L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data

J. Chen, L. Song, M. Wainwright, M. Jordan

UC Berkeley
Georgia Institute of Technology
Ant Financial
Voleon Group

https://openreview.net/pdf?id=S1E3Ko09F7

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- Interpretability of models is challenging
- Shapley value approach has been used, but it is computationally challenging (exponential to number of features)
- Propose approximate Shapley value calculation which takes linear time when data can be structured as graph
- Two methods: L-Shapley and C-Shapley (Local and Connected)

Background

- Importance score of a feature subset

$$v_x(S) := \mathbb{E}_m \left[-\log \frac{1}{\mathbb{P}_m(Y \mid x_S)} \;\middle|\; x \right]$$ where

$x \in \mathbb{R}^d, S \subset \{1, 2..., d\}, x_S = \{x_j, j \in S\}$

- Class specific importance: use degenerate conditional distribution

$$\hat{\mathbb{P}}_m(y \mid x) := \begin{cases} 1 \text{ if } y \in \arg\max_{y'} \mathbb{P}_m(y' \mid x), \\ 0 \text{ otherwise.} \end{cases}$$

- Estimate conditional distribution through empirical average or plug-in estimation using a reference point

Background

Shapley Value

- Marginal contribution of feature *i* on feature subset *S*

$$m_x(S, i) := v_x(S) - v_x(S \setminus \{i\}).$$

- Aggregate marginal contribution of *i* over ALL subsets that contain *i* to get Shapley value

$$\phi_x(\mathbb{P}_m, i) := \frac{1}{d} \sum_{k=1}^{d} \frac{1}{\binom{d-1}{k-1}} \sum_{S \in \mathcal{S}_k(i)} m_x(S, i).$$

Background

Shapley Value

- Need to account for all 2^{d-1} subsets which contain i – very computationally expensive
- Monte Carlo and weighted linear regression used in past to speed up
  - In practice, could need prohibitively large samples to avoid high variance
  - Required sample size gets exponentially larger as feature vector size increases

Methods

- In many applications, features can be associated to nodes of a graph – features distant in graph have weaker interactions
- Use k-neighborhood of a node

$$\mathcal{N}_k(i) := \{j \in V \mid d_G(i,j) \leq k\}$$

Methods

L-Shapley (Local)

- Same as Shapley, except for kth-order L-Shapley only consider features in k-neighborhood

Definition 1. *Given a model $\mathbb{P}_m$, a sample x and a feature i, the* L-Shapley estimate of order k *on a graph G is given by*

$$\hat{\phi}_x^k(i) := \frac{1}{|\mathcal{N}_k(i)|} \sum_{\substack{T \ni i \\ T \subseteq \mathcal{N}_k(i)}} \frac{1}{\binom{|\mathcal{N}_k(i)|-1}{|T|-1}} m_x(T, i). \tag{5}$$

Methods

C-Shapley (Connected)

- Further reduces complexity – kth order C-Shapley only considers connected subsets within k-neighborhood

Definition 2. *Given a model $\mathbb{P}_m$, a sample x and a feature i, the* C-Shapley estimate of order k *on a graph G is given by*

$$\tilde{\phi}_x^k(i) := \sum_{U \in \mathcal{C}_k(i)} \frac{2}{(|U|+2)(|U|+1)|U|} m_x(U,i), \tag{6}$$

where $\mathcal{C}_k(i)$ denotes the set of all subsets of $\mathcal{N}_k(i)$ that contain node i, and are connected in G.

- Coefficients use Myerson value to characterize a new coalitional game over the graph G in which the influence of disconnected subsets of features are additive

Properties

- Characterize relationship between L-Shapley and Shapley values
  through absolute mutual information as measure of dependence

as a measure of dependence. Given two random variables X and Y, the absolute mutual information $I_a(X;Y)$ between X and Y is defined as

$$I_a(X;Y) = \mathbb{E}\left[\left|\log \frac{P(X,Y)}{P(X)P(Y)}\right|\right], \tag{7}$$

- Two theorems which prove that L-Shapley and C-Shapley are related to Shapley value when model obeys Markovian structure encoded by the graph
- C-Shapley is also shown to be related to the Myerson value

Experiments

- Compare on black-box models with KernelSHAP (regression-based approximation), SampleShapley, and LIME
- Omit methods requiring certain class models since this is model-agnostic

Experiments

Text Classification

- IMDB with Word-CNN, AG news with Char-CNN, Yahoo! Answers with LSTM
- Consider L-Shapley with 1-node neighborhood, C-Shapley evaluates all n-grams with $n \leq 4$
- Examine change in log-odds scores before and after masking the top features ranked by importance scores
- L-Shapley best for Word-CNN, L-Shapley and C-Shapley best for AG news, C-Shapley best for LSTM

