CS-634 Data Mining
Spring 2021
Project 2: Why should I trust you?
Group Members:
Yorleydis Oliveros, Huong Ly Ngo, Rob Mullaney
LIME Tutorial

With Machine Learning at the center of recent advances in Science and Technology, a contentious relationship between any given Machine Learning model and its respective user will impede the wide-scale adoption (and benefits) of Machine Learning technology. Therefore, whether humans are using Machine Learning classifiers as tools or are deploying Machine Learning models within other products, the LIME methodology is essential to not only build trust between end-user and model, but to provide transparency by providing a qualitative understanding of the relationship between an instance/case and the model's prediction. Ultimately, we argue that by providing explanations of features for individual predictions through the LIME and SP-LIME methods, then measuring the impact of these explanations on user trust and associated tasks, should assist in the adoption of Machine Learning models.

First off, when we provide "explanations for individual predictions", it is important to know the background of the end-user. The reason being that Humans usually have some prior knowledge about the domain of any given Machine Learning model. Therefore, we want to empower an individual's prior domain knowledge to either accept or reject a model's prediction. However, it is important to provide clear and coherent explanations that are interpretable by the end-user, given the user's limitations. Explaining predictions in an interpretable manner is an essential step in building the trust to properly use a Machine Learning model.

When providing an explanation for a prediction, there are a few essential criteria that must be met. First, the explanation of features must be interpretable. As mentioned before, this trait must take into account user limitations, but must provide some type of understanding between the input variables and prediction. Secondly, our explanation must be Locally Faithful. Although it is nearly impossible to provide an explanation that's faithful without providing a complete description of the model itself, the explanation must highlight the primary features of the model within the vicinity of the prediction. Additionally, an explainer should be model-agnostic and be applicable to any model. Finally, our explanation should provide a Global Perspective in order to continue building trust with the user. Accuracy measurements are not always ideal when evaluating a model. Therefore, we want to explain the model while utilizing a select few explanations from individual predictions that are representative of the model.

The LIME (Local Interpretable model-agnostic explanation) technique - explains any classifiers' predictions by learning an interpretable model locally around the prediction.

A benefit of using LIME is identifying an ***interpretable model*** over an ***interpretable representation*** that is locally faithful to the classifier.[1] An interpretable model is defined as a qualitative understanding between the input variables and the response. When an interpretable model is locally faithful, the criterion must correspond to how the model behaves in the area in which it's being predicted. The explanation produced by LIME is obtained through the below formula.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

It's important to note that the formula can be used with different explanation families (G), fidelity functions (L), and complexity measures (Ω). Likewise, the constant g ($g \in G$) is a member of the set G, with G being a potentially interpretable model or explanation for families, such as linear models, decision trees, or a falling rules list. However, the LIME method only provides local explanations of our model, and if we were searching for a global explanation of the model, then the SP-LIME method would need to be utilized, which will be covered following the explanation and tutorial of the LIME method.

When using LIME, we want to minimize the locality-aware loss (L) while maintaining model-agnosticity, and therefore, we cannot make any assumptions about *f*. We interpret the locality-aware loss function (L) as a measure of how unfaithful the constant g is when approximating *f* within the locality as defined by variable π. Variable π(z) is used as a proximity measure within an instance of z to x, in an attempt to define a prediction's locality around x. Thus, as interpretable inputs vary, we estimate the locality-aware loss (L) by drawing samples weighted by variable π(z). All the while, keeping the complexity measures (Ω) low enough to be interpretable by humans.

As we minimize the locality-aware loss (L), we sample instances around x' by drawing nonzero elements of x' at random. Given sample $z \in \{0, 1\}$, which contains a share of nonzero elements of x',

we obtain the model's original sample ($z \in R$) and are able to derive *f*(z), which is then used as a label for the explanation.

As previously mentioned, LIME provides an explanation for a single prediction, but if we were to search for an explanation of the model as a whole, then we would need to utilize the SP-LIME method instead. In this approach, a global explanation of the model is provided through a set of individual instances.

When using the SP-LIME method to explain a model as a whole, the instances used must be carefully selected since users are unable to examine an infinitely large number of explanations. Therefore, given a set of X instances, we start with a "Pick Step" of selecting B instances for the user to select, with B representing the budgeted number of explanations a user is willing to analyze. The "Pick Step" takes into account the explanations associated with each prediction, and should cover a diverse set of explanations that are representative of how the model behaves globally.

Given the explanations for a set of X instances (|X|=n), we construct an n * d' explanation matrix (W), representing the number of the local importance of the interpretable features for each instance. Similarly, for each column (j) in W, we represent the global importance of the given column (j) by an importance score (I) within the explanation matrix. When selecting an importance score (I), we want to see features that explain many different instances to hold higher importance scores (I), but still avoid selecting instances with similar explanations and features.

