CS-Shapley: Class-wise Shapley Values for Data Valuation in Classification

Stephanie Schoch[1], Haifeng Xu[2], Yangfeng Ji[1]

[1]University of Virginia [2]University of Chicago

NEURAL INFORMATION PROCESSING SYSTEMS

Overview:

Shapley-based data valuation utilizing a value function that differentiates between in-class and out-of-class accuracy.

Background:

Data valuation methods aim to quantify the contribution of each datum to the predictive performance of a learning model.

Shapley Value:

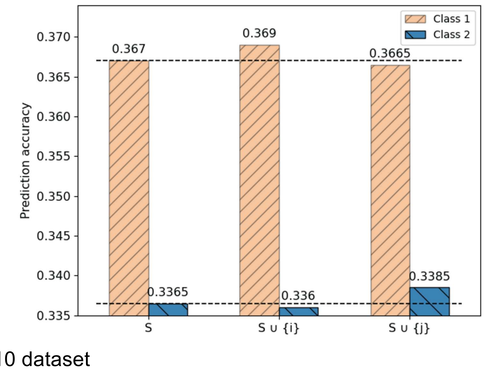
Definition: Given a value function $v(\cdot)$, training set T, and learning algorithm $\mathcal{A}$, the Shapley value for any data point i is:

$$\phi_i(T, \mathcal{A}, v) = \sum_{S \subseteq T \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

- Average marginal contribution of i to every possible subset $S \subseteq T \setminus \{i\}$.
- In data valuation, $v(\cdot)$ is typically the full development set accuracy.

Motivation:

Using full development accuracy for the value function obscures important information on ***how*** each data point contributes to training a model.



- CIFAR10 dataset
- Logistic regression classifier
- Although i and j both belong to Class 1, i increases in-class accuracy, and j decreases in-class accuracy.
- **Full development set accuracy when adding either i or j is the same.**
- **Potentially problematic for adversarial or noisy settings.**

Redesigning the Value Function:

Essentially, the redesigned value function measures contribution using in-class accuracy and uses out-of-class accuracy as a weighting (discounting) factor.

Notation:

S	Subset of training data
y_i	In-class label
$-y_i$	Out-of-class label(s)
D_{y_i}	In-class development instances
D_{-y_i}	Out-of-class development instances

In-class and out-of-class accuracy:

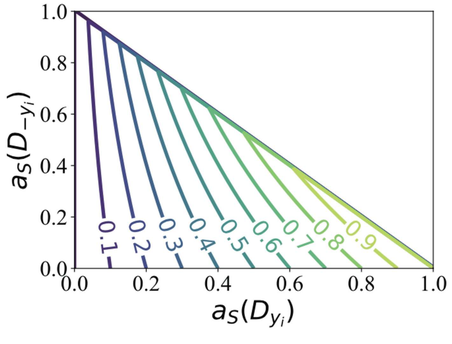
$$a_S(D_{y_i}) = \frac{\text{\# of correct predictions in } D_{y_i}}{|D|}$$

$$a_S(D_{-y_i}) = \frac{\text{\# of correct predictions in } D_{-y_i}}{|D|}$$

Value function:

$$v_{y_i}(S_{y_i}|S_{-y_i}) = a_S(D_{y_i}) \cdot e^{a_S(D_{-y_i})}$$

Prioritization of In-Class Accuracy:



Contour plot of value function v_{y_i}.

- v_{y_i} is controlled by in-class accuracy $a_S(D_{y_i})$.
- When $a_S(D_{y_i})$ is high (i.e. class is easy to learn or highly represented in the dataset), $a_S(D_{-y_i})$ has a more significant impact on v_{y_i}.
- When $a_S(D_{y_i})$ is low (i.e. class is hard to learn or not well represented in the dataset), $a_S(D_{-y_i})$ has a minimal impact on v_{y_i}.

Defining CS-Shapley:

For each class: sample an out-of-class environment $S^{(k)}_{-y_i}$, use truncated Monte Carlo on in-class data points to estimate Shapley values, repeat K times, and normalize.

Class-wise Shapley value:

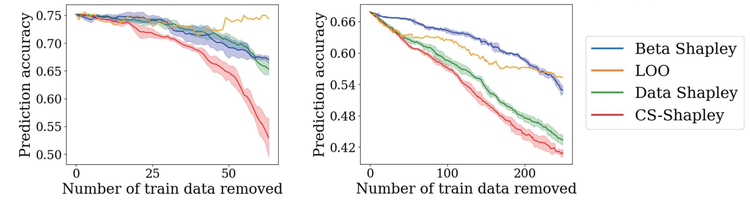
$$\phi_i|S_{-y_i} = \sum_{S_{y_i} \subseteq T_{y_i} \setminus \{i\}} \frac{v_{y_i}(S_{y_i} \cup \{i\}|S_{-y_i}) - v_{y_i}(S_{y_i}|S_{-y_i})}{\binom{n-1}{|S_{y_i}|}}$$

CS-Shapley:

$$\phi_i \approx \frac{1}{K} \sum_{S^{(k)}_{-y_i} \subseteq T_{-y_i}; k \in \{1,..,K\}} [\phi_i|S^{(k)}_{-y_i}]$$

Results:

High-value Data Removal: Removing high-value data should result in a drop in performance; quickest and largest drop indicates the best performing method. Logistic regression classifier on Diabetes (L) and Covertype (R) datasets.



Noisy Data Detection: Area under the precision-recall curve (AUC) when scanning data from lowest-to-highest value for mislabeled examples.

Dataset	Logistic Regression				SVM-RBF			
	CS	Data	Beta	LOO	CS	Data	Beta	LOO
CIFAR10	**0.450**	0.429	0.424	0.275	**0.387**	0.317	0.321	0.272
Click	**0.816**	0.689	0.797	0.149	**0.855**	0.769	0.789	0.200
Covertype	0.706	**0.766**	0.653	0.179	**0.712**	0.618	0.600	0.196
CPU	**0.785**	0.779	0.654	0.207	**0.808**	0.671	0.516	0.189
Diabetes	**0.441**	0.355	0.435	0.194	**0.412**	0.362	0.400	0.210
FMNIST	**0.570**	0.554	0.552	0.340	**0.512**	0.382	0.412	0.239
MNIST-2	**0.831**	0.815	0.806	0.280	**0.837**	0.663	0.611	0.300
MNIST-10	0.877	**0.933**	0.845	0.371	0.674	**0.747**	0.510	0.254
Phoneme	**0.575**	0.535	0.416	0.222	**0.579**	0.555	0.496	0.255

Transferability of Data Values: High-value data removal using logistic regression values and training MLP model. Want curves similar to above.