CS 634 1J2 Data Mining, Spring 2021

Team:

Parvathy Neelakanta Sarma

Rajalakshmi Vikram

Roma Dungarwal

Tutorial on LIME method

Machine learning is at the heart of many recent advances in science and technology. It is excellent in prediction accuracy, process efficiency, and research productivity. A vital concern remains whether humans directly use machine learning classifiers as tools or are deploying models within other products. Computers usually do not explain their predictions. This becomes a barrier to the adoption of machine learning models. If the users do not trust a model or a prediction, they will not use it. Therefore the issue is how to help users to trust a model.

"Why Should I Trust You?"

The interpretability for a black-box model has a novel solution. So it is essential to understand two different (but related) types of trust:

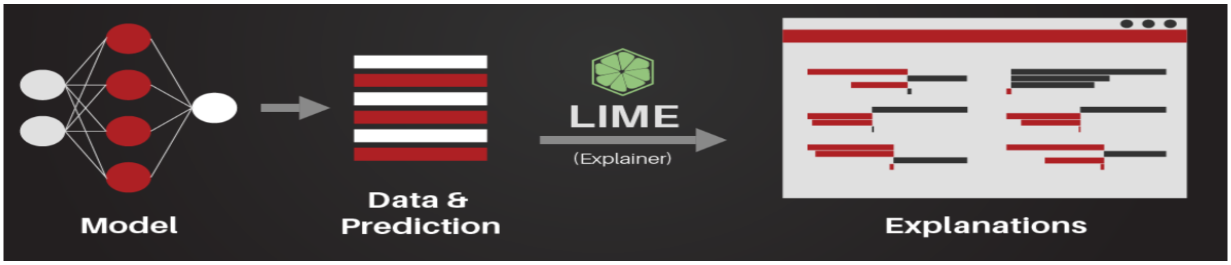
(1) trusting a prediction: a user will trust an individual prediction to act upon. No user wants to accept a model prediction on blind faith, especially if the consequences can be catastrophic.

(2) trusting a model: the user gains enough trust that the model will behave reasonably when deployed. Although in the modeling stage, accuracy metrics (such as AUC Area under the curve) are used on multiple validation datasets to mimic the real-world data, there often exist significant differences in the real-world data. Besides using the accuracy metrics, we need to test the individual prediction explanations.

There are several extraordinary solutions, including SHAP, LIME, and ELI5.

What is LIME?

Figure(a)



Local Interpretable Model-Agnostic Explanations (LIME) can explain the predictions of any classifier in "an interpretable and faithful manner. by learning an interpretable model locally around the prediction." Its approach is to gain users' trust for individual predictions and then to trust the model as a whole. It describes a prophecy that even the non-experts could compare and improve on an untrustworthy model through feature engineering.

As we insert the perturbed data into the model[Figure (a)], we receive the prediction from the data. But as a user, to understand why the prophecy(prediction), the LIME comes into place. When applied in any model, the LIME explainer gives the essential explanations and reason behind the model by using the same or similar features used by the model.

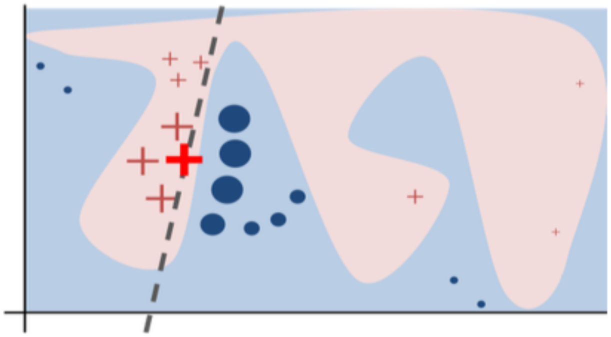
1. **For text**: It represents the *presence/absence of words*.
2. **For image**: It represents the *presence/absence of superpixels* (contiguous patch of similar pixels)

How Does LIME Work?

The blue/pink background represents the original complex model. It is not linear. The bold red cross is the individual prediction to be explained. The algorithm of LIME does the following steps:

- Generating new samples then gets their predictions using the original model, and
- Weighing these new samples by the proximity to the instance being explained (represented in Figure (b) by size).
- Then it builds a linear regression for these newly created samples, including the red cross. The dashed line is the learned explanation that is locally (but not globally) faithful.

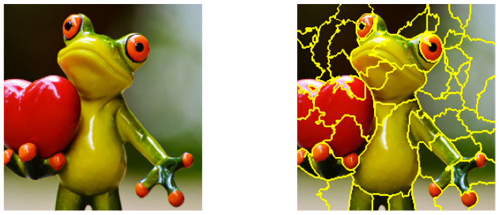
Figure (b)



Examples:

1. Lime with Images

Figure (c)



In this example, we will use LIME and see whether it can predict the image's behavior as the tree frog. We take the picture on the left and divide it into interpretable components. According to the model, we get the probability that a tree frog is in the image for each perturbed instance. We then learn a simple (linear) model on this data set, which is locally weighted. We care more about making mistakes in perturbed instances that are more similar to the original image. In the end, we present the superpixels with the highest positive weights as an explanation, graying out everything else.

