CS-301 Project Team 1 (Jincheol Jeong, Avinash Kumar, James Podeszfinski)
Dr. Pantelis Monogioudis
CS 301
April 15 2021

LIME method in "Why Should I Trust You?"

Arguably, machine learning is the greatest invention of the 21st century. It is used in many different fields such as the stock market, disease detection, and malware detection. Considering how powerful machine learning is in different industries, many people just assume that machine learning predictions are always correct. However, machine learning is only correct with properly trained models and it is really hard to verify the accuracy of the models because models are like a black box. Due to this property of machine learning models, it is hard to tell if the model made a correct prediction based on the relevant information or a random guess.

For example, let's suppose that a person came out with a model that can predict whether a person has cancer or not. This model might take various inputs such as weight, blood sugar level, MRI results, etc. Now, suppose that this model predicted that one patient has cancer. If we found that the person indeed has cancer in future examinations, then can we trust that this model is good for cancer detection?

The answer to the question above depends on how the models made the decision. If a model made a prediction that the person has cancer mainly because a person is overweight or underweight, then we cannot trust the model for obvious reasons. So, we want to change the model so that the model can determine the output based on medically meaningful factors. However, due to the black box nature of the machine learning models, this problem is not trivial to solve. In fact, it is not even easy to see why the model made such a decision.

In the paper "*Why Should I Trust You?*", the authors suggest a way to systematically determine whether the models made a decision based on correct factors or guesses turned out to be true. In the paper, the Local Interpretable Model-Agnostic Explanations (LIME) is presented to identify an interpretable model over the interpretable representation that is locally faithful to the classifier. (Ribeiro 3)

LIME, as its name suggests, uses four different components to open the black box (machine learning model) and see what is inside. After we see why models made predictions, we can decide whether the model is sound or not. As a result, we can rely on the predictions more rationally than simply relying on the predictions because it gave the correct results.

In LIME, L stands for local. However, it is very unclear what local means in LIME. So, just for the sake of understanding, it is better to see why E stands for explanations in LIME. Explanations is simply an understanding of relationships between inputs to the model and the predictions. For example, if a model is used to differentiate between husky pictures and wolf pictures, then the explanation should give us why the prediction was made. Explanations can tell us that the model classified the picture as wolf because there was snow in the background.

(a) Husky classified as wolf (b) Explanation

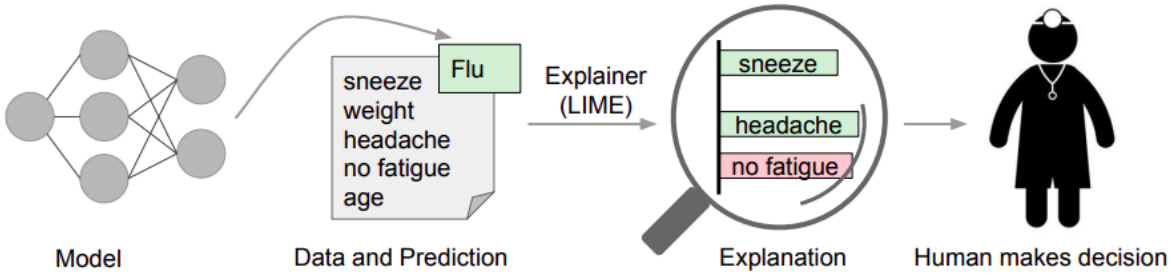
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

It appears that the LIME is useful for explaining the images. But, does it work with other types of data? The answer is yes because the M in LIME stands for model-agnostic. Model-agnostic implies that the LIME is compatible with any types of input and works with any machine learning algorithms. (this is not true in practice because the LIME library only supports text and images) For example, the LIME can be used to determine whether an email is spam or not by looking at the texts as well as determining whether a picture is a picture of husky or wolf.