Experiments

Text Classification

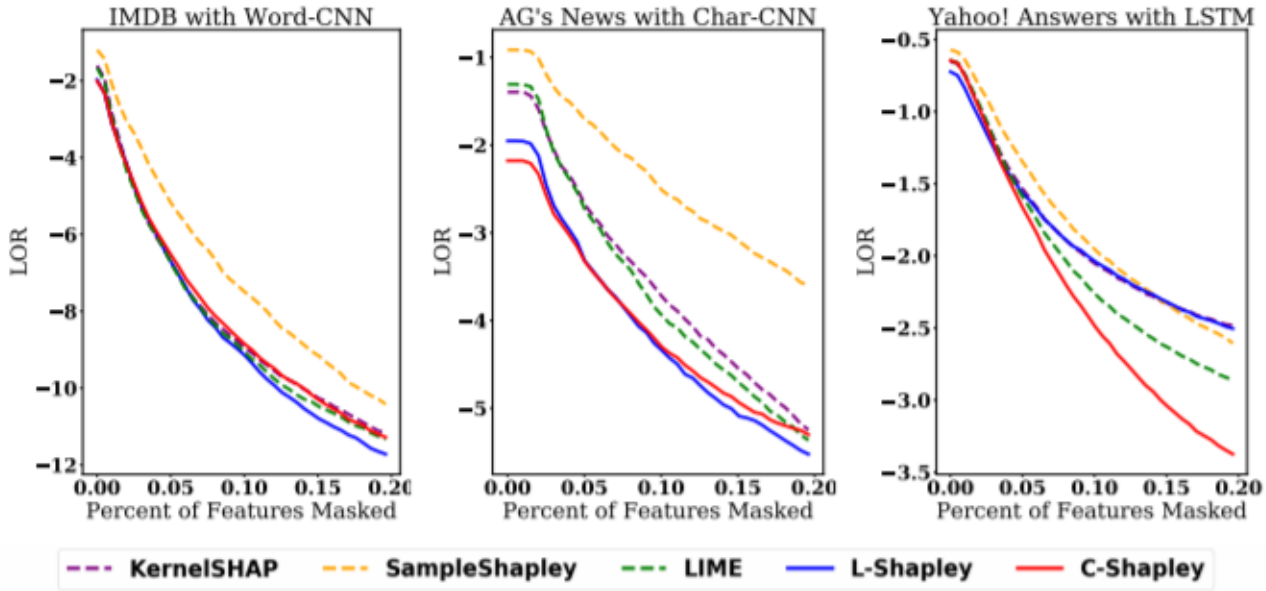

Figure 2: The above plots show the change in log odds ratio of the predicted class as a function of the percent of masked features, on the three text data sets. Lower log odds ratios are better.

Method	Explanation
Shapley	It is not heartwarming or entertaining . It just sucks .
C-Shapley	It is not heartwarming or entertaining . It just sucks .
L-Shapley	It is not heartwarming or entertaining . It just sucks .
KernelSHAP	It is not heartwarming or entertaining . It just sucks .
SampleShapley	It is not heartwarming or entertaining . It just sucks .

Table 2: Each word is highlighted with the RGB color as a linear function of its importance score. The background colors of words with positive and negative scores are linearly interpolated between blue and white, red and white respectively.

Experiments

Image Classification

- MNIST and CIFAR10 where each pixel is a feature
- LIME excluded because it requires superpixels, L-Shapley excluded because too expensive
- C-Shapley outperforms all other models

