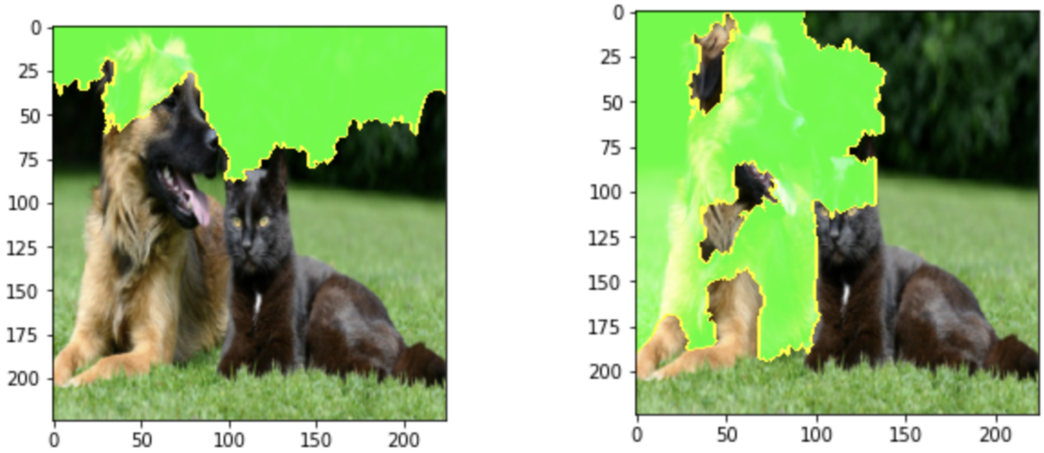
Machine Learning Project Final Report: A Theoretical Analysis of the Comparison between LIME and SHAP

Ruonan (Elizabeth) Zhao
rz1280[at]nyu[dot]edu

December 11, 2019



(a) $\sigma = 2e - 6$

(b) $\sigma = 0.75$ (default)

Figure 1: Testing the explanation on trusted model with $\pi_{x'} = \exp\left(-D^2/\sigma^2\right)$

Abstract: One of the major concerns in Machine Learning model predictions is whether we should trust the prediction. The trust of a prediction depends on the interpretability, which can be equally important as the accuracy of the prediction. Two of the famous algorithms, Local Interpretable Model-agnostic Explanations (LIME) (Riberito, et al. 2016) and SHapley Additive exPlanations (SHAP) (Lundberg-Lee 2017), ”provides the explanation of an individual prediction model as a solution of ’trusting predictions’” (1). According to Lundberg and Lee, SHAP is a more general approach than LIME. In this project, we look into the difference between the two methods to find out which one can explain the true model better.

1 Introduction

1.1 Basic Notations

Let $x \in S$ be the input variable, where S is any dataset. Let $x' \in \{0,1\}^d$ be a simplified input of x. Let $z \in \mathbf{R}^d$ be the preturbed input variable of x. Let $z' \in \{0,1\}^{d'}$ be the simplified input of z. Let $g \in G$ be the model represents the explanation of the true model, where G is a class of interpretable models. The domain of g is $\{0,1\}^d$. The measure of complexity of the explanation $g \in G$ is given by $\Omega(G)$. Let f be a function such that $f : \mathbf{R}^d \longrightarrow \mathbf{R}$. f is the probability that x belongs to a certain class. Let $\pi_x(z)$ be a proximity measure between instance z and x, and it measures how large the neighborhood around instance x we want for the explanation. Let $\mathcal{L}(f, g, \pi_x)$ be a measure of how unfaithful g is when approximating f in the locality defined by π_x. ϕ_i is defined to the feature importance, the explanation of a model, or Shapley values (2).

1.2 Properties

The following properties are collected from the SHAP paper (2).

Property 1 (Local Accuracy).

$$f(x) = g(x') = \sum_j \phi_j x'_j$$

The explanation of the model at the simplified input x' should at least equal to the original model at the original input x. The model may not have good global approximations.

Property 2 (Missingness). *For any j,*

$$x'_j = 0 \implies \phi_j = 0$$

The missingness shows that the particular explanation will not show up if its simplified input is absent.

Property 3 (Consistency). *Let $f_x(z') = f(h_x(z'))$ and denote $z'_{\{z_i=0\}}$ be z' where $z'_i = 0$. For any two models f^1 and f^2, if*

$$f^1_x(z') - f^1_x\left(z'_{\{z_i=0\}}\right) \geq f^2_x(z') - f^2_x\left(z'_{\{z_i=0\}}\right)$$

for all $z' \in \{0,1\}^d$, then

$$\phi_i(f^1, x) \geq \phi_i(f^2, x)$$

The consistency tells us that if two original models are different at the missing simplified input, then the model explanation will be sensitive to the difference.

1.3 Linear LIME

The explanation, $\phi_j(x)$, of linear LIME is given by the following function (1):

$$\xi(x) = \operatorname*{argmin}_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

where $\mathcal{G}$ is a class of the linear explanation model g, and

$$\begin{aligned} g(z') &= \sum_j \phi_j z'_j \\ \pi_x(z) &= e^{-D(x,z)^2/\sigma^2} \\ \mathcal{L}(f,g,\pi_x) &= \sum_{z,z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \\ \Omega(g) &= 0, \end{aligned}$$

where $z' \in \{0,1\}^d$ and D is the measure of distance between x and z.
LIME is model-agnostic, i.e., the choice of model f is irrelevant (1).

1.4 SHAP

The feature importance of Linear SHAP is given by

$$\phi_j(f_x) = \sum_{z' \subseteq x'} \frac{|z'|!(d - |z'| - 1)!}{d!} \left(f_x(z') - f_x(z' \setminus x')\right),$$

where $|z'|$ is the number of non-zero entries in z', $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x', and $f_x(z') = f(h_x(z')) = E[f(z)|z_s]$ (2).

