Attribution for Data with Graphical Feature Dependencies

Layerwise Relevance Propagation

4/29/2024

Jared Winslow

Interpretable Machine Learning

Motivation

Considerations and Advancements:

- Adversarial Examples [1]
- LRP on Tabular Data [7]
- Dominant Sets on MI Graphs [5]

Overview

Project Steps:

- Dependency Measures
- Graph Metrics
- Data Generation
- Layerwise Relevance Propagation

Dependency Measures

From most information to least:

1. Joint Distribution
2. Bayesian Network
3. Interaction Information
4. Mutual Information
5. Correlation

Dependency Measures

Bayesian Network:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \mathrm{Parents}(x_i))$$

Mutual Information (MI):

I(X; Y) = H(X) + H(Y) - H(X, Y)

Interaction Information:

I(X; Y; Z) = H(X) + H(Y) + H(Z) - H(X, Y) - H(X, Z) - H(Y, Z) + H(X, Y, Z)

Entropy:

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$$

Graph Metrics

Metrics for individual features in dependency graph:

- Relative eigenvector centrality
- Other centralities (e.g., betweenness, etc.)

Metrics for graph-level feature dependency:

- Average eigenvector centrality
- Entropy of the eigenvector centrality distribution
- Graph clustering coefficient (i.e., proportion of triplets)
- Number of dominant sets (i.e., cliques)

Data Generation: Correlation

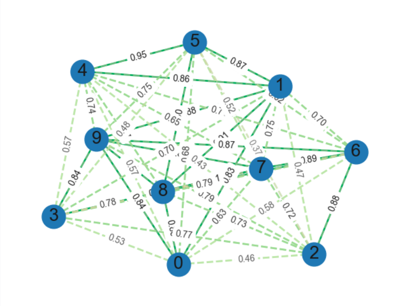
Simulating correlation matrices:

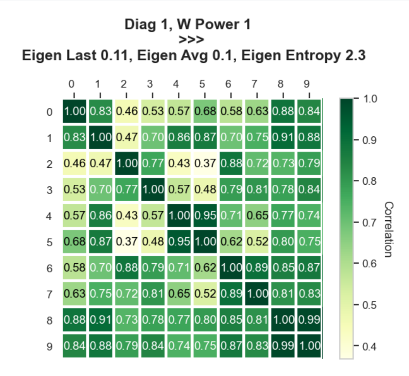
$$C = WW^T + D$$

$$N = diag(C)^{-1/2}$$

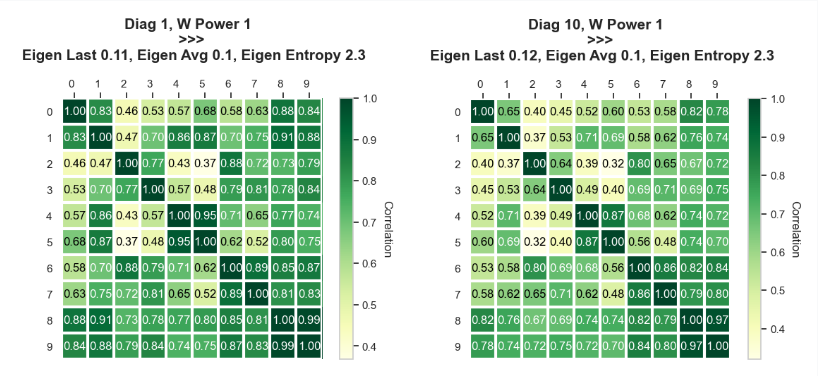
$$\Sigma = NCN$$

Data Generation: Graph Metric to Correlation

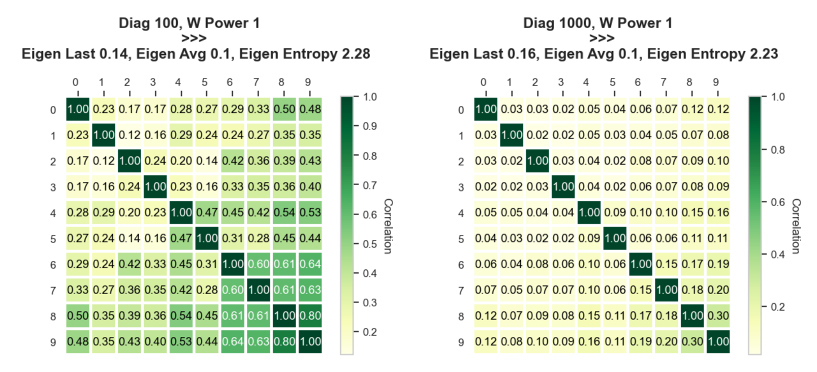




Data Generation: Graph Metric to Correlation



Data Generation: Graph Metric to Correlation



Data Generation: Linearly Related Features

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

$$X_{ij} = \begin{cases} X_{ij} & \text{if } j \neq 4 \text{ or } Z_i = 1, \\ 0 & \text{if } j = 4 \text{ and } Z_i = 0, \end{cases}$$

where $Z_i \sim \text{Bernoulli}(0.5)$ independently for each sample i.

$$\mathbf{y_1} = \mathbf{X}\beta + \epsilon$$
$$\mathbf{y_2} = f(\mathbf{X}\beta) + \epsilon$$

$$y_i = y_i + 10 \cdot X_{i4} \quad \text{for all } i \in \text{outlier indices}$$

Data Generation: Nonlinearly Related Features

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

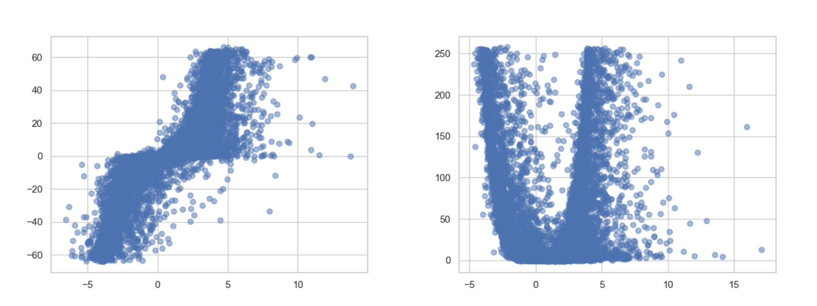
$$\mathbf{X}^{\mathbf{nl}} = X_j X_k + f_l(t)$$

$$X_{ij}^{nl} = \begin{cases} X_{ij}^{nl} & \text{if } j \neq 4 \text{ or } Z_i = 1, \\ 0 & \text{if } j = 4 \text{ and } Z_i = 0, \end{cases}$$

$$\mathbf{y_1} = \mathbf{X}^{\mathbf{nl}}\beta + \epsilon$$
$$\mathbf{y_2} = f(\mathbf{X}^{\mathbf{nl}}\beta) + \epsilon$$

$$y_i = y_i + 10 \cdot X_{i4}^{nl} \quad \text{for all } i \in \text{outlier indices}$$

Data Generation: Nonlinearly Related Features



Layerwise Relevance Propagation

Gamma rule:

$$R_i^{(l)} = \sum_j \left(\frac{a_i(w_{ij} + \gamma w_{ij}^+)}{\sum_j a_i(w_{ij} + \gamma w_{ij}^+)} R_j^{(l+1)} \right)$$