Experiments

Image Classification

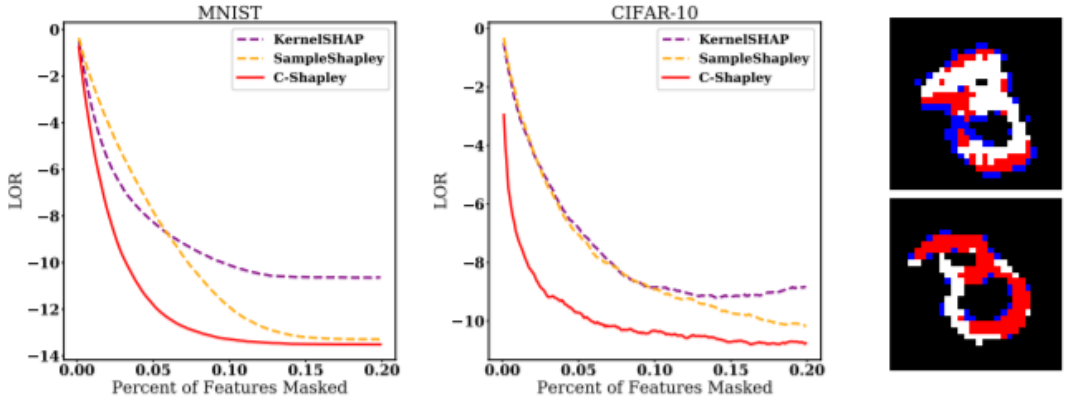

Figure 3: Left and Middle: change in log-odds ratio vs. the percent of pixels masked on MNIST and CIFAR10. Right: top pixels ranked by C-Shapley for a "3" and an "8" misclassified into "8" and "3" respectively. The masked pixels are colored with red if activated (white) and blue otherwise.

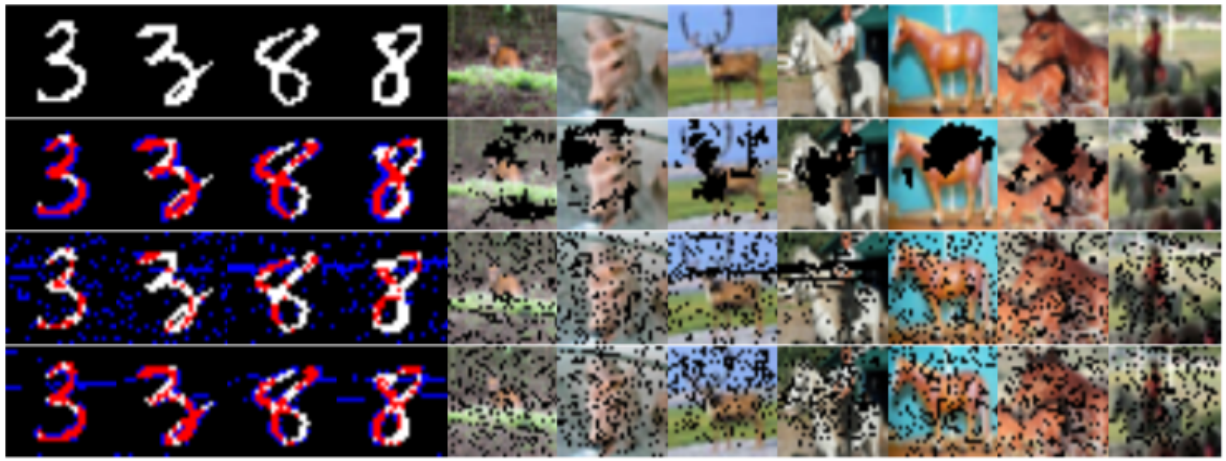

Figure 4: Some examples of explanations obtained for the MNIST and CIFAR10 data sets. The first row corresponds to the original images, with the rows below showing images masked based on scores produced by C-Shapley, KernelSHAP and SampleShapley respectively. For MNIST, the masked pixels are colored with red if activated (white) and blue otherwise.

Experiments

Evaluation with Humans

- Use Amazon Mechanical Turk to compare L-Shapley, C-Shapley, and KernelSHAP on IMDB reviews
- Examine if humans can decide with only top words or with top words masked
- Asked to rate positive or negative
- Three text types: raw reviews, reviews with only top-ten words ranked by each algorithm, reviews with top words masked (all other words are replaced with [MASKED])
- Humans perform best on only top words, C-Shapley results in best confidence and accuracy, L-Shapley harms the most when masking words

Experiments

Evaluation with Humans

Algorithm	Modification	Consistency	Standard Deviation	Abs. Score	Words Masked
Raw	None	0.880	0.960	0.811	N/A
L-Shapley	Selected	0.970	0.891	1.118	N/A
	Masked	**0.615**	**1.077**	**0.474**	**14.36%**
C-Shapley	Selected	**0.990**	**0.500**	**1.441**	N/A
	Masked	0.830	0.778	0.743	14.75%
KernelSHAP	Selected	0.960	0.627	1.036	N/A
	Masked	0.660	0.818	0.492	31.60%

Table 3: Results of human evaluation. "Selected" and "Masked" indicate selected words and masked reviews respectively. Results are averaged over 200 samples. (The best numbers are highlighted.)

Conclusion

- Proposed L-Shapley and C-Shapley for instancewise feature scoring on graphically structured data
- Demonstrated superior performance of these algorithms

References

- https://openreview.net/pdf?id=S1E3Ko09F7