Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations

Abstract

Methods for explaining black-box machine learning models aim to increase the transparency of these model and provide insights into the reliability and fairness of such models. However, the explanations themselves could contain significant uncertainty that undermines users' trust in the predictions and raises concern about the model's robustness. Focusing on a particular local explanations method, Local Interpretable Model-Agnostic Explanations (LIME), we demonstrate the presence of three sources of uncertainty, namely randomness in the sampling procedure, variation with sampling proximity, and variation in explained model credibility across different data points. Such uncertainty is present even for black-box models with high test accuracy. We investigate the uncertainty in the LIME method on synthetic data and two public data sets, newsgroups text classification and recidivism risk-scoring.

Local Interpretable Model-Agnostic Explanations

Given a black box model f, and a target point x to be explained, LIME samples neighbors of x and their black-box outcomes and chooses a model g from some interpretable functional space G by solving

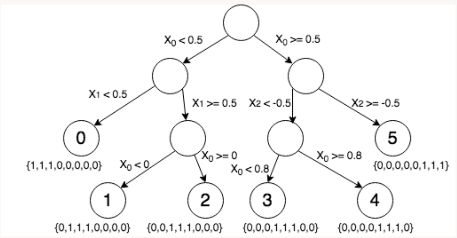
$$\text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

where π_x is some probability distribution around x and $\Omega(g)$ is a penalty for model complexity.

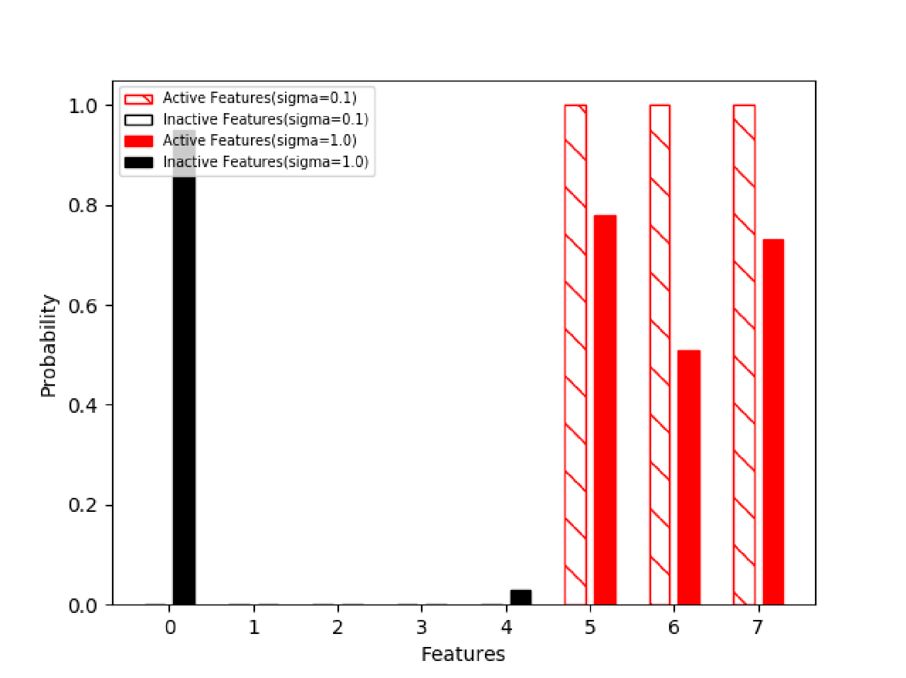
Sources of uncertainty in LIME

- Sampling variance in explaining a single data point;
- Sensitivity to choice of parameters, such as sample size and sampling proximity;
- Variation in explanation on model credibility across different data points.

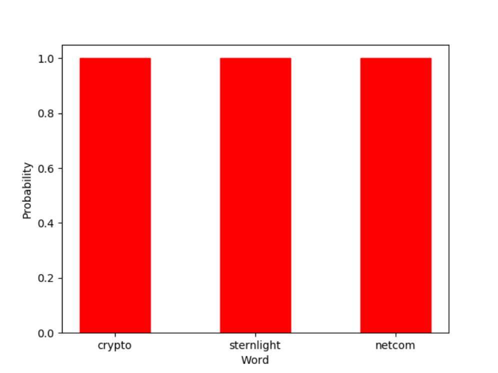
Simulation Setting



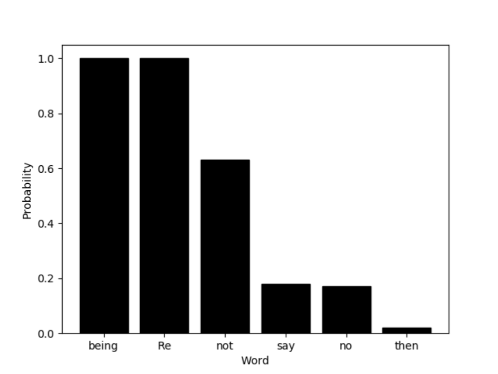
Simulation Setting



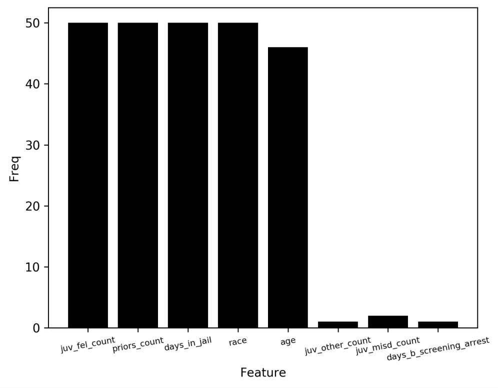
Text Data Example 1



Text Data Example 2



COMPAS Example



Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, Madeleine Udell