Kavin Elamurugan, Nick Abadiotakis, Cole Kandel

CS301

Professor Pantelis

4/22/2021

The LIME Method

Why it is important for people to be able to trust a ML model

Machine Learning is one of the fastest growing and most important recent developments in technology. However, while the main point of machine learning is to have very little human interaction with the machine, people are still an integral part of the system. Humans generally use machine learning classifiers as tools to help the user make a decision, or implement a model in some other items. But in order for a person to use any machine learning model, the person must first trust that model. What if a model that predicts whether a picture of a wolf is basing its prediction on the color white being in the picture? While it may accurately find pictures of wolves, it would not be able to tell if a picture was not a wolf if it had the color white. Ultimately, if the users do not trust a model or a prediction, they will not use it. The user must both be able to trust an individual prediction made by the model, as well as the method by which it makes that prediction. However, to know how a model makes its prediction is difficult as generally machine learning algorithms are abstract and treated as a mystery black box. This is where the LIME method can be of use.

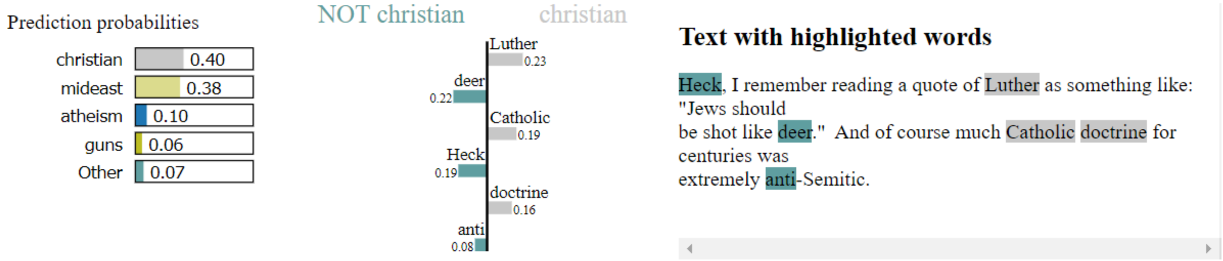
What is LIME

LIME stands for Local Interpretable Model-Agnostic Explantation. LIME is an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model. It takes an individual prediction, and tries to fit a generally linear line to it, thus “Local”, and returns an “Interpretable" model. It does this since typically it is very hard for humans to understand the decision making of a neural network which might have a million nodes. LIME works on any classifier type, Naive Bayes, Random Forest, Neural Network, and as such it is “Model-Agnostic” . Anyone using LIME can be able to pick any classifier and see which generalizes better in the real world. Further, they are able to greatly improve an untrustworthy classifier by doing feature engineering using LIME.

How LIME works.

In a Machine Learning Model, there are generally thousands of features that are accounted for. However, some features that are globally important may not be important in the local context, and vice versa. For example, when classifying a text, some keywords, like “orthodox” or “Jesus” might be important for classifying a text about christianity, however in the context of classifying a Islamic text, those key words would not be so helpful. Thus by using keywords that are used in the individual test data makes the explanation more local, since a global explanation would be difficult to understand. This ties into the next point that the explanation must also be interpretable and model-agnostic. As mentioned before, a globally faithful interpretation would be difficult to interpret, thus by simplifying the model, it becomes easier to find reasons as to why a result was given, regardless of the actual features used by the model. “For example, a possible interpretable representation for text classification is a binary vector indicating the presence or absence of a word, even though the classifier may use more complex (and incomprehensible) features such as word embeddings.”(Ribiero) In image 1, the text shown contains a few highlighted works which are key in determining what type of text this is. The words Luther, Catholic, and doctrine have a high correspondence with christian text, whereas the words deer, Heck and anti bear an non christian connotation. Notice that the word Jews and Semetic are used in this text, but as the classification was christian or not christian, the words are not considered locally. While they might be considered by the model globally, a faithful explanation of their effect on the prediction made by the model would be too complex.

(1)

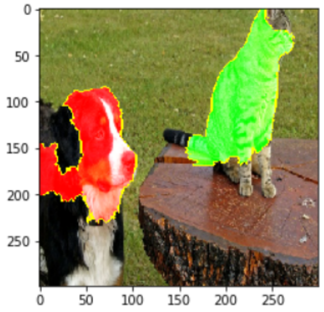


Text with highlighted words

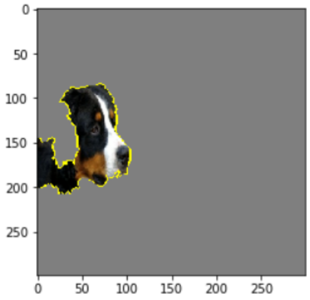
Heck. I remember reading a quote of Luther as something like: "Jews should
be shot like deer." And of course much Catholic doctrine for centuries was
extremely anti-Semitic.

“Likewise for image classification, an interpretable representation may be a binary vector indicating the “presence” or “absence” of a contiguous patch of similar pixels (a super-pixel), while the classifier may represent the image as a tensor with three color channels per pixel.”(Ribiero) Similarly to the text classifications, the classification of whether the image is that of a cat or dog is being decided here. The portion of image 2 that is red and can be seen in image 3 is being considered by the model locally. The same happens with the green portion that is also seen in image 4. Both of these pixel groupings have two different connotations. The red would make the model believe the picture is that of a dog, and the green would make the model believe it is a picture of a cat. However, once again it is important to note that the remainder of the image, while not considered locally, is considered globally by the model. But that image data is less significant to the outcome of the model, and in addition trying to create an explanation that faithfully interprets the gray space would be too complicated for humans to understand.

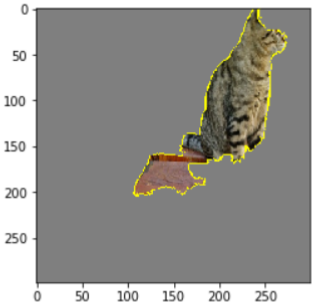
(2)



(3)



(4)



How to use LIME

Lime can be imported into any ipy notebook with the !pip install lime command. Otherwise, it can be found on the lime github by a user of the name marcotcr. The github had many links and tutorials as to the specifics of Lime, however the basics are as follows. After importing lime into a project, the next step is generally to create an explainer object based on which type of classifier you are using, text or image. Then after passing the explainer the test data and the probability which was obtained using the classifier, one would use some of the various methods of the explainer to show the classifiers decision making. For example, the explainer has an as_list method, which shows a list of weighted features which were important in the prediction. Another option is the show_in_notebook method, which outputs the same data as a graph.

Conclusion

Lime is a powerful tool in order to better understand a ML model. And since it is unreasonable to use a model if we can't trust it, the best way to circumvent this issue is to slowly try and understand it. Through the use of LIME, even incredibly complex models can be understood, and over time, we might be able to fully trust a ML model.