Explanation evaluation metrics

1. Introduction

After obtaining an explanation of any artificial intelligence model, how do we know if the explanation is good or not? There are several methods of ensuring the transparency and trustworthiness of those models. In this document we will cover only those that we want to utilize in our project.

Human judgment - after gathering explanations, we assemble a group of humans (dependable on application of the model sometimes we want experts, sometimes possible users/clients) and ask them to evaluate the explanations subjectively. Usually we will ask humans to rate the relevance, clarity and overall usefulness of the explanations.

Metrics - there are also several metrics that can help us evaluate the explanations. We want to focus on metrics designed specifically for GNNs, which include: fidelity, characterization score and fidelity curve AUC (introduced in “GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks” paper) and GEF metric (introduced in “Evaluating Explainability for Graph Neural Networks” paper).

2. Fidelity

Fidelity metric can be used for both phenomenon and model explanations. It evaluates the contribution of the explanatory subgraph to the initial prediction. Fidelity$_+$ is obtained by removing the explanatory subgraph from the entire graph, while fidelity$_-$ is obtained by giving only the explanatory subgraph to the model. For the phenomenon fidelity is measured with respect to the ground-truth node label, for the model it is measured with respect to the outcome of the GNN model.

Phenomenon:

$$fid_{+} = \frac{1}{N}\sum_{i=1}^{N} |\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_{C\setminus S}} = y_i)|$$

$$fid_{-} = \frac{1}{N}\sum_{i=1}^{N} |\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_{S}} = y_i)|$$

Where y_i is the original prediction of graph i, N is the number of graphs, $\hat{y}_i$ is the prediction made by the gnn, $\hat{y}_i^{G_{C/S}}$ is prediction made by the gnn on graph without the explanatory subgraph, $\hat{y}_i^{G_S}$ is prediction made by the gnn on the explanatory subgraph and $\mathbb{1}$ is the indicator function.

Model:

$$fid_{+} = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i^{G_{C\setminus S}} = \hat{y}_i)$$

$$fid_{+} = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i^{G_{S}} = \hat{y}_i)$$

It can be misleading, but let's reiterate: for phenomenon explanation we compare model prediction and ground truth, for model explanation we compare model prediction with model prediction.

3. Fidelity AUC (area under curve)

Fidelity AUC is represented by the function:

$$f(x) = \frac{fid_{+}}{1 - fid_{-}}$$

4. Characterization score

Characterization score is another metric that utilizes fidelity scores. It returns component wise characterization score:

$$charact = \frac{w_{+} + w_{-}}{\frac{w_{+}}{fid_{+}} + \frac{w_{-}}{1 - fid_{-}}}$$

Where w are weights that should be picked from range (0, 1). Original paper used a value of 0.5 for both weights.

5. Graph explanation faithfulness

GEF evaluates how faithful explanations are to an underlying GNN predictor. The metric is calculated using formula:

$$GEF(y, \widehat{y}) = 1 - e^{(-KL(y||\widehat{y}))}$$

Where y is prediction probability vector obtained by processing the original graph through net, $\widehat{y}$ is prediction probability vector obtained from the masked subgraph, and KL is the Kullback-Leibler divergence score that quantifies the distance between the two probability distributions.

For discrete probability distributions P and Q defined on the same sample space, where X is the relative entropy from$Q$ to P, KL divergence is defined as:

$$KL(P||Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)})$$

6. References

- K. Amara, R. Ying, Z. Zhang, Z. Han, Y. Shan, U. Brandes, S. Schemm, C. Zhang: GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks, 2022
- C. Agarwal, O. Queen, H. Lakkaraju, M. Zitnik: Evaluating Explainability for Graph Neural Networks, 2022
- Pytorch geometric documentation - metrics