L-SHAPLEY AND C-SHAPLEY: EFFICIENT MODEL INTERPRETATION FOR STRUCTURED DATA

Jianbo Chen* Le Song†,§ Martin J. Wainwright*,‡ Michael I. Jordan*

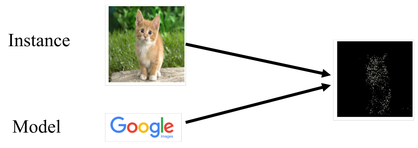
UC Berkeley* , Georgia Institute of Technology† , Ant Financial§ , Voleon Group‡

ABSTRACT

We focus on instancewise feature attribution where the data have a graph structure, and the contribution of features to the target variable is well-approximated by a graph-structured factorization. In such settings, we develop two algorithms with linear complexity for instancewise feature importance scoring on black-box models. We establish the relationship of our methods to the Shapley value.

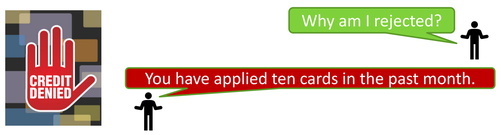
OBJECTIVE

For a given instance, assign a vector of importance scores for each feature.

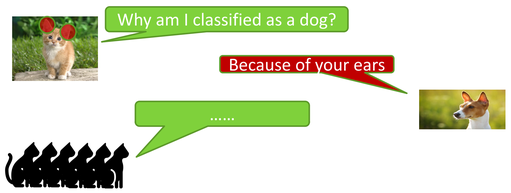


MOTIVATION

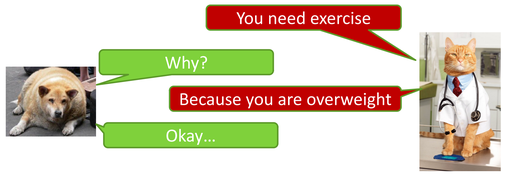
- Reasons for decision making



- Debugging a model

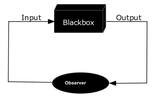


- Increase trust in decision making

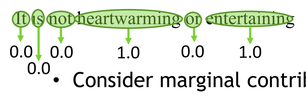


DESIRED PROPERTIES

- Black-box



- Interaction
  - Estimate features one after another



$\phi_{f,x}(\text{“not”}) = f(\text{“not”})$

  - Consider marginal contribution

It is not heartwarming or entertaining
It is not heartwarming or entertaining

$\phi_{f,x}(\text{“not”}) = f(\text{“not heartwarming”}) - f(\text{“heartwarming”})$

- Fairness
- **Scalability**

SHAPLEY VALUE

- A fair way to combine marginal contributions
  - [Additivity] $\sum_{i=1}^{d} \phi_{f,x}(i) = f(\{1,\dots,d\}) - f(\emptyset)$
  - [Equal Contributions] $\forall S, f(S \cup \{i\}) = f(S \cup \{j\}) \Rightarrow \phi_{f,x}(i) = \phi_{f,x}(j)$
  - [Monotonicity] $\forall S, f(S \cup \{i\}) - f(S) \geq g(S \cup \{i\}) - g(S) \Rightarrow \phi_{f,x}(i) \geq \phi_{g,x}(i)$
- The unique distribution satisfying the three properties:

$$\phi_{f,x}(i) = \frac{1}{d} \sum_{S \subset [d]} \frac{1}{\binom{d-1}{|S|-1}} (f(S \cup \{i\}) - f(S))$$

- Complexity : $\Theta(2^d)$
- Existing work:
  - [Strumbelj and Kononenko 2010]
  - [Datta, Sen and Zick 2016]
  - [Lundberg and Lee 2017]

