Machine Learning for Graphs and Sequential Data Exercise Sheet 14

Graphs: Limitations

Randomized Smoothing

For the sake of simplicity, we consider a slightly different setup than in the lecture. In this exercise, we assume no knowledge about $f_\theta(\mathbf{x})$ respectively $g(\mathbf{x})_c$ (usually we would estimate a lower bound of $g(\mathbf{x})_c$ via Monte Carlo sampling, but here we do not).

We use the same sparsity-aware randomization scheme $\phi(\mathbf{x})$ as in the lecture:

$$g(\mathbf{x})_c = \mathcal{P}(f(\phi(\mathbf{x})) = c) = \sum_{\tilde{\mathbf{x}} \text{ s.t. } f(\tilde{\mathbf{x}})=c} \prod_{i=1}^{n^2} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) \tag{1}$$

with

$$\mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) = \begin{cases} p_d^{\mathbf{x}_i} p_a^{1-\mathbf{x}_i} & \tilde{\mathbf{x}}_i = 1 - \mathbf{x}_i \\ (1-p_d)^{\mathbf{x}_i}(1-p_a)^{1-\mathbf{x}_i} & \tilde{\mathbf{x}}_i = \mathbf{x}_i \end{cases} \tag{2}$$

and the number of nodes n. For an illustration we refer to Slide 15 "Smoothed Classifier for Discrete Data"

Problem 1: Given an arbitrary graph $\mathbf{x}$, and a perturbed one $\mathbf{x}'$ where $\mathbf{x}'$ differs from $\mathbf{x}$ in exactly one edge. What is the worst-case base classifier $h^*(\mathbf{x})$? In this context, we refer to the worst-case base classifier $h^*(\mathbf{x})$ as the classifier that has the largest drop in classification accuracy between $g(\mathbf{x})_c$ and $g(\mathbf{x}')_c$. Or in other words, $h^*(\mathbf{x})$ results in the most instable smooth classifier if we switch a single edge. This motivates the importance of analyzing robustness for graph neural networks (or other models with discrete input data).

Problem 2: How many of the possible graphs $\tilde{\mathbf{x}}$ does the worst-case base classifier assign the label c (see Problem 1)? To be more specific, we are looking for a term reflecting the absolute number and not a ratio?

Problem 3: What is $g(\mathbf{x}')_c$, $g(\mathbf{x})_c$, and $g(\mathbf{x}')_c - g(\mathbf{x})_c$ for the worst-case base classifier $h^*(\mathbf{x})$ (see Problem 1)? Please derive the equations (given $p_a + p_d < 1$). Subsequently, we would like to know the precise values for $p_a = 0.001$ and $p_d = 0.1$.