Better Sampling In Lime To Defense Against Adversarial Attacks

Trieu Vy Tran, Anh Khoa Pham

LUH|AI

Leibniz Universität Hannover

Introduction

- Users demand transparency of ML models, which drives developement of post-hoc explanation methods like LIME.
- Lime's reliance on perturbation sampling troduces serious weakness: perturbation alters data distribution.
- Manipulation is possible, resulting in bised or discriminatory decisions.
- Owners of sensitive prediction could hide socially biases present in the model.
- Variational autoencoders (VAE) as advanced sampling technique is introduced in LIME, making the modified explanation method gLIME more resistant to manipulation attempts.

COMPAS dataset

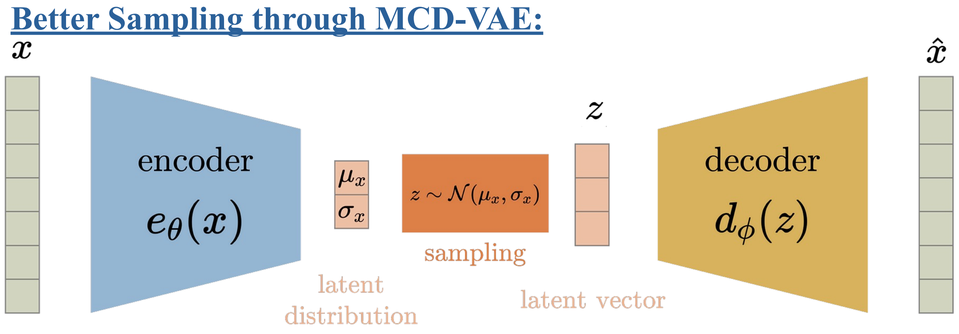
- Dataset used to determine the crime recurrence risk of a defendant.
- Dataset includes information such as criminal history, time in prison,age, gender, race,...of 6172 defendants.
- The sensitive feature in the dataset is „race", the adversarial model will be biased on this feature.

Generators

Traditional Sampling Technique:

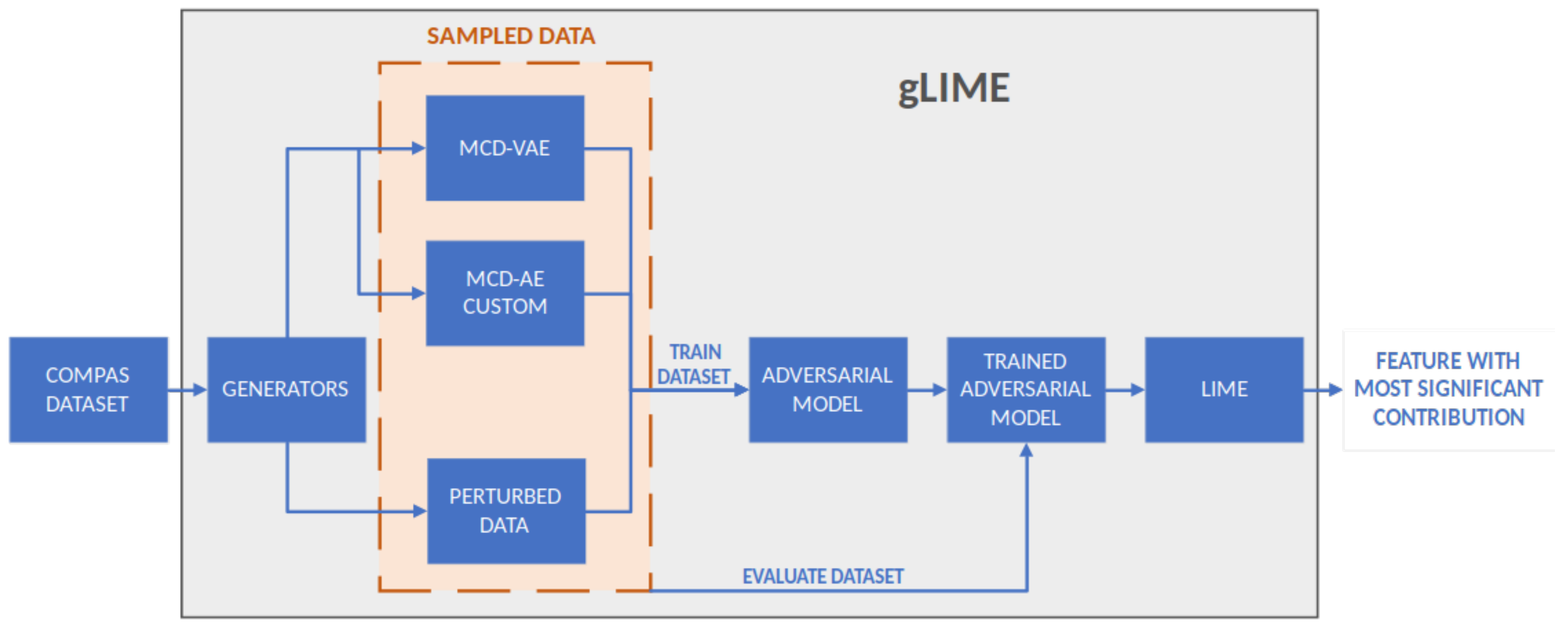
- Perturbed instances are generated by adding Gaussian noise to each feature of x independently.

