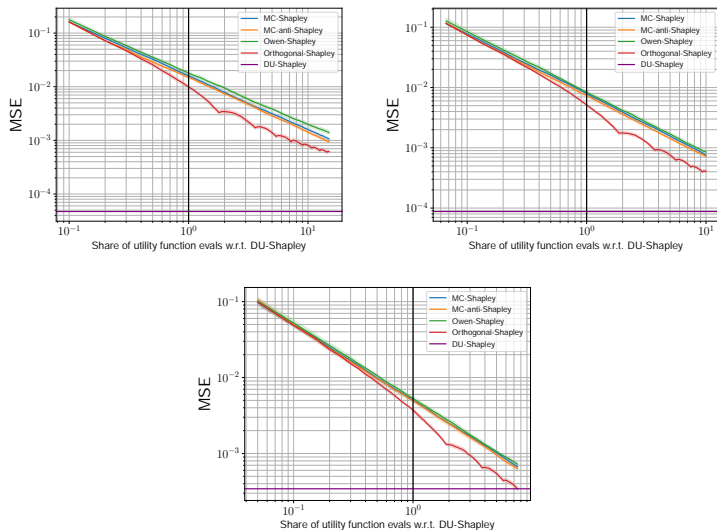


Figure: DU-Shapley vs MC-based approximations on synthetic datasets. $I = 10$, $I = 15$, and $I = 20$. $n_i^b, n_i^r \sim U(\{20, \dots, 10^3\})$.

OUTLINE

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

Conclusions

- We study **data-set valuation** problem
- Mixing ML and GT we model this problem as the sum of cooperative games
- We have an **efficient Shapley value approximation**
- We have **theoretical guarantees** for our method
- Our method **outperforms** state of the art Monte Carlo approximation schemes

Future work

- Study the heterogeneity per bin
- Design mechanism to incentivise data-sharing

Thanks

www.gaimss24.org