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 confidence 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).

The classifier with the *largest drop in classification accuracy between $g(\mathbf{x})_c$ and $g(\mathbf{x}')_c$* can be formalized as a minimization problem $h^*(\mathbf{x}) = \arg\min_{h(\mathbf{x})\in\mathcal{H}} g(\mathbf{x}')_c - g(\mathbf{x})_c$. In the following we consider a random order of edges and hence we may assume w.l.o.g. that all edges are identical but the last edge. Hence, from (1) it follows:

$$\begin{aligned}
\min_{h(\mathbf{x})\in\mathcal{H}} g(\mathbf{x}')_c - g(\mathbf{x})_c &= \min_{h(\mathbf{x})\in\mathcal{H}} \left(\sum_{\tilde{\mathbf{x}} \text{ s.t. } h(\tilde{\mathbf{x}})=c} \prod_{i=1}^{n^2} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}'_i) \right) - \left(\sum_{\tilde{\mathbf{x}} \text{ s.t. } h(\tilde{\mathbf{x}})=c} \prod_{i=1}^{n^2} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) \right) \\
&= \min_{h(\mathbf{x})\in\mathcal{H}} \sum_{\substack{\tilde{\mathbf{x}} \text{ s.t.} \\ h(\tilde{\mathbf{x}})=c}} \left[\left(\prod_{i=1}^{n^2-1} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}'_i) \right) \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}'_{n^2}) - \left(\prod_{i=1}^{n^2-1} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) \right) \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}_{n^2}) \right] \\
&= \min_{h(\mathbf{x})\in\mathcal{H}} \sum_{\tilde{\mathbf{x}} \text{ s.t. } h(\tilde{\mathbf{x}})=c} \left(\prod_{i=1}^{n^2-1} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) \right) \underbrace{\left(\mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}'_{n^2}) - \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}_{n^2}) \right)}_{\Delta_{\tilde{\mathbf{x}}}}
\end{aligned}$$

$\Delta_{\tilde{\mathbf{x}}} = \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}'_{n^2}) - \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}_{n^2})$ resolves to two cases (each case occurs 50% of the time): (1) $1-(p_a+p_d)$ and (2) $p_a + p_d - 1$. To minimize $g(\mathbf{x}')_c - g(\mathbf{x})_c$ we now choose $h^*(\mathbf{x})$ to predict c for all cases where $\Delta_{\tilde{\mathbf{x}}} < 0$ (assuming $p_a + p_d \neq 1$). Hence, $\Delta = \Delta_{\tilde{\mathbf{x}}}$ for $\tilde{\mathbf{x}}$ s.t. $h(\tilde{\mathbf{x}}) = c$.

We conclude the worst-case base classifier $h^*(\mathbf{x})$ exactly classifies exactly 50% of the random graphs $\tilde{\mathbf{x}}$ with c (note that in the general case $g(\mathbf{x})_c \neq 1/2$). In the case where one edge is removed from $\mathbf{x}'$ (relatively to $\mathbf{x}$) and $p_a + p_d < 1$, the worst case base classifier $h^*(\mathbf{x})$ predicts c for all graphs where this edge is not missing (e.g. $h^*(\mathbf{x}) = c$ and $h^*(\tilde{\mathbf{x}}) \neq c$).

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?

Since we have n^2 edges there are 2^{n^2} possible adjacency matrices (each adjacency matrix represents one graph). Since we predict 50% with class c, we have a total of $2^{n^2}/2 = 2^{n^2-1}$ graphs resulting in c.

This clearly shows that enumerating all possible $\tilde{\mathbf{x}}$ is infeasible also for very small graphs.

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 $p_d = 0.1$.

Since $p_a + p_d < 1$ we conclude that $\Delta = p_a + p_d - 1$.

$$\begin{aligned}
\min_{h(\mathbf{x})\in\mathcal{H}} g(\mathbf{x}')_c - g(\mathbf{x})_c &= \min_{h(\mathbf{x})\in\mathcal{H}} \sum_{\tilde{\mathbf{x}} \text{ s.t. } h(\tilde{\mathbf{x}})=c} \left(\prod_{i=1}^{n^2-1} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i) \right) \underbrace{\left(\mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}'_{n^2}) - \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}_{n^2})\right)}_{\Delta} \\
&= \min_{h(\mathbf{x})\in\mathcal{H}} \Delta \underbrace{\sum_{\tilde{\mathbf{x}} \text{ s.t. } h(\tilde{\mathbf{x}})=c} \prod_{i=1}^{n^2-1} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i)}_{=1} \\
&= p_a + p_d - 1
\end{aligned}$$

Please note that $\sum_{\tilde{\mathbf{x}} \text{ s.t. } h(\tilde{\mathbf{x}})=c} \prod_{i=1}^{n^2-1} \mathcal{P}(\tilde{\mathbf{x}}_i|\mathbf{x}_i)$ can be understood as a sum over the entire sample space of a product of $(n^2 - 1)$ Bernoulli random variables (i.e. sum over all possible combinations). Due to the basic laws of probability it must sum up to one.

Using $\Delta = \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}'_{n^2}) - \mathcal{P}(\tilde{\mathbf{x}}_{n^2}|\mathbf{x}_{n^2})$ s.t. $h^*(\tilde{\mathbf{x}}) = c$, we can easily go back and fourth between $g(\mathbf{x}')_c$, $g(\mathbf{x})_c$, and $g(\mathbf{x}')_c - g(\mathbf{x})_c$. Consequently, the the worst-case base classifier, with the given flip probabilities $p_a = 0.001$ and $p_d = 0.1$, has the following probabilities:

- $g(\mathbf{x}')_c = p_a = 0.001$
- $g(\mathbf{x})_c = 1 - p_d = 0.9$
- $g(\mathbf{x}')_c - g(\mathbf{x})_c = p_a + p_d - 1 = -0.899$

Please acknowledge that a smooth classifier might predict the right class c with high probability $g(\mathbf{x})_c = 1 - p_d = 0.9$, but flipping a single edge can result in $g(\mathbf{x}')_c = p_a = 0.001$. Hence, the probability of the smooth classier drops by around 90%.