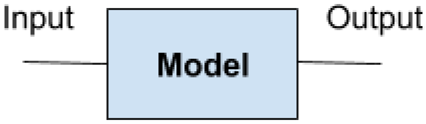
Model Explaination with LIME

Tejas Gaikwad
MT19AI021
Dependable AI

LIME stands for Local Interpretable Model-agnostic Explanation. It is basically the study of change of the model when input is changed.



It helps us to explain why a machine learning model has taken a certain decision. Well now, the question comes why do we even want to know the explainability behind a certain decision made by the model?. The simple answer for this can be, without a good understanding of the methods we are very likely to hold our decisions on some false basis, thus there is a necessity of interpretability. LIME can be applied to any learning model. LIME approaches to understand the model by perturbing the input of data samples and understanding how the predictions change. It modifies a single data sample by tweaking the feature values and observes the resulting impact on the output. This is similar to what humans are expecting while observing the output of a model. The interpretable models are trained on small perturbations of the original instance and provide a local approximation.

Examples of interpretable representations are e.g. a BoW vector for NLP or an image for computer vision. The output of LIME is a list of the probability of each element like words in case of NLP example, reflecting the contribution of each feature to the prediction of a data sample. This provides local interpretability, and it also allows to determine which feature changes will have the most impact on the prediction.

The problems associated with the LIME is as follows,

1. Can take some time to train the model
2. The model may not be locally Linear
3. The interpretable feature may not be a useful predictor

So, LIME is a great tool to explain what machine learning classifiers (or models) are doing. It leverages simple and understandable ideas and does not require a lot of effort to run.