Better Sampling through MCD-VAE:



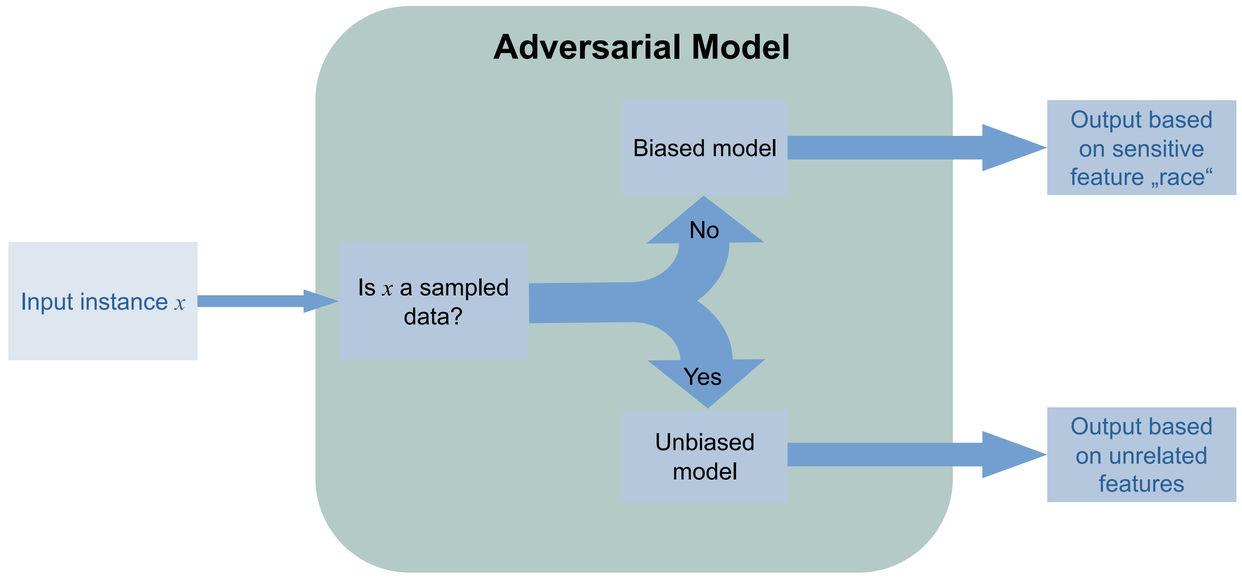
- Variational Autoencoder consists of two neural networks called encoder and decoder.
- The encoder compresses the input instances and they are reconstructed to the orignial values with the encoder.
- Three sampled dataset were generated: one dataset through perturbation, two datasets generated by MCD-VAE, one with parameters from original paper, one with our custom parameters.
- Datasets are split into train and evaluate datasets. Train datasets are used to train the adversarial models, resulting in trained adversarial models. The evaluate datasets were utilized as input for the trained adversarial models.

gLIME



- The COMPAS dataset undergoes processing through generators (perturbation and MCD-VAE) to produce sampled datasets.
- The train datasets are utilized to train the adversarial models, resulting in trained adversarial models. The evaluate datasets serve as input for the trained adversarial models.
- Subsequently, the output is passed through LIME to determine the features with the most significant contribution.

Adversarial Model



- The input instance x is decided by the decision model (Random Forest Classifier) as either a sampled data or a original data.
- If the decision model decides that the instance is from original distribution, the biased model will be used. Otherwise, the unbiased model will be used when the instance is decided to be a sampled instance.
- The output of the biased model is based on sensitive feature "race", while the output of the unbiased model is biased on unrelated features.

Key Insights

- MCD-VAE could perform sampling on a given instance to generate sampled instances.
- Adversarial model with biased model depending on sensitive feature could act as a black box model in applying explanation methods such as LIME.
- Re-implementation an algorithm from a paper is still challenging, especially for algorithms that are not kept up-to-date.
- LIME depends on generated instances, which are affected by the quality of the generators, LIME could be more robust with optimizied generators.
- Hyperparameters optimization plays a significant role in machine learning method.
- Team work is an importance aspect.

Results

- MCD-VAE is proven not to be a suitable approach to generate sample data.
- The desired results have not been met.
- MCD-VAE could be implemented diffrently from the original approach. Although it still not successful in the re-implementation, it could be promising, since autoencoders only need to optimize its reconstruction loss.
- Optimizing hyperparameters of MCD-AE and MCD-VAE is still a challenge in general.
- The unbiased model of the adversarial model depends on randomness of the unrelated features, its performance is not ensure to be stable.

Reference

Vreš, D. and Robnik Šikonja, M. (2020) Better sampling in explanation methods can prevent dieselgate-like deception Submitted to International Conference on Learning Representations
https://arxiv.org/pdf/2101.11702.pdf