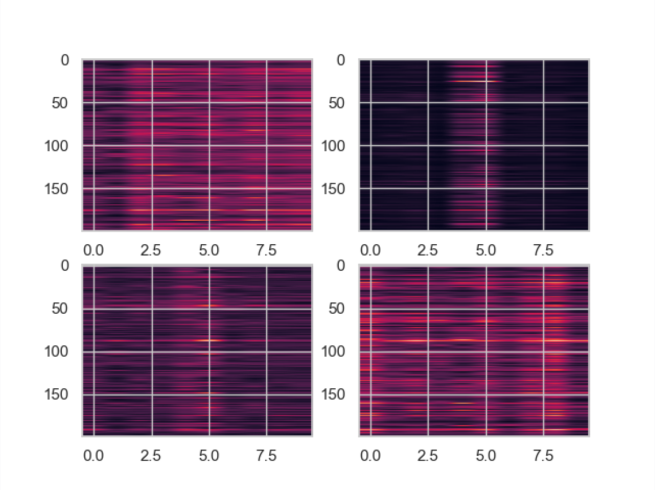
Epsilon rule:

$$R_i^{(l)} = \sum_j \left(\frac{w_{ij}}{\sum_i w_{ij} + c\epsilon_1 + \epsilon_2} R_j^{(l+1)} \right)$$

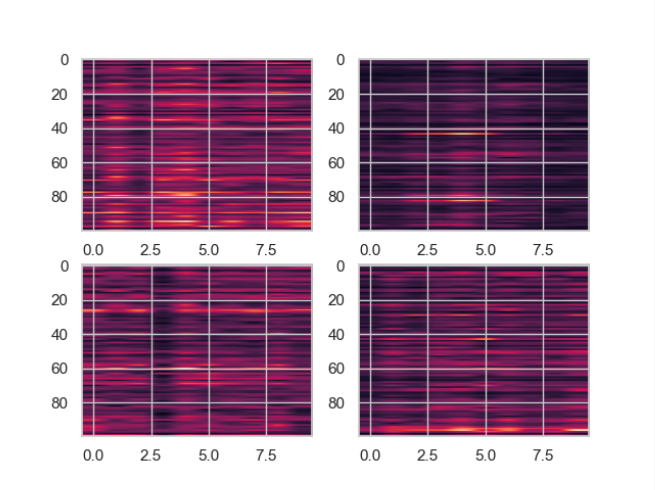
where $\epsilon_1 = sqrt(\sum_i w_{ij}^2)$

$$R_i^{(l)} = \sum_j \left(\frac{(w_{ij})^2}{\sum_i (w_{ij})^2} R_j^{(l+1)} \right)$$

Layerwise Relevance Propagation: Explanations



Layerwise Relevance Propagation: Outliers



References I

[1] T.R. Dieter and H. Zisgen. “Evaluation of the Explanatory Power Of Layer-wise Relevance Propagation using Adversarial Examples”. In: *Neural Process Lett* 55 (2023), pp. 8531–8550. DOI: 10.1007/s11063-023-11166-8.

[2] Maximilian Kohlbrenner et al. *Towards Best Practice in Explaining Neural Network Decisions with LRP*. 2020. arXiv: 1910.09840 [cs.LG].

[3] S. Lapuschkin. *Opening the machine learning black box with Layer-wise Relevance Propagation*. https://www.semanticscholar.org/paper/Opening-the-machine-learning-black-box-with-Lapuschkin/c601b185b6080ded463d3c236fa4f9f849f0435b. Accessed: 2024-04-28.

References II

[4] G. Montavon et al. “Layer-Wise Relevance Propagation: An Overview”. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700. Lecture Notes in Computer Science. Cham: Springer, 2019. DOI: 10.1007/978-3-030-28954-6_10.

[5] M. Pavan and M. Pelillo. “A new graph-theoretic approach to clustering and segmentation”. In: *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Madison, WI, USA: IEEE, 2003, p. 1211348. ISBN: 0-7695-1900-8. DOI: 10.1109/CVPR.2003.1211348.

[6] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. *Learning Important Features Through Propagating Activation Differences*. 2019. arXiv: 1704.02685 [cs.CV].

References III

[7] Ihsan Ullah et al. "Explaining Deep Learning Models for Tabular Data Using Layer-Wise Relevance Propagation". In: *Applied Sciences* 12.1 (2022), p. 136. DOI: 10.3390/app12010136.

Thank you!

https://github.com/jaredwins99/