Layer-Wise Relevance Propagation for Tabular Data with Graphical Dependencies

Jared Winslow

May 10, 2024

1 Introduction

Machine learning (ML) and deep learning have made grand strides in general predictive performance in the last decades. The use of ML is now becoming commonplace in high-stakes domains like medicine and criminal justice, where the threat of making erroneous or biased decision-making leads to larger repercussions. With model decisions having greater sway than before, understanding models' internals and auditing their behavior is ever more important. Within the burgeoning field of interpretable machine learning (IML) and explainable AI (XAI), researchers using a suite of interpretability methods, ranging from model-specific to model-agnostic, have attempted to identify the data features driving decisions in these models (Allen, Gan, and Zheng 2023). State-of-the-art deep learning models, however, are some of the most opaque models, and the best we can hope for at present are post-hoc explanations. Layer-wise relevance propagation (LRP) is a post-hoc, model-specific, local (attributes feature importance to individual observations rather than the global dataset) interpretability method that backpropagates activations in neural networks.

LRP has primarily been applied to computer vision models to produce saliency maps. Saliency maps illuminate, often literally more than figuratively, pivotal pixels or data features in model predictions. They are intuitively viewable, but assessing these maps is mostly subjective and ad-hoc. Much work has questioned the robustness of LRP as a feature attribution method and saliency maps in general, likening them to edge detectors (Adebayo et al. 2020). Validation attempts by corrupting the model and perturbing the data have frequently failed. Recently, researchers crafted adversarial images that would look indistinguishable from an apparent class but would be predicted by models as another class, and the saliency map constructed by LRP remained unchanged (Dieter and Zisgen 2023).

There has been limited research, however, on applying LRP to tabular data, also referred to as unstructured data (real-valued continuous data without the structure of e.g., image, audio, text data). One notable example is that researchers applied LRP to a one-dimensional convolutional neural network trained on mixed tabular data (Ullah et al. 2022). LRP was able to discern reasonable features as important, and the researchers validated these results by selecting those features and using them in simpler models. Given that we do not know the ground truth in such a situation, we cannot fully assess the correctness of LRP. Furthermore, it is unknown how LRP can be applied to less structured neural networks, especially neural network regression models trained on tabular data with a dependence structure.

In this paper, we apply LRP to a simple multi-layer perception (MLP) trained on a range of simulated numeric data, and evaluate how well LRP discerns the important features for the entire dataset and individual observations.

2 Background

The basic idea of LRP is to decompose the output prediction $f(x)$ into relevance scores R_i distributed across the input features x_i, expressed recursively by:

$$R_i^{(l)} = \sum_j \frac{a_{ij}^{(l)} w_{ij}^{(l+1)}}{\sum_i a_i^{(l)} w_{ij}^{(l+1)}} R_j^{(l+1)} \tag{1}$$

where $R_i^{(l)}$ is the relevance of neuron i in layer l, $a_i^{(l)}$ is the activation of neuron i in layer l, and $w_{ij}^{(l+1)}$ are the weights between layers l and $l+1$.

LRP can be applied to diverse models and datasets. Conversely, it has been recently shown that many methods in the IML and XAI literature suffer significant limitations due to stringent assumptions. For instance, Shapley values as a method makes additivity assumptions and LIME makes feature independence assumptions (Covert, Lundberg, and Lee 2022). One of deep learning's biggest assets is its ability to manage complex feature interactions and still make useful predictions. Deep-learning-based interpretability methods and LRP in particular have the potential to perform feature attribution even amidst feature interactions. Therefore, LRP is especially applicable towards models expressible enough to learn linear and nonlinear dependence structures.

Tabular data is often said to be unstructured because it lacks the grid and relational structure that images and text have respectively. Without preset structure, however, we can easily impose a customizable dependence structure. This dependence structure can be composed of conditional distributions, interaction information, mutual information, or correlation. To simplify simulations and avoid the necessary estimation that would be required for something like mutual information, we restrict our attention to correlation.

Once a correlation structure is constructed, we can analyze it more holistically by using graph metrics from network analysis. Notably, we are not only interested in the strength of correlation between two variables, but also the effect of correlation on other variables farther along a chain or network of correlation. One important metric from graph theory is eigenvector centrality, which measures the degree (or weighted degree) of nodes recursively, and can be represented mathematically by:

$$\lambda v_i = \sum_{j \in N(i)} a_{ij} v_j \tag{2}$$

where λ is a constant (the largest eigenvalue of the adjacency matrix A), v_i is the eigenvector centrality of node i, a_{ij} represents the edge weight from node i to node j, and $N(i)$ denotes the set of neighbors of node i.

