***LIME Methodology***

***Section: CS 301 - 102***

***Group Members:*** Shanil Rahul Pathak

Akash D Patel

Mann M. Patel

Trusting prediction models and their behaviors is a task with no sure shots of landing a viable explanation in some instances. When the consequences of predictions are beyond just individual feelings or preferences, the results can be catastrophic. That is why, apart from trusting individual predictions, there is also a need to evaluate a model before deploying it for definitive usage. Explaining a prediction implies presenting textual or visual artifacts that provide a qualitative understanding of the relationship between the instance’s components and the model’s pre­­diction.

Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm that is capable of explaining predictions of any regressor or classifier by approximating viable outcomes locally with an interpretable model. An interpretable model must provide a qualitative understanding between the input variables and the response. Models must also take into account the user’s limitations and thus a linear model, a gradient vector model or an additive model may or may not be interpretable.

Interpretability essentially means that the explanations have to made easy for the users to understand, which is not absolute truth for the feature space used by different models since it may use too many input variables. Hence, LIME’s explanations use a data representation that is different from the original feature space.

An essential criterion for interpretable models is that they must provide local fidelity. Local fidelity however, does not imply Global fidelity since features that are important in the local context may not be important in the global context and vice-versa. Another essential criterion is that an explainer must be able to explain any model and thus be model-agnostic. As an eventuality of any accepted model, a global perspective is very important to ascertain trust in that model. LIME produces explanations by approximating the black-box model using an interpretable model in the neighborhood of the instance we want to explain.

LIME is focusing on identifying an interpretable model over the interpretable representation that is locally faithful to the classifier. Alongside, SP-LIME is a method that selects a set of representative instances with explanations to address the “trusting the model” problem, via submodular optimization.

Speaking of black-box models, that is the metaphor that roots back to the early days of cybernetics and behaviorism. It typically means to refer to a system for which we can only observe the inputs and outputs, but not the internal workings. Secondly, the reason why the metaphor of black-boxes is relevant is with respect to the systems we are trying to learn, such as human vision. In some ways, human behavior is unusually transparent, in that we can actually ask people why they did something, and obtain explanations.

We may take into consideration a model denoted by f: Rd 🡪 R. In classification, f(x) is the probability that x belongs to a certain class. We further use πx(z) as a proximity measure between an instance z to x, so as to define locality around x. Finally, let L(f, g, πx) be a measure of how unfaithful g is in approximating f in the locality defined by πx. To we make sure the persistence of both local fidelity and interpretability, we must minimize the unfaithfulness, L(f, g, πx) while having Ω(g) be low enough to be interpretable by an observer. LIME explanation can be portrayed as follows: ∑(x) = argming∈G L(f, g, πx) + Ω(g). LIME, in other words, relies on the assumption that every complex model is linear on a local scale. It presents an explanation that is locally faithful, where the locality is captures by πx.

There is also a trade-off between the aspects of fidelity and interpretability. We define an explanation as a model g ∈ G, where G is a class is a class of potentially interpretable models and g acts over the availability of the interpretable components. Ω(g), πx(z), L(f, g, πx) play out in such a manner that we intend on minimizing how unfaithful g is in approximating f.

In application, the method of approach that LIME relies towards is that primarily it generates N samples of the interpretable version of an instance to explain y. The we recover those observations in the original feature space using a mapping function. The black box model then predicts the outcome of every perturbed observation. Furthermore the weights of every perturbed observation are calculated. Using all of that information as training data, better predictions can be made. Since this method produces the explanation for an individual prediction, the complexity does not depend on the size of the training set and instead on time to compute predictions and on the number of samples.

The LIME method has been wide adopted within the text and image analysis, part thanks to the explainable information illustration. In this case, the reasons square measure delivered within the sort of fragments of Associate in Nursing image/text, and users will simply realize the justification of such explanations. The underlying intuition for the tactic is simple to understand: a less complicated model is employed to approximate an additionally complicated one. By employing a less complicated model, with a smaller variety of explainable instructive variables, predictions square measure easier to clarify. The LIME methodology may be applied to complicated, high-dimensional models.

Another vital purpose is that, as a result of the glass-box model is chosen to approximate the black-box model, and not the information themselves, the tactic doesn't management the standard of the native work of the glass-box model to the information. Thus, the latter model could also be dishonorable. Finally, in high-dimensional information, information points square measure distributed. shaping a “local neighborhood” of the instance of interest might not be easy.