Juan Ceballos

CS301-104

4/25/21

LIME Method Tutorial

LIME(Local Interpretable Model-Agnostic Explanations) is an algorithm that approaches a solution to a big problem in machine learning. This problem is that there is no way for a user to trust a prediction and model. It is easier for a user to trust a model if the model can be presented in a way for a human to understand its behavior. LIME seeks to present these models faithfully, but also in a fashion that is easier for humans to interpret by first explaining individual predictions. Explaining multiple predictions will help gain trust for a model.

A classification model is vulnerable to failure when performing real world data. Models are built using prepared datasets that may not correctly represent real world datasets that the model will be given. This could for example cause an issue with the model's accuracy sensitivity. Giving positives for incorrect reasons which when used in a field such as medicine is a situation that should be avoided. Finding this issue may prove difficult which is why LIME provides a solution. Explaining the model in an interpable fashion could help a human understand the flaws in a particular algorithm when having to choose from multiple. For example, let's say you had two models that had to predict if an article was about fruits or vegetables. Both correctly predict the article is about vegetables, but the first model put most importance into the word SALAD, while model 2 put more importance into the words GREEN, FOOD, and RECIPE. While both are correct, a human can look at this and reasonably assume that model 2 is more sustainable to face a false positive due to it placing more importance into ambiguous terms. Understanding the

model's process for its prediction is key to building trust in that model. When it comes to what is considered interpable is something that also requires consideration. That's why LIME needs to be able to function with any model, but also be able to accurately represent the model at least in terms of how it was used in an individual prediction.

The first part of LIME is interpretable data representations, how the data needs to be presentable to humans regardless of how the model works such as text classification represented by the existence and absensence of key terms or image classification which can be represented by existence/absence of particular pixels. $\xi(x) = g \epsilon G\ L(f, g, \pi_x) + \Omega(g)$ is the formula explained by LIME which gives flexibility to different explanation families, fidelity functions, and complexity measures. The goal is to minimize L which represents the loss of faithfulness to the model while Ω represents how complex the explanation is. LIME has to remain open to use from different models and as such, assumptions cannot be made of the model so in order to reduce loss, L is approximated by using samples weighted by π_x. Using low and high weights with samples far and close to x. Using this will give an explanation while remaining locally faithful to the model.

Using a linear model for the loss function would look similar to this kind of function: $\Sigma\, exp(-D(x,z)^2/\sigma^2)(f(z) - g(z'))^2$. Using this for text classification, the interpretable representation would be a group of words limited by K, which would be used as a constant. Using Lasso to determine K and getting weights from least squares, the explanation's complexity is determined by how long it takes to compute the model with N samples. So for example 1000 trees with 5000 samples would take around 3 seconds if unoptimized.

Again this just a means of explaining one prediction, LIME also seeks to apply trust to a model as a whole. To build trust for a model, multiple prediction explanations are necessary for a person to get a better understanding of the model's behavior. A person however may not have the time to go through every explanation. So the goal is to be selective with the explanations given. So the idea is to give the best range of the model's capabilities with the fewest reasonable amount of explanations. So to do this a matrix needs to be made which contains explanation importance per instance. So for a model using text the function would be $I_j = \sqrt{\Sigma_{i=1}^{n} W_{ij}}$ in which it presents the importance of an instance and from those instances those with more importance take priority. The right set of instances to choose also depends on range and diversity. It's best to avoid as much overlap as possible. Using this method LIME is able to properly budget a person's time while still remaining faithful to the model's capabilities.

While LIME gives methods to apply explanations to individual predictions and models, there also needs to be testing of the efficiency of how much trust these explanations are capable of creating. If LIME doesn't build significant trust to a model as to view a model without using LIME then the method would fall apart. It's found that when simulating users, that LIME has a higher recall when compared to the parzen method. In terms of trusting a model, the simulated users choose the correct model when pressed which classifier was worth trusting more between two sets of explanations. When using human users, they were given an intentionally bad model and were asked if they should trust the model. When presented with no explanation, a lot less than half trusted the model. When given an explanation, a significant portion of those that trusted the model was reduced.

LIME as a method provides versatility to be used by multiple classifiers and models while still remaining faithful to how these models behave. It provides users to gain trust in the efficiency of a prediction and model by giving understandable explanations to how a prediction and model came to its conclusion.