Eigenvector centrality essentially measures the importance of each node given the importance of each of its neighbors. The node with the largest eigenvector centrality therefore has the greater influence on other nodes (measured globally). A correlation matrix is a weighted adjacency matrix where the pairwise correlations are weighted edges. Under this interpretation, the feature with the largest eigenvector centrality is the feature with the greatest (correlational) influence.

Beyond that, we can measure the entropy over the distribution of eigenvector centralities to capture the skewness of correlation within the variables. If the eigenvector centrality entropy is high, then all the variables have roughly the same eigenvector centrality within the correlation structure. This could mean that the variables are all heavily correlated, but to the same level, or that the variables are weakly correlated, but again to the same level.

3 Simulation Methodology

We start with 1) uncorrelated features X that linearly generate a target variable y, and then progressively incorporate 2) split feature domains, and 3) correlation.

For each of the simulations, there are 10 features (zero-indexed 0-9) all generated from a standard multivariate normal distribution, with 5 of the features contributing to the generation of y, with parameter values of 10, 5, 2, 1, and 1 respectively. This means that feature 0 is two times as important as feature 1. The other 5 features represent noise features with parameters values of 0. Together, y is simulated to be a linear function of X plus additional noise. Incorporating feature-related and target-related noise captures the fact that in real world problem we often times both have irrelevant features and do not have relevant features.

For the next simulation, to simulate a split domain, the features are partitioned into even and odd indices. In the even indices, y is generated in the same way. In the odd indices, y is then generated from the original features plus the last feature 8, with a parameter of 10. To help the MLP better learn the patterns, we also add an additional indicator variable for odd and even.

To simulate correlation matrices, we can construct covariance matrices and normalize them. The required positive semi-definiteness can be attained by taking the outer product of an arbitrary matrix, and then taking the sum of this new positive semi-definite matrix with a diagonal matrix to create a full rank matrix. The inverse square root of the diagonal represents the normalizing factor.

$$C = WW^T + D$$

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

$$\Sigma = NCN$$

Leveraging this formulation for different levels of skewness for W and different magnitudes for D, we create various correlation matrices. Specifically, to generate novel correlation structures, W is chosen to be an increasing sequence 1,2,3,...,10 and the skewness of W measures the power to which we raise W. The goal of specifying these hyperparameters is to create correlation matrices that can be parametrized by the aforementioned eigenvector centrality metrics and measure their impact on the attributions made by LRP.

Below is a visual showing correlation matrices as function of W and D. As the skewness of W is increased (downward), correlation is consolidated around the last feature. As the strength of D increases (rightward), the matrix becomes more independent and the maximum correlation possible is decreased. In this sense, the skewness impacts the eigenvector centrality entropy and the diagonal strengths impacts the maximum eigenvector centrality.

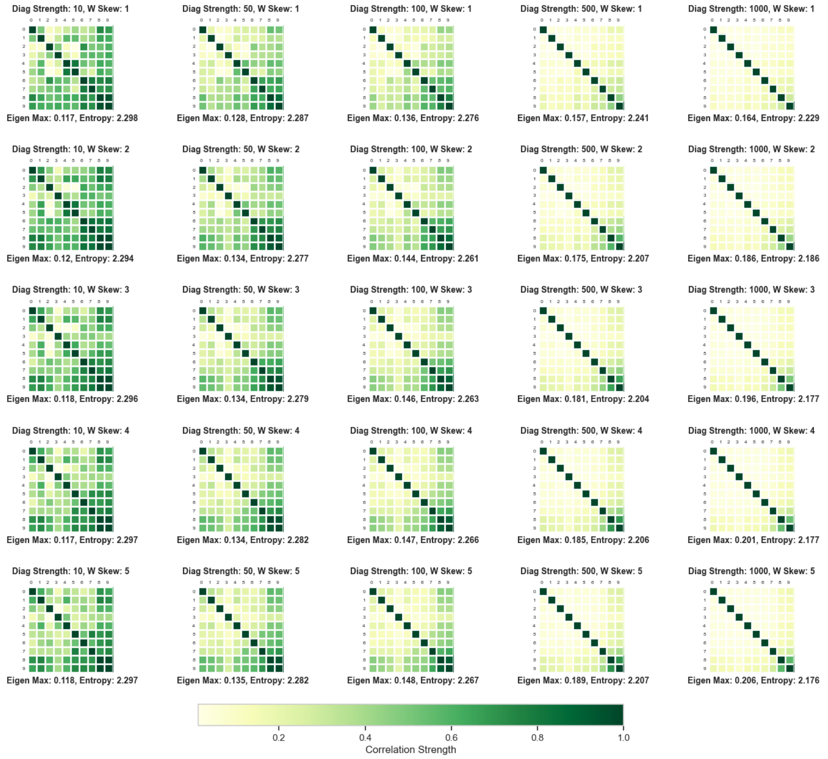

Figure 1: Correlation matrices as a function of max eigenvector centrality and eigenvector centrality entropy

