The Shapley Value in Machine Learning

Sophie Greenwood · March 7, 2023 · CPSC 532L Presentation

Learning Goals

- Become **familiar with popular ML applications** of the Shapley value
- **Interpret** the Shapley value and its properties in ML contexts
- Recognize the **need for efficient computation** of the Shapley value
- Understand key **properties** of common **estimators / approximations**

Outline

- **Applications** of the Shapley value in machine learning
  - Data Valuation
  - Feature Attribution
  - Additional applications
- Efficient **computation**
- Limitations

Applications of the Shapley value in ML

(Supervised) Machine Learning

Features

Label

Observations

i	Name	Degree	Position	Has dog named Peach?	Cooperate in PD?
1	Ruiyu	MSc	Student	N	N
2	Lironne	MSc	Student	Y	Y
3	Sophie	BSc	Student	N	N
4	Narun	MSc	TA	N	Y
5	Prayus	BSc	Student	N	Y
	Shruthi	MSc	Student	N	?

Data Attribution

Data Attribution

- Identify the **contribution of a data point** to performance
- Applications:
  - Data valuation/pricing
  - Assessing data quality
  - Identifying poisoned or mislabeled data
  - Explaining the model
- Several existing approaches: leave-one-out, influence functions
- Most recently: **data Shapley**

Data Attribution – a coalitional game!

- Players (N): training set
- Characteristic function (v):

$$v(S) = \text{performance of the model trained on } S \subseteq N$$

- Performance
  - measured by log-likelihood, accuracy, …
  - on a test *dataset* or an individual test *data point*