To avoid redundant coverage of features, we define a coverage function (c) which calculates the total importance of features appearing in at least one instance within a new set V, given the explanation matrix (W) and importance score (I). The coverage function is defined below, along with the algebraic goal of the "Pick Step" which is to find the set of W and I, which maximizes the coverage function.

Leveraging submodularity, an algorithm that iteratively adds the instance with the highest marginal coverage gain to the solution, offers a constant-factor approximation guarantee of 1-1/e to the optimum, optimizing our algorithm.

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$$

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}}\, c(V, \mathcal{W}, I)$$

Finally, we conduct simulated user experiments in an attempt to evaluate the effect of explanations on trust between the end-user and model. Specifically, we seek to answer whether the explanations are faithful to the model, whether the explanations aid users in cementing trust of the model's predictions, and whether the explanations are helpful when evaluating the model as a whole.

Within *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* by Ribiero, et. al., sentiment analysis was performed on two separate datasets (books and DVDs) to classify product reviews as either positive or negative by using Decision Trees, Logistic Regression, Nearest Neighbors, or a Random Forest, while dividing each dataset with 75% of instances going into a train set and the remaining 25% of instances going into a test set. To explain individual predictions, the LIME approach is compared against the Parzen method, a "Greedy procedure" (iteratively removing heavily weighted features influencing predicted class until the prediction changes) and against a random procedure that randomly picks K (K = 10) features as an explanation using a random selection (RP) or a submodular selection (SP). Next, we measure the faithfulness of the explanations generated by each procedure (Parzen, Greedy, Random, and LIME) by measuring the fraction of features that are found within the explanations generated by each procedure.

The next step after measuring the Recall of each explanation, is to simulate user trust of individual predictions. Ribiero, et. al. simulate this step through development of an oracle "trustworthiness" by labeling test set predictions as "untrustworthy" if the prediction changes from removing insignificant or untrustworthy features for the given instance. Ultimately, this simulated step mimics the users' trust of each explanation provided by comparing the results of the test set to the trustworthiness oracle.

The aim of the final simulated user experiment is to evaluate whether the explanations generated are appropriate for model selection. This is done by simulating user selection between two different models with similar accuracy on validation data, while adding 10 artificially “noisy” features into our datasets to recreate situations where the models hold informative value, but contain spurious correlations as well. The simulated user experiment is attempting to evaluate if a user can recognize the optimal classifier based on the explanations of B instances from the validation set. In the paper, we see that the accuracy of the LIME method consistently outperforms the “greedy” method when tasked with picking the correct classifier while B varies. Furthermore, if Submodular Pick of LIME and the “greedy” method outperform their Random Pick peer.

After building the studied models and simulating user experiments, Ribiero et. al. begin to evaluate the usefulness of feature explanations provided through the LIME and SP-LIME methods with Human subjects. Within the paper, the evaluation with human subjects seeks to answer (1) if users can choose which of two classifiers generalizes better, (2) if users can perform feature engineering to improve the model, based on the explanations generated, and (3) if users are able to identify and describe classifier irregularities by looking at explanations. These questions are answered by presenting explanations to the users to help them decide which classifier is more accurate when generalizing each instance of our religion dataset between “Christiantiy” and “Atheism”.

Using Support Vector Machines with a Radial Basis Function kernel, trained on 20 datasets from news organizations and hyperparameters optimized through cross-validation, human subjects are asked to identify the better algorithm based on the explanations provided from the “greedy”, iterative approach against the LIME approach. It’s worth noting that the number of words in each explanation is limited to K, and the total number of documents a person inspects (B) is set to 6. The explanations are then produced by the greedy or LIME approaches with each instance selected by Random Pick (RP) or Submodular Pick (SP).

The results in the paper under figure 9 (shown below) show that all of the methods are helpful at identifying the better classifier. However, the Submodular Pick (SP) outperforms when compared to the Random Pick (RP), but LIME continues to outperform the “greedy” method both in cases.

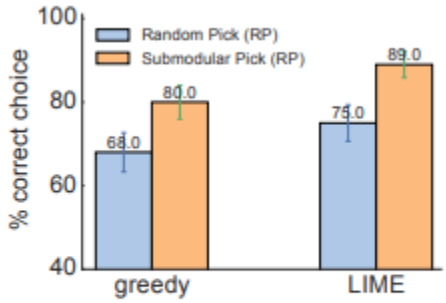

Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

Next, Ribiero et. al. then sought to answer whether users could perform feature engineering by asking users to identify which words should be excluded from explanations for subsequent retraining of the worst classifiers from their previous step. The explanations and instances shown to users tasked with Feature Engineering are provided through the SP-LIME and RP-LIME methods. The below graph shows the average accuracy on the religion dataset after each round of Feature Engineering (removing unnecessary words).

