Explanation Variation Exp

Jiajun Zhu

May 17, 2023

Idea ¦ Observation

CXPlain, an explanation method with uncertainty, inspires us to test the explantion variation of current methods. There is a natural assumption: whether the more important nodes have lower std value.

We observe that for signal nodes, the importance scores' variation is perfectly decreasing with the importance scores' magnitude.For background nodes, the correlation varies a lot across different classifiers.

Another observation is that if auc performance is relatively high (>90), then use negative importance scores' std value's can get comparable performance.

Methods & Setting

- Synmol + EGNN, only take method PGExplainer (efficient) for example
- Ten different (erm) classifiers are trained (with different seeds)
- Each node of every positive sample is explained by 10 PGExplainers

Experiment Results

Table 1 shows the Spearman Correlation between avg/std IMP scores on x nodes. Except for classifier 1/3/5 which fails to provide high explanation auc (see Table 2), i.e. cannot distinguish signal nodes and bkg nodes, others' coef on all nodes are negative. And the coef on signal nodes are all -1.

Table 1: Spearman Correlation of avg/std node IMP

seed	all_nodes	sig_nodes	bkg_nodes
0	-0.627	-0.975	-0.04
1	0.315	-0.177	0.359
2	-0.987	-0.936	-0.983
3	0.394	0.871	-0.026
4	-0.51	-0.97	0.204
5	0.242	-0.6	0.541
6	-0.473	-0.956	-0.127
7	-0.863	-0.973	-0.638
8	-0.844	-0.958	-0.591
9	-0.802	-0.966	-0.542

Table 2 provide some information about the classifier and explainer performance. Interestingly regarding -std as the IMP score can get comparable performance except classifier 1/3/5 (and 6).

Table 2: Explanation AUC of avg/-std node IMP

seed	exp_auc_avg	exp_auc_std	clf_acc
0	0.927	0.906	0.973
1	0.771	0.346	0.984
2	0.823	0.816	0.984
3	0.204	0.737	0.984
4	0.947	0.907	0.978
5	0.732	0.42	0.989
6	0.905	0.699	0.995
7	0.906	0.905	0.984
8	0.917	0.906	0.978
9	0.953	0.903	0.984