4 Model Specifications and Attribution Results

We fix the model for each of the simulation cases. Specifically, we use a feedforward neural network with three hidden layers of size 64, 32, and 16 respectively with a dropoff of 30 percent. We train the model over 5000 epochs with the Adam optimizer using a learning rate of 0.0001 and default values for the other hyperparameters.

In applying LRP, we use an epsilon rule for the highest hidden layer, a gamma rule for the middle layers, and a w squared rule for the input layer.

In the most basic linear case, LRP successfully identifies the most important features globally (Figure 2), albeit with many sensitivities. Black feature-observation pairs represent negative attributions, white represent positive, and orange represent neutral. As can be seen in Figure 2, LRP attributes the most importance to feature 0, and attributes some importance to features 1 and 2. If we were to take the absolute value, this would be mostly correct. Conversely, LRP applied to the model is sensitive to whether the features are standardized. The standardized feature attributions are much less interpretable. The main problem, however, is that the feature attributions for individual observations are not consistent.

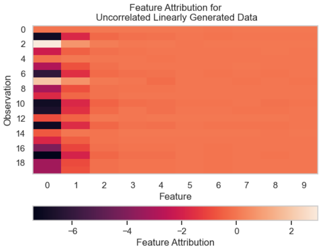

Figure 2: Unstandardized and with Bias in LRP Calculation

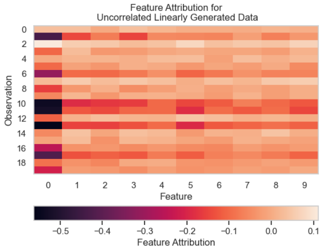

Figure 3: Standardized and with Bias in LRP Calculation

With a heterogeneous domain (Figures 4 and 5, now absolute value), LRP again successfully identifies the most important features globally, but fails to successfully attribute features locally in many cases. We see that feature 8 is less important for some observations, but we do not see that it is only relevant to odd indices.

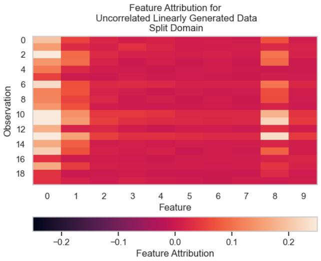

Figure 4: Feature 8 Influences Odd Indices

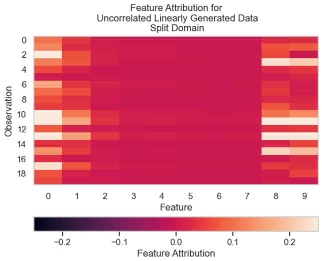

Figure 5: Including an Indicator Variable (Feature 9)

Similar results are found when there is correlation. However, the more correlation there is, the less LRP is able to attribute features.

