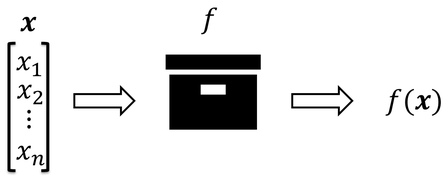
Provably Accurate Shapley Value Estimation via Leverage Score Sampling[1,2]

Christopher Musco and R. Teal Witter

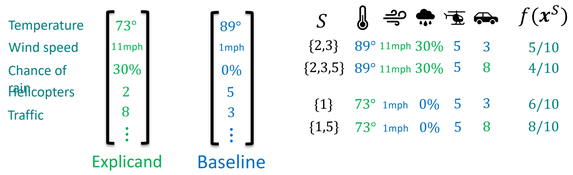
New York University

Explainable AI

Goal: Explain the output of an opaque function.



Example: How does changing traffic (an input feature) change the enjoyment of a bike ride (the output)?



Our answer will attribute the model output to each input using…

Shapley Values

For a set function $v\colon 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

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

$$= \frac{1}{n} \sum_{k \in [n-1]} \underbrace{\frac{1}{\binom{n-1}{k}} \sum_{S \subseteq [n] \setminus \{i\} : |S| = k} v(S \cup \{i\}) - v(S)}_{\text{Average over sets of size } k}$$

Average over all sizes k

We will set $v(S) = f(\boldsymbol{x}^S)$.

Regression Formulation

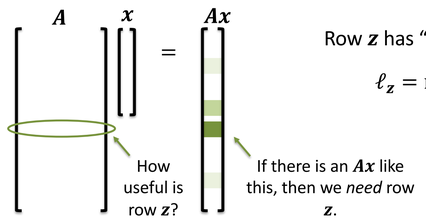
For every v, there is a matrix $\boldsymbol{A} \in \mathbb{R}^{2^n - 2 \times n}$ and $\boldsymbol{b} \in \mathbb{R}^{2^n - 2}$ so that

$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{bmatrix} = \boldsymbol{\phi} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{b}||_2 + \mathbf{1} \frac{v([n]) - v(\emptyset)}{n}$$

We can compute Shapley values by solving a regression problem!

Leverage Score Sampling

If the regression problem is too large, how should we sample?



Row $\boldsymbol{z}$ has "leverage":

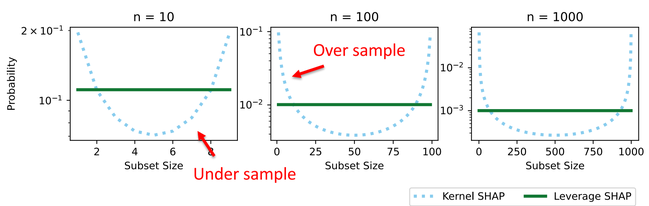
$$\ell_{\boldsymbol{z}} = \max_{\boldsymbol{x}} \frac{(\boldsymbol{Ax})^2_{\boldsymbol{z}}}{||\boldsymbol{Ax}||_2^2}$$

Leverage SHAP

Algorithm: *Leverage SHAP* estimates Shapley values by solving a regression problem sampled via leverage scores $\ell_{\boldsymbol{z}} = \binom{n}{||\boldsymbol{z}||}^{-1}$.

Leverage SHAP vs Kernel SHAP

Distribution Over Sampled Subset Size



Error Over 100 Runs

	IRIS	California	Diabetes	Adult	Correlated	Independent	NHANES	Communities
Kernel SHAP								
Mean	0.00261	0.0208	15.4	0.000139	0.00298	0.00324	0.0358	130.0
1st Quartile	5.69e-07	0.0031	3.71	1.48e-05	0.00166	0.00163	0.0106	33.5
2nd Quartile	9.52e-06	0.0103	8.19	3.86e-05	0.00249	0.00254	0.0221	53.6
3rd Quartile	0.00181	0.029	20.1	0.000145	0.00354	0.00436	0.0418	132.0
Optimized Kernel SHAP								
Mean	3.28e-14	0.00248	2.33	1.81e-05	0.000739	0.000649	0.00551	21.8
1st Quartile	2.12e-14	0.000279	0.549	2.16e-06	0.00027	0.000187	0.000707	5.85
2nd Quartile	3.55e-14	0.00138	1.26	5.43e-06	0.000546	0.000385	0.0024	13.0
3rd Quartile	4.22e-14	0.0036	3.03	1.63e-05	0.00101	0.000964	0.00665	25.1
Leverage SHAP								
Mean	3.28e-14	0.000186	0.63	5.21e-06	0.000458	0.000359	0.00385	14.7
1st Quartile	2.12e-14	1.91e-05	0.0631	6.3e-07	0.000139	9.51e-05	0.000333	3.6
2nd Quartile	3.55e-14	8.31e-05	0.328	2.33e-06	0.000376	0.000235	0.00149	8.9
3rd Quartile	4.22e-14	0.000231	0.769	7.09e-06	0.000617	0.000556	0.00401	15.3

Theoretical Guarantees

With $O\left(n \log n + \frac{n}{\epsilon}\right)$ samples, the Leverage SHAP solution $\widetilde{\boldsymbol{\phi}}$, with probability 9/10, satisfies

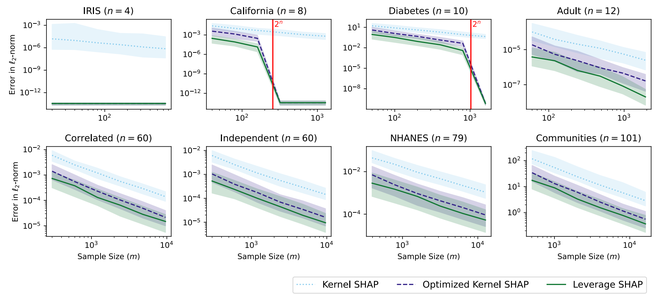
$$||\boldsymbol{A}\widetilde{\boldsymbol{\phi}} - \boldsymbol{b}||_2^2 \leq (1+\epsilon)||\boldsymbol{A}\boldsymbol{\phi} - \boldsymbol{b}||_2^2$$

and, for $\gamma = \frac{||\boldsymbol{A}\,\boldsymbol{\phi} - \boldsymbol{b}||_2^2}{||\boldsymbol{A}\,\boldsymbol{\phi}||_2^2}$,

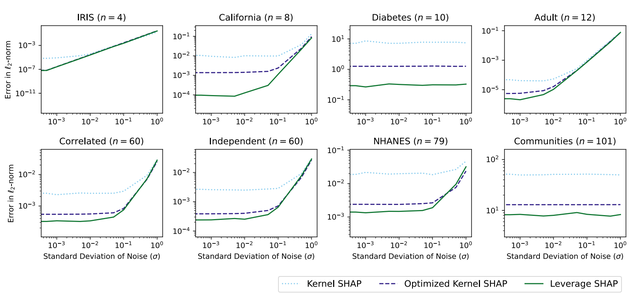
$$||\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}||_2^2 \leq \epsilon\,\gamma ||\boldsymbol{\phi}||_2^2.$$

Performance by…

Sample Size



Noise



Noise is relevant because set functions can often only be approximated e.g., $v(S) = \mathbb{E}_{\boldsymbol{x}^S}[f(\boldsymbol{x}^S)]$

Notes

[1]Our work is available on arXiv at https://arxiv.org/abs/2410.01917

[2]This work was supported by the National Science Foundation under Grant No. 2045590 and Graduate Research Fellowship Grant No. DGE-2234660