Ahmed Mohamed, Khoa Nguyen, Michael Vythilengam

Professor Monogioudis

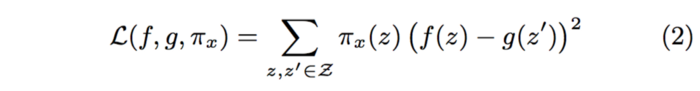
CS 301-104

25 April 2021

Summary on the LIME method

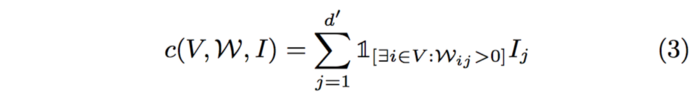
In recent years, we have seen great advancements in machine learning. We have learned how to create and manipulate the way machines learn things and how they should react to real world scenarios. Overall, it has shown promising results for the future. As each year goes by, we get closer and closer to a futuristic world as it's portrayed in movies. We may not have robots with human-like intelligence yet, but we do see many companies like Facebook, Amazon, Apple, Tesla, and many others integrating AI into their products and services. However, to have reached this point, many years of ongoing research and testing have been conducted. When we look at the root of machine learning, you will see that it's all about the datasets that are fed into these different learning algorithms. Many of which we learned about in class. Each algorithm (or method) has its own pros and cons, and each different algorithm tries to improve certain aspects that past algorithms failed in. One thing about all these different algorithms is that at some point, we must be able to trust the predictions that the models come up with. However, the main issue with these models is that they are still black boxes and understanding how these predictions are made is very important in machine learning. This is where the LIME method comes into action by making the models more interpretable by producing local faithfulness.

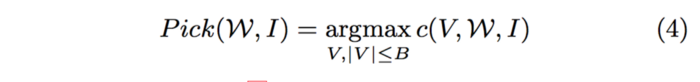
The LIME (Local Interpretable Model-Agnostic Explanations) algorithm has the capability of explaining any prediction from any “classifier or regressor in a faithful way, by approximating it locally with an interpretable model” and then SP-LIME is a method that selects a set of explanations that will make sure a model behaves as it it was human logic when it comes to decision making. Thanks to this method, anybody whether you're an expert in machine learning or have little to no knowledge about it, you will be able to understand how these predictions were made no matter what classifier was used to come up with them. What makes LIME a good model explainer are the following qualities:

1. Explanations are interpretable - a representation that can be easily understood by humans no matter what features are present and making the predictions whether certain features are present or absent depending on the type of representation (text or image). For text, the predictions are based on the absence or presence of certain words while with images the predictions are based on the absence or presence of superpixels (“a contiguous patch of similar pixels”)
2. Local fidelity - we replicate the model’s behavior locally. To do so we ensure interoperability and local fidelity by minimizing the locality-aware loss which is done with the following equation:
3. Sampling for local explanation - to be able to minimize the locality-aware loss without making assumptions about *f* we use random uniform sampling x` which is later on weighted by Pi(x) making it possible to optimize the above equation
4. Sparse linear explanation - we set K as a limit on the number of words (or superpixels if an image is being used) that will be used for the explanation. The complexity is not based off the size of the dataset but rather the time it takes to compute f(x) and the number of samples N

Submodular Pick (SP-LIME) as mentioned before, picks a set of explanations which will provide global understanding of the model making it easier for us to be able to ensure our trust into a model or not. To use this method, we first represent the time and patience humans will have as budget *B* “that denotes the number of explanations they are willing to look at to understand the model”. We then move onto the pick step where the task of selecting specific instances out the set *B* for user inspection is conducted. These instances should be “non-redundant explanations that represent how the model behaves globally”. We then come up with the following

* An explanation matrix *W* which will represent local importance of the interpretable components
* Let I(j) denote the global importance of this component in the explanation space
* Given *W* an *I* we compute the total importance of the features that appear at least once in an instance and add it to set *V*



* We then maximize the weighted coverage function

In the article we see an example experiment where multiple different methods were used to classify two datasets (books and DVDs) by whether they had good or bad reviews. They trained “decision trees, logistic regression with L2 regularization, nearest neighbors and supposed vector machines with RBF kernel”. Then four different approaches were used to explain individual predictions. These four are:

1. LIME: explained beforehand
2. Parzen: explains the black box classifier globally “by taking the gradient of the prediction probability function”. We then take K features with the highest absolute gradients as the explanation
3. Greedy: which essentially removes any features that contribute the most to the predicted class until the prediction changes (or reach the max of *K* instances).
4. Lastly a *random* procedure that randomly picks *K* features as an explanation

As we saw in the results for sparse logistic regression and decision trees, LIME consistently provided a >90% recall which demonstrates that the explanations are faithful to the models. Furthermore, they did another test where they manipulated the features to be either trustworthy or untrustworthy to see which of the four classifiers would perform better. It was shown that LIME greatly outperformed the other three methods which allows it to maintain both high precision and high recall. We then got into the topic of how even non-experts would be able to use LIME to identify irregularities in these explanations when it came to the “Husky vs Wolf” experiment. Even though some of the predictions were incorrect, knowing why it made that incorrect prediction is what helps us trust the classifier and also demonstrate what should be fixed.

In conclusion, it is important to understand how exactly machine learning works to be able to one day fully trust and go with the predictions/decisions that they make. LIME gives experts and non-experts a better sense of understanding by faithfully explaining the predictions of any model in an interpretable manner. While we also introduced SP-LIME which is a method that selects representative and non-redundant predictions to be able to provide a global view of the model to users. All of these explanations are essential to providing experts and non-experts an insight into these classifiers and how we can further improve our trust in them. This will definitely help us improve machine learning at a faster pace while minimizing most discrepancies.