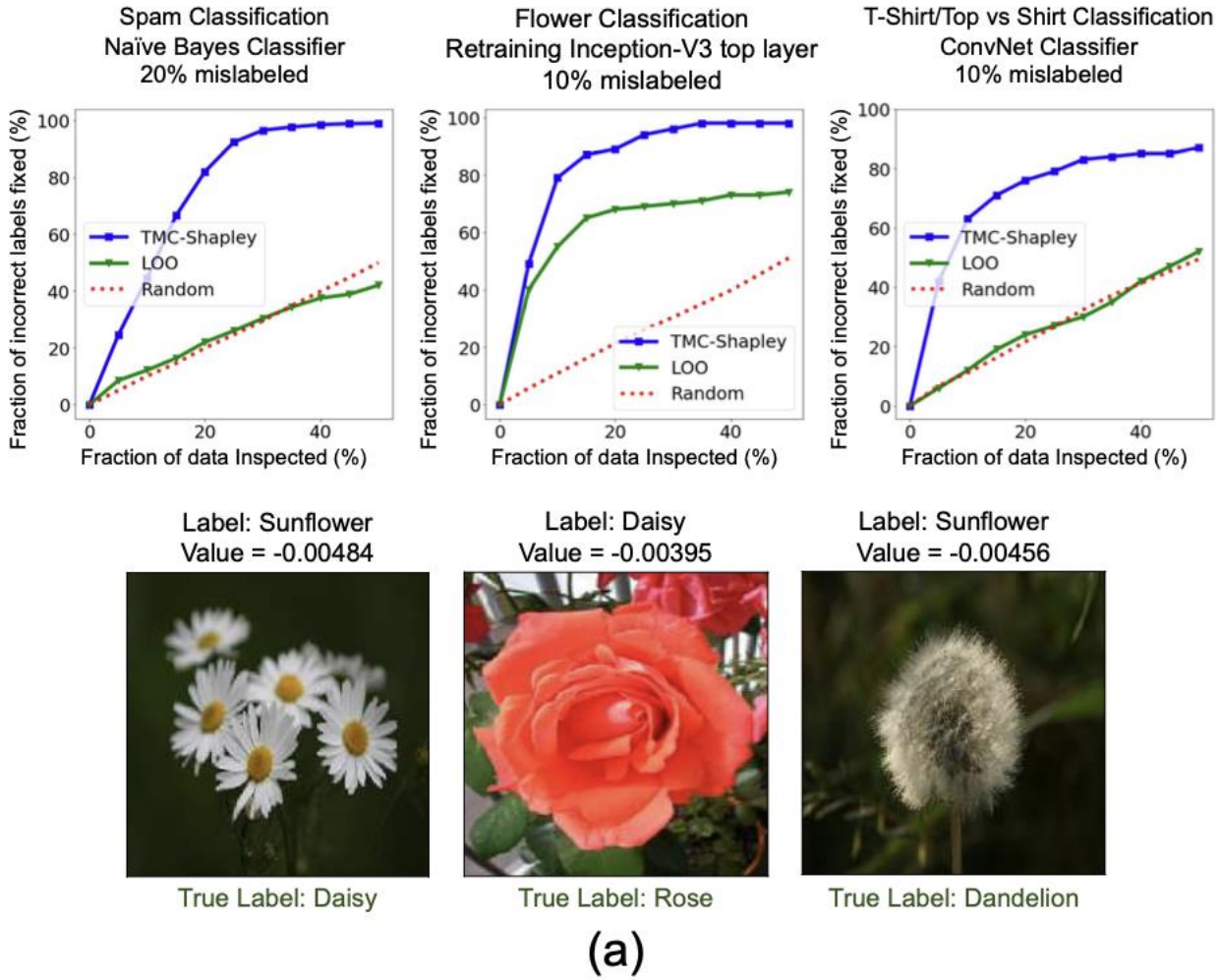
Data Shapley

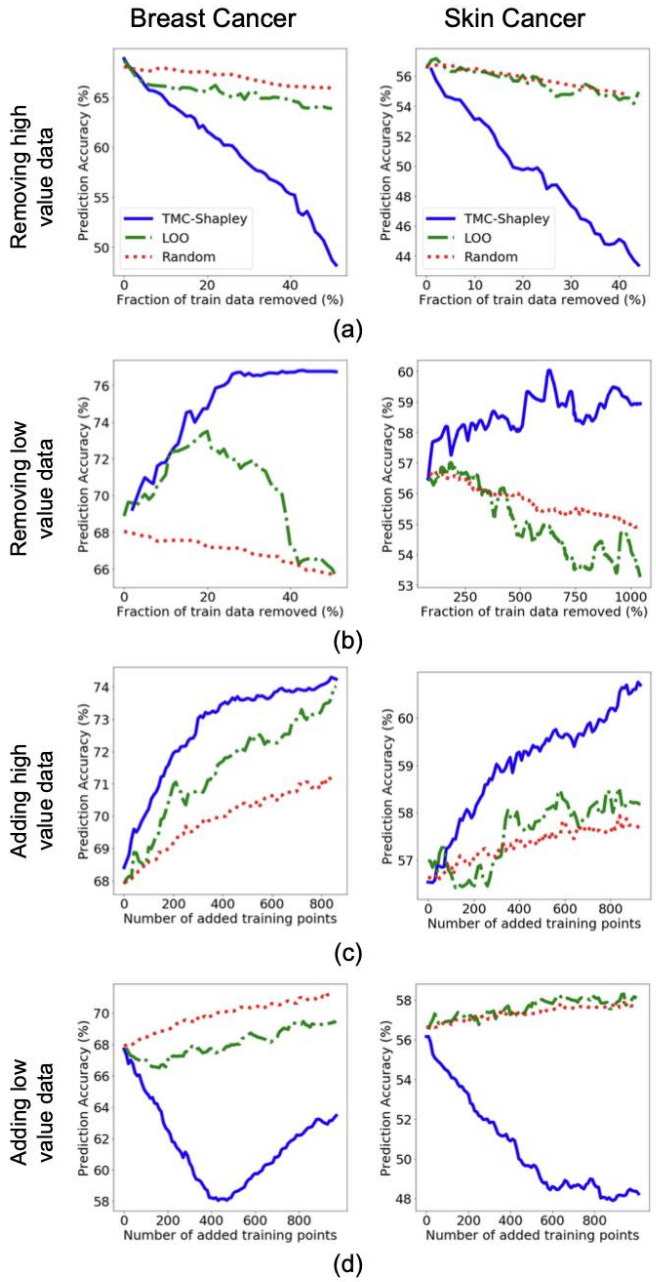
- Desirable **fairness properties**: analogues of Shapley axioms!