GRAPH STRUCTURE

$G = (V, E)$

node $i \Leftrightarrow$ feature i

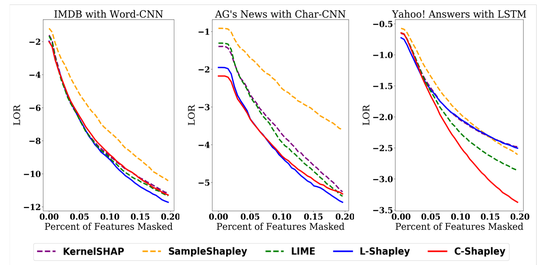
edge $e = (i, j) \Leftrightarrow$ interaction between i, j

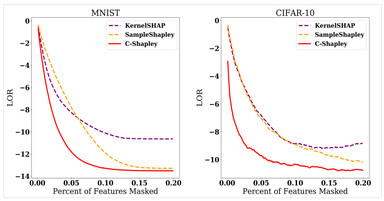
$d_G(\ell, m) =$ number of edges in shortest path joining ℓ to m

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

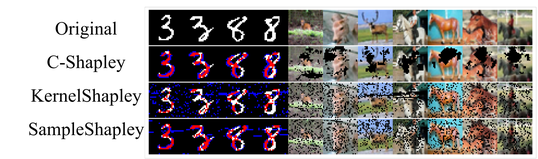
EXPERIMENTS

- Text data: IMDB, AG's News, Yahoo! Answers
- Image data: MNIST, CIFAR-10
- Models: Word-CNN, Char-CNN, LSTM, 2d-CNN
- Comparing methods: KernelSHAP, SampleShapley, LIME
- Metric: Log odds ratio with top features masked





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 .



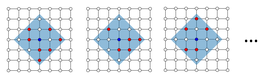
L-SHAPLEY AND C-SHAPLEY

Example: The movie tells the story of a little girl who … It is not heartwarming or entertaining

$f(\text{“the story of … not”}) - f(\text{“the story of”}) \approx f(\text{“not”}) - f(\emptyset)$

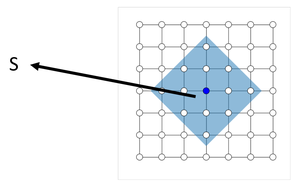
Observation: The words distant from them have a weaker influence on the importance of a given word in a document.

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



Complexity: $\Theta(2^{k^n} \cdot d)$

Theorem 1. *Suppose there exists a feature subset $S \subset \mathcal{N}_k(i)$ with $i \in S$, such that feature i is approximately independent of features outside S, conditioned on features within S and the model output Y. Then $\hat{\phi}^k$ is approximately equal to ϕ in expectation.*



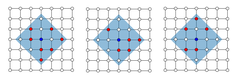
Example: It is not heartwarming or entertaining

$f(\text{“It … not heartwarming”}) - f(\text{“It … heartwarming”}) \approx f(\text{“not heartwarming”}) - f(\text{“heartwarming”})$

Observation: “It not heartwarming” rarely appears in real data and may not make sense to a human or a model trained on real-world data.

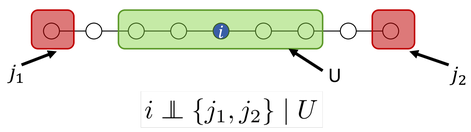
C-Shapley: $\tilde{\phi}_{f,x}^k(i) := \sum_{U \in \mathcal{C}_k(i)} \frac{2}{(|U|+2)(|U|+1)|U|} (f(U \cup \{i\}) - f(U))$

$\mathcal{C}_k(i)$ contains connected subsets of $\mathcal{N}_k(i)$ that contain i



Complexity: $\Theta(k^{2n} \cdot d)$

Theorem 2. *Besides the condition in Theorem 1, for any connected subset $U \subset S$ containing i, feature i is approximately independent of features disconnected with U, conditioned on features in U and the model output Y. Then $\tilde{\phi}^k$ is approximately equal to ϕ in expectation.*



$i \perp\!\!\!\perp \{j_1, j_2\} \mid U$