Resistance of accumulated local effects to data poisoning attacks

Mateusz Błajda, Michał Krutul, Maciej Nadolski

Tools for analyzing model performance are broadly used to understand ML models better. Two of them are PDP and ALE[2]. In [1] it was shown that PDP can be fooled by poisoning data in that way, so PDP is generating a different plot from created original data, but keeping some features intact. That project consisted of two variants of attack focusing on different goals:

- targeted attack - poison data, so its explanation looks as similar as given target function
- robustness test - change data, so its explanation is as different from the original one as it is possible

In this project we ran experiments on gradient tasks similar to those conducted in [1], but replaced PDP with our implementation of ALE. Experiments were conducted on the heart dataset, using gradient algorithm.

Results:
PD is clearly easier to fool than ALE, we can achieve almost exactly the target explanations. With ALE it’s harder to fool it so that it fits the target, however we can still influence it significantly. (see the images)

What was done:

- rewrote functions to keras/tensorflow for easier use and cleaner code
- implementation of ALE
- adjusting ALE to work with tensorflow
- libraries dependencies updates to newer versions
- experiments run on Heart dataset using gradient algorithm

The difficulties:

- gradients did not want to calculate - we are not an experts in tensorflow

Next steps:

- Implementation of populational version of ALE for genetic algorithm
- Implementation of ALE++
- Easier comparison of ALE with PDP
- Run more experiments with different hyperparameters
- Add bike-sharing dataset referenced in ALE paper

References:
[1] Fooling Partial Dependence via Data Poisoning, Baniecki et al 2021, https://doi.org/10.48550/arxiv.2105.12837
[2] ALE paper: https://arxiv.org/abs/1612.08468