Akhil Sreedhara, Jaymee Hyppolite, Mihir Rana

Professor Pantelis Monogioudis

Introduction to Data Science - CS301

25 April 2021

Lime Tutorial

In the realm of artificial intelligence (AI) and machine learning (ML), there are quite a few quandaries and conundrums that must be accounted for. One of the biggest being the reasoning behind predictions within machine learning models, and understanding the reasoning is substantially important in establishing trust for AI., "which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one." (Riberio et al, 1). One technique to help understand model reasoning is the LIME method. LIME stands for Local Interpretable Model-Agnostic Explanations. When we analyze the reasoning behind predictions, there are certain factors that we wish to be addressed in such explanations — and each letter in the LIME acronym directly reflects something that is desired in explanations. Local refers to local fidelity — this means that the explanation is geared to reflect the behavior around the ML classifier around the instance being predicted. Interpretable refers to how the model explicates the data to the point that a human can understand it. Lastly, LIME is able to explain any model without having to analyze the data, making it model-agnostic. Another important aspect of what is provided by the LIME method — that isn't directly intuitive by the name — is the global perspective. The explainer is able to explain a representative set to the user so that the user has a global perspective of the model.

All of the processes within the LIME method are done with the intention of building trust in a model. In order to build trust, multiple cross-validations are run, and hold-out set validations are performed. These simulations together provide an aggregated view of model performance over unknown data. "Unfortunately, the important role of humans is an oft-overlooked aspect in the field. Whether humans are directly using machine learning classifiers as tools or are deploying models within other products, a vital concern remains: if the users do not trust a model or a prediction, they will not use it" (Riberio et al., 1). In order to trust the models, we need to understand why some predictions are correct, and others are wrong. Also, we need to have the ability to trace our model's decision path. In machine learning, we can define trust in two ways. The first is trusting a prediction, which means whether or not a user trusts a single prediction well enough to take action based on it. "We argue that explaining predictions is an important aspect in getting humans to trust and use machine learning effectively if the explanations are faithful and intelligible." (Riberio et al, 2). The second definition has to do with trust within a model, which means whether or not a user trusts a model to behave reasonably when deployed. The two definitions are different but related — and both are directly impacted by a user's ability to understand a model's behavior.

In order to address these definitions of trust and user understanding, we have to analyze what makes LIME a good model explainer. The two biggest characteristics of an explainer are interpretable data representation and fidelity-interpretability trade-offs. — in addition to sampling for local exploration and sparse linear explanation — Beginning with the former, we can see that LIME utilizes a representation of data that is easily understood by people/users, irrespective of the features that are used by the model. This kind of interpretable representation can — and will — vary depending on the type of data that is supplied. For tabular data, the

interpretable representation would take the form of a weighted combination of columns. For text, it would represent the absence or presence of words. And for images, it would represent the presence or absence of pixels. The second characteristic again is fidelity-interpretability trade-offs. To ensure that both the interpretability and local fidelity locality-aware loss is minimized while keeping a measure of model complexity low enough to be understood by people.

The usage of LIME is very simple and follows a formula that goes by $\xi(x) = \text{argmin}\ g \in G\ L(f, g, \pi x) + \Omega(g)$ (Riberio et al, 2). In this formula, G represents the explanation, and $g \in G$ represents a model that can be presented to the user with visual elements. $L(f, g, \pi x)$ represents the local fidelity, better described as the measure of how bad g is at approximating f, how close g's prediction is to f, based on the weights of πx. $\Omega(g)$ measured the complexity of g or how un-humanly interpretable the result of g is. All of this together produces the explanation that lime outputs. As a more clear non-formulaic explanation, it is possible to represent each part of the equation in a specific problem set. For example, in a problem measuring the likelihood of a message being spam or not, it is possible to break down the equation in a way that can suit the problem. For $L(f, g, \pi x)$, one can use a loss function like the square loss to approximate the relative closeness of the interpretable model g, which can be some linear model that will be weighted by πx to determine the significance of each point. $\Omega(g)$ will account for the interpretability of the model g, and $\text{argmin}\ g \in G$ will present a set of interpretable models, G, so that the most appropriate model would be applied to interpret the model to be explained, f.

LIME can be further generalized to the whole by using a modified version of the LIME formula above with submodular picking. The Submodular Pick LIME or SP-LIME is essentially a method to show exclusive information to the user that is produced by the model, f, that has the

likelihood of having a larger impact by being analyzed. This step is added into the formula by accounting for the added value that new information has to the user that is analyzing the current model. LIME does this by creating an explanation matrix, which is a matrix that contains features and the number of instances. A feature that is encountered many times will have higher importance, so the picking process would be very likely to pick an instance that would contain this feature. To avoid redundancy, the pick feature would then pick another instance that would not contain features in the first selected instance which allows broader coverage of data in the explanation.

LIME's ability to explain a single prediction is valuable and provides understanding into how reliable a classifier is to a user, but it is not adequate enough to assess and evaluate trust in a model as a whole as SP-LIME attempts to. LIME aims to describe a model's prediction to features that are perceivable by people. In order to achieve this, users need to run the LIME explanation model on a representative and diverse set of instances — yielding a nonredundant explanation set that provides a global representation of the model. "We propose to give a global understanding of the model by explaining a set of individual instances. This approach is still model agnostic and is complementary to computing summary statistics such as held-out accuracy. Even though explanations of multiple instances can be insightful, these instances need to be selected judiciously, since users may not have the time to examine a large number of explanations." (Riberio et al, 5). So after running the explainer on all instances and predictions, computing the global importance of individual components, and maximizing the coverage function — what is returned is the representative nonredundant explanation set that has all of the desirable properties of a good model explainer.

Bibliography

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. doi:10.18653/v1/n16-3020