Machine Learning for Graphs and Sequential Data Exercise Sheet 05

Robustness of Machine Learning Models II

Problem 1: On slide 15 of the robustness chapter, we have defined an optimization problem for untargeted attacks, i.e. we aim to have the sample $\hat{\boldsymbol{x}}$ classified as **any** class other than the correct one:

$$\min_{\hat{\boldsymbol{x}}} \mathcal{D}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda \cdot L(\hat{\boldsymbol{x}}, y)$$

The loss function is defined as:

$$L(\hat{\boldsymbol{x}}, y) = \left[Z(\hat{\boldsymbol{x}})_y - \max_{i \neq y} Z(\hat{\boldsymbol{x}})_i \right]_+ ,$$

where $[\boldsymbol{x}]_+$ is shorthand for $\max(\boldsymbol{x}, 0)$ and $Z(\boldsymbol{x})_i = \log f(\boldsymbol{x})_i$ (i.e. log probability of class i. Here, $L(\hat{\boldsymbol{x}}, y)$ is positive if $\hat{\boldsymbol{x}}$ is classified correctly and 0 otherwise.

Provide an alternative loss function to turn this attack into a targeted attack, i.e. we aim to have the sample $\boldsymbol{x}$ classified as a *specific* target class t.

$$L(\hat{\boldsymbol{x}}, t) = \left[\max_{i \neq t} Z(\hat{\boldsymbol{x}})_i - Z(\hat{\boldsymbol{x}})_t \right]_+$$

This loss is positive if $\hat{\boldsymbol{x}}$ is classified as a class that is **not** t, and is zero otherwise.

Problem 2: Recall from slide 41 the MILP constraints expressing the ReLU activation function. Show that a continuous relaxation on $\boldsymbol{a}$ leads to the convex relaxation constraints on slide 54. That is, we replace the constraint $\boldsymbol{a}_i \in \{0, 1\}$ with $\boldsymbol{a}_i \in [0, 1]$.

We can combine the first two constraints on slide 41:

$$\begin{aligned} \boldsymbol{y}_i &\leq \boldsymbol{x}_i - \boldsymbol{l}_i(1 - \boldsymbol{a}_i) \\ \boldsymbol{y}_i &\leq \boldsymbol{u}_i \cdot \boldsymbol{a}_i \end{aligned}$$

by expressing it as

$$\boldsymbol{y}_i \leq \min(\boldsymbol{x}_i - \boldsymbol{l}_i(1 - \boldsymbol{a}_i), \boldsymbol{u}_i \cdot \boldsymbol{a}_i)$$

Note that we are free to choose any value for $\boldsymbol{a}_i$ between 0 and 1. We want to choose $\boldsymbol{a}_i$ so that it leads to the loosest-possible constraint on $\boldsymbol{y}_i$, since this leads to the maximum 'leeway' to optimize the objective function. More formally,

$$\boldsymbol{y}_i \leq \max_{\boldsymbol{a}_i} \min(\boldsymbol{x}_i - \boldsymbol{l}_i(1 - \boldsymbol{a}_i), \boldsymbol{u}_i \cdot \boldsymbol{a}_i)$$

Further note that the two terms in the $\min(\cdot, \cdot)$ are two linear functions in $\boldsymbol{a}_i$. For $\boldsymbol{l}_i < 0$, the former is a function with negative slope in $\boldsymbol{a}_i$. We only need to consider the case $\boldsymbol{l}_i < 0$, since if $\boldsymbol{l}_i \geq 0$, we know that the unit is *stably active* and therefore linear.

The second term in the $\min(\cdot,\cdot)$ is a function with positive slope in $\boldsymbol{a}_i$ if $\boldsymbol{u}_i > 0$. Again, we only need to consider this case since $\boldsymbol{u}_i \leq 0$ implies that the unit is *stably inactive* and therefore $\boldsymbol{y}_i = 0$.

Consequently, the function $\min(\boldsymbol{x}_i - \boldsymbol{l}_i(1-\boldsymbol{a}_i), \boldsymbol{u}_i \cdot \boldsymbol{a}_i)$ maximal at the intersection of the two linear functions. Solving for $\boldsymbol{a}_i$ we get:

$$\begin{aligned} \boldsymbol{x}_i - (1-\boldsymbol{a}_i)\boldsymbol{l}_i &= \boldsymbol{a}_i\boldsymbol{u}_i \\ \Leftrightarrow \boldsymbol{a}_i &= \frac{\boldsymbol{x}_i - \boldsymbol{l}_i}{\boldsymbol{u}_i - \boldsymbol{l}_i} \end{aligned}$$

Plugging the expression of $\boldsymbol{a}_i$ into one of the original constraints, e.g. $\boldsymbol{y}_i \leq \boldsymbol{a}_i \cdot \boldsymbol{u}_i$ we get:

$$\begin{aligned} \boldsymbol{y}_i &\leq \frac{\boldsymbol{x}_i - \boldsymbol{l}_i}{\boldsymbol{u}_i - \boldsymbol{l}_i}\boldsymbol{u}_i \\ \Leftrightarrow \boldsymbol{y}_i(\boldsymbol{u}_i - \boldsymbol{l}_i) - \boldsymbol{u}_i\boldsymbol{x}_i &\leq \boldsymbol{u}_i\boldsymbol{l}_i, \end{aligned}$$

and therefore we have recovered the constraint of the convex relaxation.