- Ex: Points i, j with identical contributions have attributions $\phi_i = \phi_j$

Use the Shapley value to quantify a point's contribution

Data Shapley: Results





Figures from "Data Shapley: Equitable Valuation of Data for Machine Learning" (Ghorbani and Zou 2019)

Other Solution Concepts – Beta Shapley

- Shapley values can be seen as computing

$$\phi_i = \sum_{k=1}^{n} \frac{1}{n} \Delta_i(k)$$

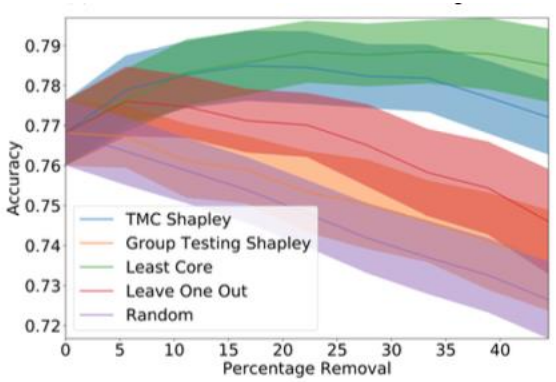
Average marginal contribution among subsets of size k

- **Semi-values:** compute

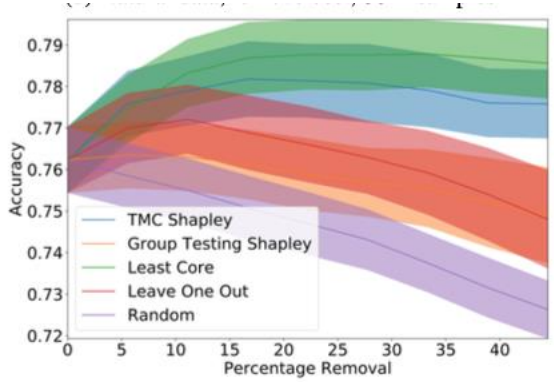
$$\phi_i = \sum_{k=1}^{n} w_k \Delta_i(k)$$

- Satisfy symmetry, dummy player, and additivity!?

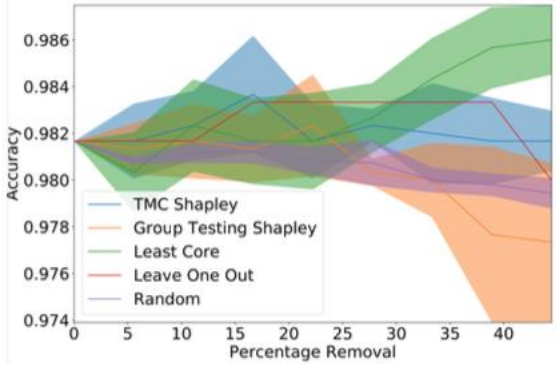
Other Solution Concepts – The Core



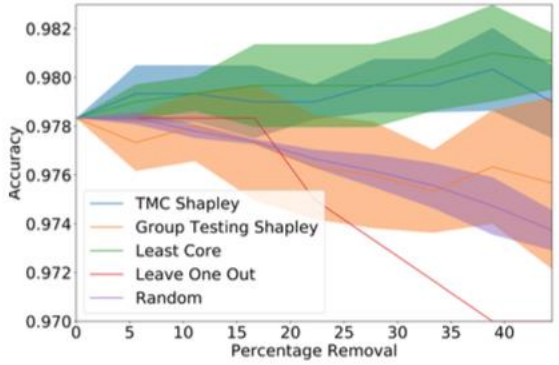
(e) Synthetic data, remove worst, $10K$ samples



(f) Synthetic data, remove worst, $50K$ samples



(g) Natural data, remove worst, $10K$ samples



(h) Natural data, remove worst, $50K$ samples

Figures from "If You Like Shapley Then You'll Love the Core" (Yan and Procaccia 2021)

Feature Attribution

Feature Attribution

- Players: **features** of the data
- Characteristic function:

$$v(S) = \text{performance of the model trained with features } S \subseteq N$$

- Similar axiomatic motivation

- Original application: Feature selection (Cohen 2007)
- Popular application: Interpretable ML (Lundberg and Lee 2017)
  - SHAP: **11,563 citations**