1.4.1 Linear SHAP

Since the true SHAP value is challanging to compute numerically, a Linear SHAP approximation is provided from the following,

$$\phi_j(f,x) = \phi_j(x_j - E[x_j]),$$

where f is a linear model (2).

2 LIME v. SHAP

SHAP paper attacks LIME by claiming that property (1-3) fails to hold when the parameters $\mathcal{L}, \pi_x$, and Ω are chosen "heuristically" (2). Therefore, LIME does not always work. To see the claim holds, our first step is to empirically examine it by experimenting with different parameters. Once we have the desired result, we can start to analyze the claim mathematically.

2.1 An Empirical Approach

The idea is simple. We use a previously well-trained model with the same set of data from the open source of LIME Image Explainer, and test the LIME method on a similar testing image. We observe that the LIME algorithm uses a default kernel $\pi_x = \exp\left(-D^2/\sigma^2\right)$ with a fixed kernel width $\sigma = 0.75$. By modifying the kernel width and/or kernel while keeping other parameters fixed, we expect to see a failed result, i.e., the prediction explains everything else other than the target. In addition, by the previous definition of the kernel, we restrict it to be nonnegative and be less and equal to 1.

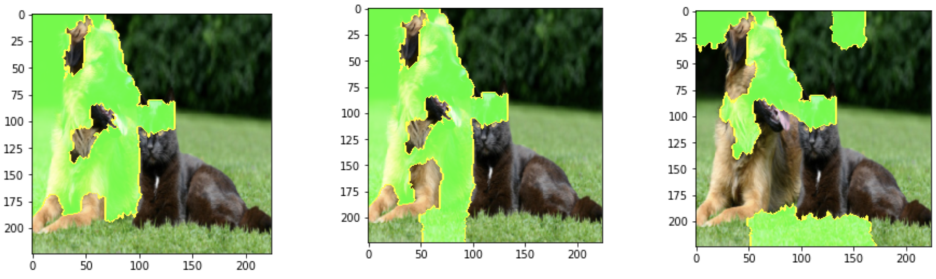
2.2 A Theoretical Intuition

By the end of the project deadline, we are unable to provide a rigorous proof for the claim above [2]. Nevertheless, we shall intuitively discuss about why LIME does not work for a certain choice of π_x based on our observations in section 3.

Assume we are given by the default kernel $\pi_x(z) = exp(-D(x,z)^2/\sigma^2)$. We sample the simplified x' with an arbitrarily small distance σ around x. If the distance between the original input x and the perturbed input z is nonzero, then $\pi_x(z)$ is arbitrarily close to 0. Therefore, the LIME loss function $\mathcal{L}(f,g,\pi_x)$ [1.3] will approach to 0, whereas the squared distance between the true model and the explanation model $(f(z) - g(z'))^2$ can be large. Therefore, we might have a contradiction of the local accuracy [1]. For more mathematical details, see Appendix [6].

3 Results

By testing on a trusted model with the default kernel and an extremely small value of kernel width $\sigma = 2e - 6$, the prediction indeed **fails** to explain the dog (Figure 1(a)).



(a) $\sigma = 2000$

(b) $\sigma \approx 1.17$

(c) $\sigma = 2e - 6$

Figure 2: Testing the explanation on trusted model with Cauchy kernel $\pi_{x'} = \frac{1}{1+D^2/\sigma^2}$ with different values of σ.

We also tried modifying the kernel in the source code with some common ones (3), such as Cauchy kernel. It may be a special case, even though we can get some interesting results, LIME can work regardless how small the kernel width is.

4 Conclusions and Future Works

Our experimental results show that, with some extremely small values of kernel width σ, linear LIME method does not work well in general. We still need to analyze the violation of local accuracy, missingness, and consistency of LIME through rigorous mathematical proofs. We also do not know for how large of the scale of local accuracy is being violated can cause the failure of the LIME explanation. Although SHAP can work better than LIME, the SHAP paper (2) only provides experimental results of the comparison between the linear LIME and the kernel SHAP on decision tree models. A more detailed proof of linear SHAP working well on linear models is desired. Furthermore, We are curious about whether linear SHAP can work on nonlinear models, and we shall look into it in future.

5 Acknowledgements

Special thanks to my amazing teaching assistants, Mark Goldstein, Aahlad Manas Puli, Mukund Sudarshan, who patiently answered my questions and provided me inspirations and tons of useful tips for this project.

References

[1] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[2] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[3] César R. Souza. Kernel functions for machine learning applications, 2010. Online; accessed December-13-2019.

6 Appendix

Some mathematical details of the discussion about the failed explanation of LIME [2.2] is shown below.

Let $D(x, z) > 0$. Then, $\forall x, z \in \mathbf{R}^d$,

$$\pi_x(z) = e^{-\frac{D^2(x,z)}{\sigma^2}} \longrightarrow 0 \quad \text{as} \quad \sigma \longrightarrow 0$$

Then, $\forall z, z' \in Z$,

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(z)(f(z) - g(z'))^2 = \sum_{z,z' \in Z} 0 \cdot (f(z) - g(z'))^2 = 0$$

Although $\mathcal{L}$ reaches its minimum 0, we have no information about the squared distance $(f(z) - g(z'))^2$. Thus, we cannot guarantee that the explanation model can approximate the true model, i.e., $g(z') = \sum_j \phi_j z'_j \neq f(z)$, for some z' and z. This can give us an insight of the local accuracy property [1] being violated. However, we hesitate to confirm that it is a solid proof. We will keep working on it in the future.