Figure (d)

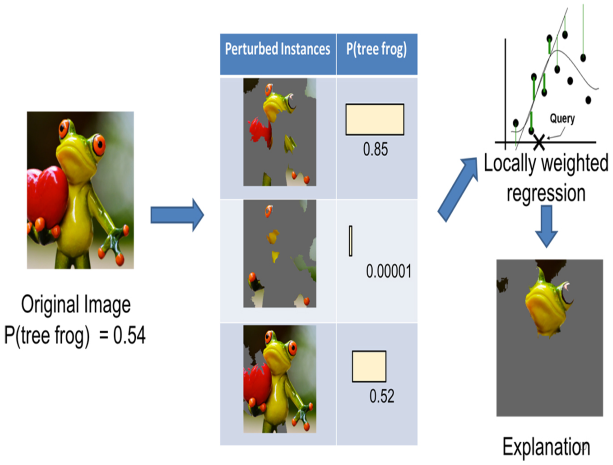
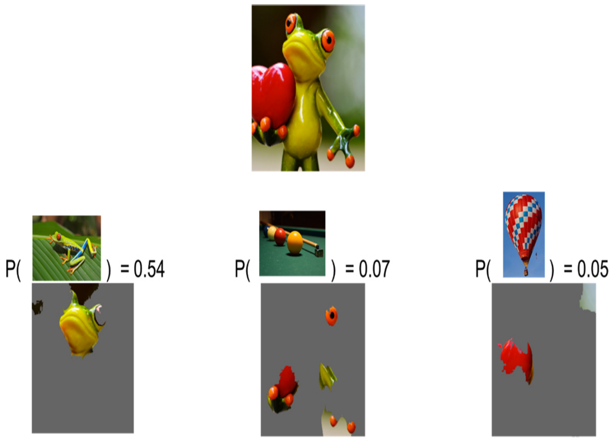


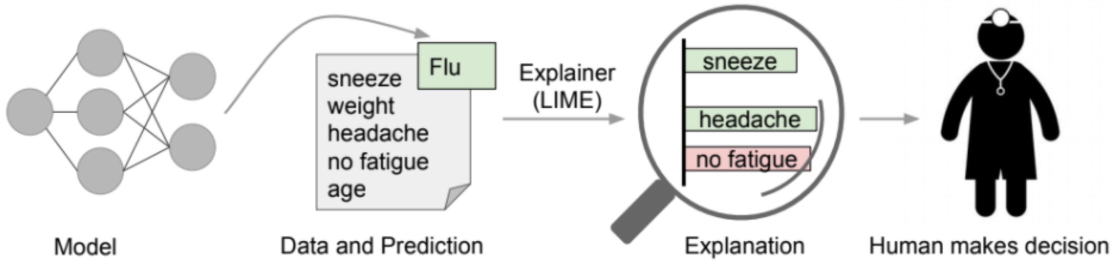
Figure (e)



The classifier predicts "tree frog" as the most likely class, followed by "pool table" and "balloon" with lower probabilities. The explanation reveals that the classifier primarily focuses on the frog's face to explain the predicted class. It also sheds light on why "pool table" has non-zero probability: the frog's hands and eyes bear a resemblance to billiard balls, especially on a green background. Similarly, the heart reaches a resemblance to a red balloon.

Lime with Text

Figure (f)



The above is an example of LIME, where a doctor is trying to find the reliability behind the model's results. The LIME explains the intuition and concept behind the model with respective features of the model used. A patient shows up with particular symptoms, e.g., sneeze, headache, no fatigue, age. The model predicts that the patient has Flu, and the LIME explains the reasons for the same, i.e., the positive symptoms such as sneeze and headache are associated with Flu [green], and no fatigue is against the signs of Flu [red]. However, the majority of the symptoms are positive features of Flu hence the diagnosis. This way, LIME interprets the concepts behind the model without any assumptions.

Advantages:

- The explainer to be model-agnostic; we can perturb the input dataset and see how the predictions change. This turns out to benefit in terms of interpretability because we can perturb the input by changing components that make sense to humans (e.g., words or parts of an image), even if the model uses much more complicated component features.
- The resulting explanations are short and easily human-understandable.
- The *fidelity measure* gives us an idea of how reliable the interpretable model is in explaining the black box predictions in the neighborhood of the data instance of interest.

Disadvantages:

- For each application, you must try *different kernel settings* and see for yourself if the explanations make sense.
- If we try to repeat the sampling process, then explanations can differ for very close data instances.
- LIME explanations can be manipulated by the data scientist to hide biases. The possibility of manipulation makes it more difficult to trust explanations generated with LIME.

Conclusion:

LIME is a great tool that helps data scientists figure out why their predictive models fail and explain individual predictions. It is model-agnostic, leverages understandable and straightforward ideas, and does not require a lot of effort to execute. As always, even when using LIME, it is still essential to correctly predict the output.

References:

1. Explain Your Model with LIME. Compare SHAP and LIME | by Dr. Dataman | Dataman in AI | Medium
2. https://christophm.github.io/interpretable-ml-book/lime.html#images-lime
3. https://homes.cs.washington.edu/~marcotcr/blog/lime/
4. https://towardsdatascience.com/lime-explaining-predictions-of-machine-learning-models-1-2-1802d56addf9
5. https://www.kdnuggets.com/2016/08/introduction-local-interpretable-model-agnostic-explanations-lime.html