SHAP Applications

Interpretable and accurate fine-grained recognition via region grouping
Z Huang, Y Li - Proceedings of the IEEE/CVF Conference …, 2020 - openaccess.thecvf.com
We present an interpretable deep model for fine-grained visual recognition. At the core of our method lies the integration of region-based part discovery and attribution within a deep …

Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls
DC Feng, WJ Wang, S Mangalathu… - Journal of Structural …, 2021 - ascelibrary.org
RC shear walls are commonly used as lateral load-resisting elements in seismic regions, and the estimation of their shear strengths can become simultaneously design-critical and …

[HTML] Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility
F Wang, S Huang, R Gao, Y Zhou, C Lai, Z Li, W Xian… - Cell discovery, 2020 - nature.com
The COVID-19 pandemic has accounted for millions of infections and hundreds of thousand deaths worldwide in a short-time period. The patients demonstrate a great diversity in …

Additional Applications

Other applications

- Multi-agent RL
- Ensemble pruning
- Federated learning
- Other topics in explainable AI

Efficient Computation

Food for thought

- Data valuation:
  - Dataset size: 10 points
  - How many **models** do we need to **train**?

Food for thought

- Data valuation:
  - Dataset size: 10 points
  - How many **models** do we need to **train**?

- Need to estimate **efficiently** and **accurately**

Monte Carlo Sampling

- Sample permutation, update the Shapley values, repeat
- Many variations on this (restricted, stratified sampling, etc.)
- For fixed number of iterations, $O(|N|)$

Linear Regression

- **Trick:** SV is the solution to a weighted linear regression problem
- Find an approximate (biased) estimator in $O(|N|)$
- Unbiased estimator exists, but has **high variance**

Other

- Multilinear Extension
- Structure-specific (e.g. Voting game approximation)
- ML-specific (e.g., Gradient Shapley)
- …

Discussion & Conclusion

Limitations

- Shapley value axioms don’t necessarily hold when approximated
- Sometimes produce **incorrect results**
- More **ordinal** than **cardinal**

Conclusion

- The Shapley value is a **powerful tool** in a variety of ML problems
  - Guarantees fair solutions
  - Excellent performance in practice

- Other solution concepts are less widespread but promising

- Wide array of algorithms exist for efficient computation

References

Yongchan Kwon and James Zou. “Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning.” Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, 2022.

Amirata Ghorbani and James Zou. “Data Shapley: Equitable Valuation of Data for Machine Learning.” Proceedings of the 36th International Conference on Machine Learning, 2019.

Benedek Rozemberczki, Lauren Watson, Peter Bayer, Hao-Tsung Yang, Oliver Kiss, Sebastian Nilsson, and Rik Sarkar. “The Shapley Value in Machine Learning.” Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. *A Principled Approach to Data Valuation for Federated Learning*. In: Yang, Q., Fan, L., Yu, H. (eds) Federated Learning. Lecture Notes in Computer Science(), vol 12500. Springer International Publishing, 2020.

Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P. Friedlander, Changxin Liu, and Yong Zhang. “Improving Fairness for Data Valuation in Horizontal Federated Learning.” Proceedings of the IEEE 38th International Conference on Data Engineering (ICDE), 2022.

Tom Yan and Ariel D. Procaccia. “If You Like Shapley Then You’ll Love the Core.” Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021.

Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. “SHAQ: Incorporating Shapley Value Theory into Multi-Agent Q-Learning.” Proceedings of the 36th Conference on Neural Information Processing Systems, 2022.

Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. “Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning.” Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021.

Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. “Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments.” Proceedings of the 39th International Conference on Machine Learning, 2022.

Benedek Rozemberczki and Rik Sarkar. “The Shapley Value of Classifiers in Ensemble Games.” Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021.

Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” Proceedings of the 31st Conference on Neural Information Processing Systems, 2017.

Shay Cohen, Gideon Dror, and Eytan Ruppin. “Feature Selection via Coalitional Game Theory.” Neural Computation, 2007 Jul;19(7):1939-61.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. “FastSHAP: Real-Time Shapley Value Estimation.” Proceedings of the International Conference on Learning Representations, 2022.