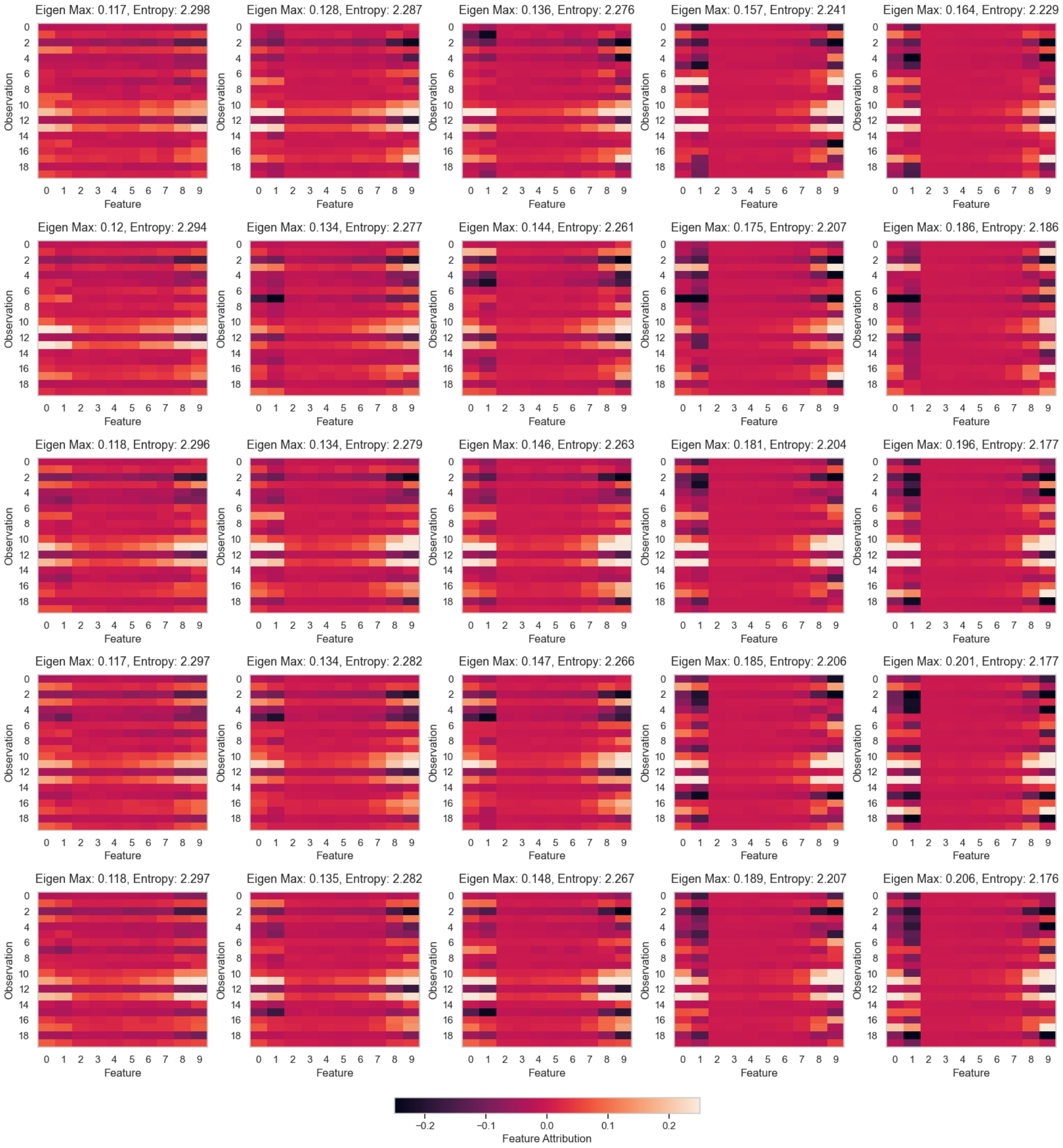

Figure 6: LRP Heatmaps as a Function of Max Eigenvector Centrality and Eigenvector Centrality Entropy

LRP was most impacted by general correlation. The larger the eigenvector centrality of the most important feature, the easier it was for LRP to discern the importance of that feature. The entropy did not seem to have much of an effect.

5 Discussion

LRP is able to identify importance features very roughly, however it fails to perform feature attribution for individual observations.

There are many limitations of this analysis. While simulation is iteratively made more complicated, the diversity of types of simulation is limited.

In future work, it would be valuable to analyze LRP using different dependency metrics and nonlinear relationships.

References

Adebayo, Julius et al. (2020). *Sanity Checks for Saliency Maps.* arXiv: `1810.03292 [cs.CV]`.

Allen, Genevera I., Luqin Gan, and Lili Zheng (2023). *Interpretable Machine Learning for Discovery: Statistical Challenges & Opportunities.* arXiv: `2308.01475 [stat.ML]`.

Borisov, V., K. Broelemann, E. Kasneci, et al. (2023). "DeepTLF: robust deep neural networks for heterogeneous tabular data". In: *International Journal of Data Science and Analytics* 16, pp. 85–100. DOI: `10.1007/s41060-022-00350-z`. URL: `https://doi.org/10.1007/s41060-022-00350-z`.

Covert, Ian, Scott Lundberg, and Su-In Lee (2022). *Explaining by Removing: A Unified Framework for Model Explanation.* arXiv: `2011.14878 [cs.LG]`.

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

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

Lapuschkin, S. (n.d.). *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.

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

Montavon, Grégoire (2024). *LRP Tutorial.* `https://git.tu-berlin.de/gmontavon/lrp-tutorial`. Accessed: 2024-04-28.

Pavan, M. and M. Pelillo (2003). "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, p. 1211348. ISBN: 0-7695-1900-8. DOI: `10.1109/CVPR.2003.1211348`.

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

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