Explanations of Black Box Models based on Directional Feature Interactions

Aria Masoomi[1], Davin Hill[1], Zhonghui Xu[2], Craig P. Hersh[2], Edwin K. Silverman[2], Peter Castaldi[2], Stratis Ionnidis[1], Jennifer Dy[1]

[1] Northeastern University, Department of Electrical and Computer Engineering [2] Brigham and Women's Hospital, Channing Division of Network Medicine

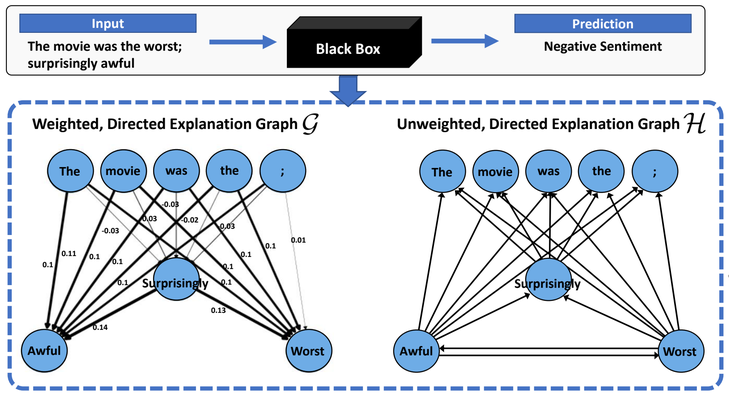
Abstract

As machine learning algorithms are deployed ubiquitously to a variety of domains, it is imperative to make these often black-box models transparent.

Several recent works explain black-box models by capturing the most influential features for prediction per instance; such explanation methods are univariate, as they characterize importance per feature. However, these univariate methods do not account for *feature interactions*.

- We propose a method to extend any given univariate removal-based explanation to a bivariate explanation model that can capture asymmetrical feature interactions, represented as a directed graph
- Analyzing this graph gives a semantically-rich interpretation of black boxes. Beyond the ability to identify most *influential features*, this graph can identify *directionally redundant features*, as well as *mutually redundant features.*
- Extensive experiments show that our explanation graph outperforms prior symmetrical interaction explainers as well as univariate explainers with respect to post-hoc accuracy, AUC, and time.

Illustrative Example



Mutual Redundancy

Worst & Awful are Mutually Redundant

Directional Redundancy

Awful makes {The, movie, was, ;, surprisingly} redundant

Top Influential Features

1. Worst
2. Awful
3. Surprisingly

Figure 1: Bivariate Shapley builds instance-wise explanation graph $\mathcal{G}$ and $\mathcal{H}$. **Graph $\mathcal{G}$:** An edge i → j represents the conditional influence of word j when word i is present. We can use the PageRank algorithm to identify the top influential features. **Graph $\mathcal{H}$:** an edge i → j indicate that word i makes the word j redundant. We identify strongly connected components, source nodes, and sink nodes to find mutually redundant and directionally redundant features.

Experiments

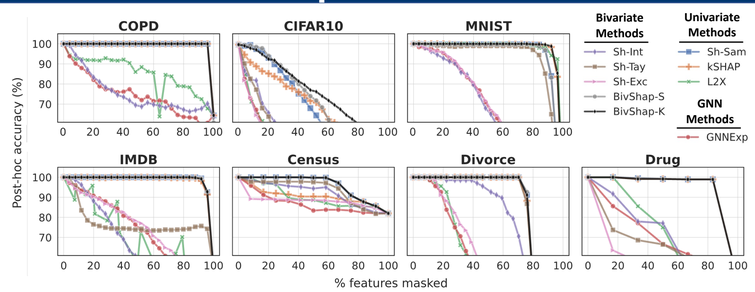


Figure 2: Post-hoc accuracy evaluated on Mutual Redundancy masking derived from graph H. Strongly connected nodes in H are randomly masked with increasing mask sizes until a single node remains, represented by the final marker for each dataset.

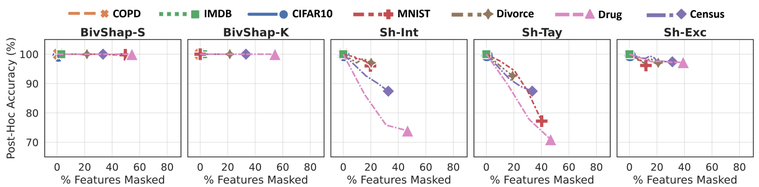


Figure 3: Post-hoc accuracy evaluated on Mutual Redundancy. Mutually Redundant features are iteratively masked until one feature remains in each group. We see that masking redundant features identified by Bivariate Shapley has minimal impact to accuracy.