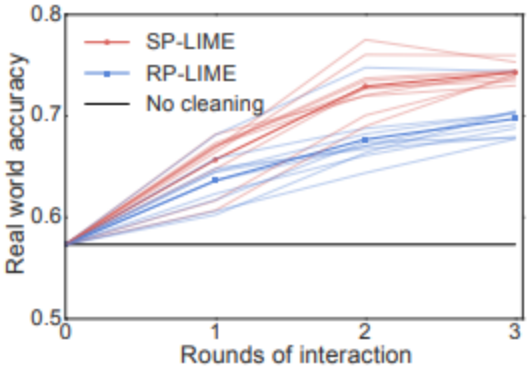

Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

In the final step of the experiment, Ribiero et. al. attempt to measure the impact of providing feature explanations to users by presenting human subjects with a classifier that differentiates images of Huskies from images of Wolves. The classifier used in this experiment was intentionally trained to be a "bad" classifier as a means to evaluate whether the subjects are able to detect its poor performance.

What is becoming more or less a Psychology experiment, Ribiero et. al. present each human subject with a balanced set of 10 images of Huskies and Wolves. Within the set, one incorrectly classified Husky is included (shown below), with the other 9 images being correctly classified. Users are then asked whether they trust the algorithm and why. Additionally, users are asked how they think the algorithm is able to distinguish between photos of Wolves vs. Huskies. After the users respond to the questions, the users are presented with the same images but with explanations included, and are then tasked to answer the same questions. After collecting the initial round of responses, 3 independent evaluators are tasked with reading user responses and determining if the test user mentions anything about snow or background as a feature behind the model, with the results shown in table 2.

(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.

Looking at the results in table 2, the presence of explanations provided through the LIME methodology clearly improved the decision-making abilities of the end user. Before providing explanations, more than a third trusted the "bad" classifier, with less than half of the users suggesting that an image's background or the presence of snow as a feature of the classifier. However, after examining the explanations, nearly all users correctly identified the background or snow as an important factor of the model, with nearly all users no longer trusting the "bad" classifier.

In conclusion, by providing an explanation of model features through the LIME methodology, Ribiero et. al. exhibit its substantial impact on user decision making. Although the results showcased a large drop in the trust of the classifier, users still managed to improve their performance of the associated task, which was whether they could correctly identify the snow or background as a potential feature.

Given the overall results of the test, it is easy for one to argue that the LIME methodology is counterproductive to the adaption of Machine Learning and harms the trust between end-user and model. However, we believe that the presence of explanations through the LIME will increase the rate of adoption of Machine Learning within the real world because the Ribiero et. al.'s experiment exhibits the benefits (an improvement in identification of a good or bad model) from providing explanations for individual cases, which thereby provides clarity into "black-box" models. Although table 2 shows a drop in trust between the users and the "bad" model, the paper is a great example of how to prevent implementation of "bad" Machine Learning models or classifiers that could have a potentially adverse impact on the communities and end-users who stand to benefit from improved efficiencies or automation of menial tasks within their local businesses or local government agencies.

CONCLUSION AND FURTHER STUDIES

To summarize, LIME is a method of explanation of a prediction or description of any machine learning model. It is crucial to data analysts to make their models explainable and transparent. Sometimes, we just get driven away by the results and forget about how we come up to those outputs. The questions are how we can understand a model and how can we be confident to trust the prediction of that model. LIME is highly recommended for model interpretability because of being local-faithful and model-independent or model-agnostic, just like the name of this technique. But we must stretch out that interpretable explanations and interpretable representation are different, and this leads to the fidelity-interpretability trade -off when we apply LIME to a model. The data representation should be understandable to human beings regardless of features. The example that we can mention here is : text classification task as a binary vector as output (present word/missing word), regardless of the number of input features. But LIME is local faithful, which means that the local explanations must fit the predictions that we obtain from the chosen model. The more features we feed to the local explanation, the more accurate are the outputs and will match closer to the original prediction but then the explanation will be more complex as there are more features which contribute to this explanation.

But LIME, as its name already described, explains models only locally. In order to understand a model globally, submodular pick or SP-LIME will be used. The method behind SP-LIME is to pick only the most important features from the dataset. The global importance of each feature is determined by the number of instances that each feature can explain. The more point values are explained, the more important is that feature. Unlike LIME, SP-LIME does not have redundant predictions and will provide a global view of the model to the users.

Finally, LIME/SP-LIME is a method applied to explain any machine learning system. This new domain opens the door to the explanation of variety studies such as speech, text, medical domains, recommendation systems….

Bibliography

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 9). *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* [PDF]. Seattle: University of Washington.