Improving Cooperative Game Theory-based Data Valuation via Data Utility Learning

Tianhao Wang, Yu Yang, Ruoxi Jia

Princeton University, Xi'an Jiaotong University, Virginia Tech

TL; DR: we propose to boost the efficiency in computing cooperative game theory-based data value notions by learning to estimate the performance of a learning algorithm on unseen data combinations.

Background: Data Valuation

- Goal: quantify the contribution of each training data point to a learning task.
- Example of Applications: inform the implementation of policies; filter out poor quality data and identify data sources that are important to collect in the future.

Cooperative Game Theory-based Data Valuation

- Cooperative Game: a set of players $N = \{1, \dots, n\}$, a characteristic function $v: 2^N \rightarrow R$ assigns a value to every subset $S \subseteq N$.
- In data valuation: N is dataset, each player $i \in N$ is a data point; v takes a data subset as input, and output the performance score (e.g., test accuracy) of a learning algorithm trained on the data subset (we call v data utility function).
- **Shapley Value**:

$$\phi(v)_i = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \frac{1}{\binom{n-1}{|S|}} [v(S \cup \{i\}) - v(S)]$$

- **Least Core**:

$$\min_{\psi} e \quad \text{s.t.} \sum_{i=1}^{n} \psi_i = v(N), \sum_{i \in S} \psi_i + e \geq v(S), \forall S \subseteq N$$

- The fairness properties of SV and LC provide strong motivation for using them in data valuation.
- But the exact computation of SV and LC is NP-hard in general!

Sampling-based SV/LC Estimation Heuristics

- **Heuristic sampler** takes a dataset N and outputs a set of utility samples $\{(S_i, v(S_i))\}_{i=1}^{m}$ where each $S_i \subseteq N$ sampled according to certain distributions.
- **Heuristic estimator** takes the utility samples and computes the estimation of the corresponding solution concept (i.e., Shapley or Least core).
- Example-1: Permutation Sampling estimator for Shapley value.

$$\hat{\phi}(v)_i = \frac{1}{m_{\text{perm}}} \sum_{j=1}^{m_{\text{perm}}} [v(P_i^{\pi_j} \cup \{i\}) - v(P_i^{\pi_j})]$$

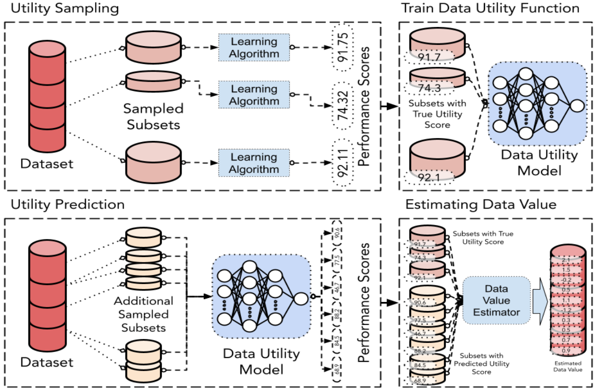
- Example-2: Monte Carlo estimator for Least Core.

$$\min_{\psi} e \quad \text{s.t.} \sum_{i=1}^{n} \psi_i = v(N), \sum_{i \in S_j} \psi_i + e \geq v(S_j), j = 1, \dots, m_{\text{train}} - 1$$

- Require many model retraining for a decent estimation accuracy.

Boosting Sampling-based Heuristics via Utility Function Learning

With the utility samples, we can potentially use a parametric model $\hat{v}$ to learn and approximate the data utility function v.
=> We can sample additional subsets with the *same* distribution followed by the heuristic sampler much more efficiently!



Theory: Shapley / Least Core Estimation Under Noisy Evaluation of Utilities

Motivation: Since the trained parametric function $\hat{v}$ may not fully recover v, we investigate the reliability of SV and LC estimated from a hybrid of m_{train} clean samples from v and m_{test} noisy samples from $\hat{v}$.

Shapley Value Estimation

1. Information-theoretic Result: one can largely improve the Shapley value estimation guarantee by computing $v(S)$ for S of very small or very large cardinality.

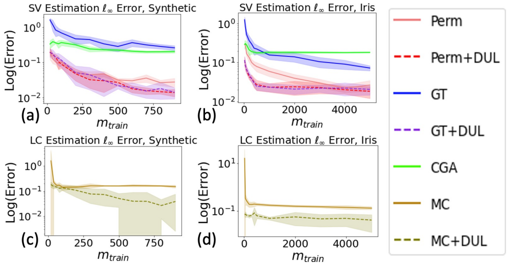
2. Result for Permutation Sampling: with smooth error distribution, the sample complexity of permutation sampling with hybrid utility samples is the same as regular permutation sampling, except for an extra irreducible error term.

Least Core Estimation

With the number of samples logarithm in the number of data points, one can still obtain a good approximation of least core with some additional irreducible error due to the error in v.

Evaluation

Estimation Error Comparison



Data Removal