	Post-hoc Accy		% Feat Masked	
Dataset	$\mathcal{H}$-Sink Masked	$\mathcal{H}$-Source Masked	$\mathcal{H}$-Sink Masked	$\mathcal{H}$-Source Masked
COPD	99.5	62.7	1.5	98.5
CIFAR10	94.6	15.0	6.2	93.8
MNIST	100.0	13.4	77.7	22.3
IMDB	100.0	54.0	3.5	96.5
Census	100.0	82.0	23.8	76.2
Divorce	100.0	51.5	22.2	77.8
Drug	100.0	48.5	43.5	56.5

Table 1: Post-hoc accuracy of BivShap-S after masking H-source nodes, representing features with minimal redundancies, and H-sink nodes, representing directionally redundant features. We see that masking sinks has little effect on accuracy. In contrast, masking H-source nodes and keeping sinks results in large decreases in accuracy, indicating that prediction-relevant information is lost during masking.

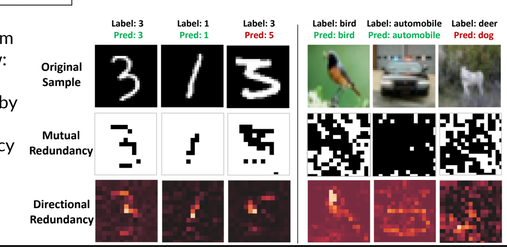
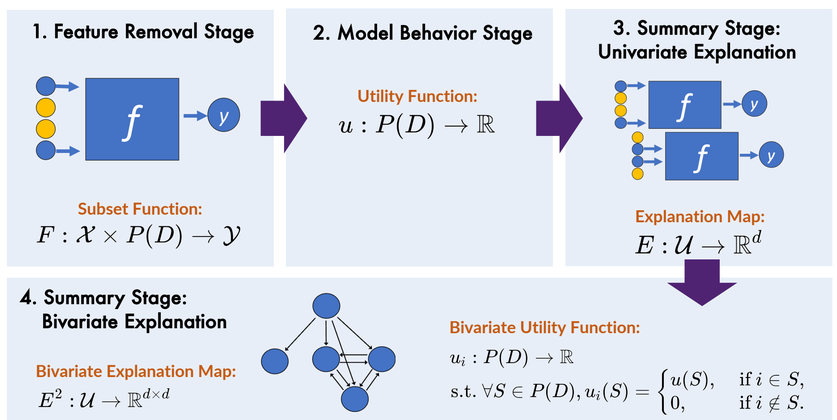


Figure 4: We explore samples from MNIST and CIFAR10. Middle row: we identify Mutually Redundant features from graph H, indicated by the white pixels in each image. Bottom row: We apply redundancy ranking on graph G and show a heatmap of PageRank scores; important nodes have higher PageRank scores.

Method

Univariate ➔ Bivariate Explanations



(Univariate) Shapley: $$E(u)_i = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!}(u(S \cup \{i\}) - u(S))$$

Bivariate Shapley: $$E^2(u)_{ij} = \sum_{j \in S \subseteq D \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!}(u(S \cup \{i\}) - u(S))$$

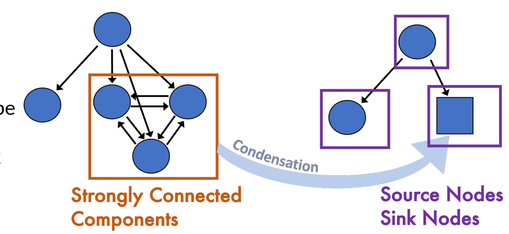
Mutual & Directional Redundancy

Given $i, j \in D$:

Mutual Redundancy: features i, j are mutually redundant if $E^2(u)_{ij} = E^2(u)_{ji} = 0$.

Directional Redundancy: i is directionally redundant with respect to feature j if $E^2(u)_{ij} = 0$.

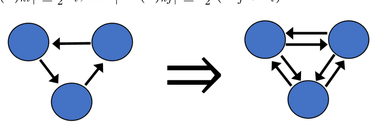
We can extend this pairwise notion to identify *groups* of Mutually and Directionally redundant features. **Mutually Redundant Features** can be identified as Strongly Connected Components. We then identify Sink and Source nodes which represent **directionally redundant** and important features, respectively.



Transitivity

Theorem 1: For $i, j, k \in D$, assume that $\max_{j \in S \subseteq N} |u(S \cup \{i\}) - u(S)| \leq \varepsilon_j$ and $\max_{i \in S \subseteq N} |u(S \cup \{k\}) - u(S)| \leq \varepsilon_i$. Then, the following inequalities hold: $|E^2(u)_{ij}| \leq \frac{d!}{2}\varepsilon_j$, $|E^2(u)_{ki}| \leq \frac{1}{2}\varepsilon_i$, and $|E^2(u)_{kj}| \leq \frac{d!}{2}(2\varepsilon_j + \varepsilon_i)$

Corollary 1.1: *Graph $\mathcal{H}$ for Shapley explanation map with a monotone utility function is transitive.*



Acknowledgements

The work described was supported in part by Award Numbers U01 HL089897, U01 HL089856, R01 HL124233, and R01 HL147326 from the National Heart, Lung, and Blood Institute, the FDA Center for Tobacco Products (CTP), and NSF IIS 1546428.

View our code and other examples on Github →