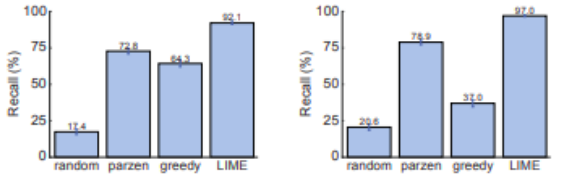
Now, it is clear that LIME gives an explanation that is very flexible with machine learning models and input types. However, we did not establish that the explanation is interpretable by humans. For example, "000111000" can be useful data for morse code, but it is not interpretable by people who do not know the morse code (basically most of the population). Fortunately, the I in LIME stands for interpretable, which means LIME gives an interpretable explanation of why the model made a prediction. So, the LIME won't give us explanations such as "01001001011010", which can be extremely useful for machines, but are useless for humans. Instead, the LIME will tell us which parameters were used and how parameters were weighted to make a prediction.



Finally, everything in LIME starts to make sense except for the L. Now, it is getting clear why L stands for local and what does local mean in LIME. As shown in the husky picture above, the LIME does not use the entire input to provide an explanation. Instead, it uses local samples from the input, then uses a locally weighted square loss function to make an explanation. This saves us a lot of time without losing too much accuracy compared to making a decision based on global data.

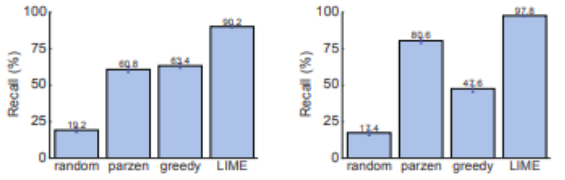
Now, we know what LIME is and how it works, so we should shift our focus to how we apply LIME to trust the model. For this, we need to simulate experiments to evaluate the utility of explanations in trust-related tasks. In particular, we address the following questions: (1) Are the explanations faithful to the model, (2) Can the explanations aid users in ascertaining trust in

predictions, and (3) Are the explanations useful for evaluating the model as a whole. (Ribeiro 6) In the paper, the authors experimented with two sentiment analysis datasets where the task is to classify product reviews as positive or negative. Then, they used LIME to randomly pick K features as an explanation (K was 10 in the paper). After that, they cross validated the result with parzen.



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

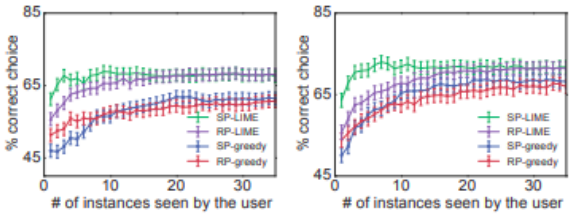


(a) Sparse LR (b) Decision Tree

Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**



(a) Books dataset (b) DVDs dataset

Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

To determine the validity of the explanation, the researchers measured how good other methods provide a valid explanation. The data set was processed with each method, then the result was compared to "answers" that were already known. Then, the researchers measured what percentage of meaningful features were included in the results. This way the researchers were able to determine whether the methods provide valid explanations or not. The results showed that LIME consistently provided over 90 percentage recall for all data sets and classifiers. (Ribeiro 6) Considering high performance of LIME, explanations appear to be faithful to the model.

Now, it is empirically proven that the LIME provides a trustful explanation to the model. However, a sound explanation does not mean that the model will give accurate predictions. So, it is natural to ask if we can trust the prediction. In the paper, the authors mixed 25 percent of wrong data to see if LIME can identify this fake data. Over 100 runs, the researchers statistically concluded that the LIME provides significantly better predictions than other methods.

Finally, it is noteworthy to check if we can trust the model as a whole (sound explanations and accurate predictions). In the research, the authors purposely added some noisy features and mixed data in order to see how well LIME performs. In this experiment, the result showed that the SP-parzen and RP-parzen performed almost as bad as the random approach. However, LIME was consistently better than other methods. With averaged over 800 trials, the authors were able to conclude that the LIME was a good method for giving the valid explanation as well as making good predictions.

Worked Cited

[1] Ribeiro, Marco Tulio, et al. "'Why Should I Trust You?": Explaining the Predictions of Any Classifier." ArXiv.org, 9 Aug. 2016, arxiv.org/abs/1602.04938.