Script for JSM 2019 Speed Presentation

Visual Diagnostics of a Model Explainer: Tools for the Assessment of LIME Explanations from Random Forests

Slide 1

My research is focused on explaining predictions made by random forest models. Recently, I have been working with the model explainer method of LIME to do this. This has led me to develop some diagnostic tools to assess LIME explanations.

Slide 2

LIME was developed in 2017 as a method to provide explanations for predictions made by black-box models. It stands for Local Interpretable Model-Agnostic Explanations. The idea is that a simple interpretable model can be used as an "explainer model" to explain the relationship between the black-box model predictions and the model features in a local region around a prediction of interest. This model can then be interpreted and used to determine the key features driving the prediction. These figures show a linear model weighted by proximity to the prediction of interest being used as the simple model to explain the complex model for two features. The image on the left shows that the simple model has captured the local trend around the point of interest, and the image on the right shows that simple model has identified that there is no strong trend around the point of interest.

Slide 3

While LIME can be used to determine if a model is producing trustworthy predictions, I am interested in the question of: How do we know if the LIME explanation is trustworthy? Thus, I would like to assess the explanation produced by LIME. For example, I want to consider:

- Is the simple model is a good approximation of the complex model?
- Is the explanation local as intended by LIME?
- Are the explanations consistent across different implementation methods available with LIME (such as the form of explainer model and distance metric used)?

These figures were created using the same data as the previous slides, but a different distance metric was used to fit the weighted linear regression models. While the simple model on the previous slide approximated the complex model well in the local region around the prediction of interest, this model does not, and it would provide untrustworthy results. I have been working to develop diagnostic tools to assess whether the LIME explanations are trustworthy.

Slide 4

The example data I am working with is a forensics bullet matching dataset. When a bullet is fired from a gun, grooves are created on the bullet. These markings are used to create signatures such as the ones shown in the figure on the right that can be used to identify if two bullets have been fired from the same gun if the signatures match. Features can be extracted from the signatures and input to a random forest model to predict whether two bullets were fired from the same gun.

Slide 5

Here is an example of one of the visualizations I have created to asses the LIME explanations. I have applied LIME to all predictions made by the random forest model fit to a bullet matching test dataset. This figure shows the top feature selected by LIME represented by the colors for each prediction in the data. These are shown for different implementation methods of LIME and separated by matches and non-matches. Some of the methods show clear vertical stripes that suggest that LIME is providing different explanations across methods for a case but the same explanation for all cases within a method. Some of the other methods show horizontal lines, which suggests consistency across methods and local explanations. These results indicate that it is important to diagnose LIME since the explanations may be dependent on the implementation method used and the explanations may not be “local” as they are suppose to be. Additionally, this leaves me asking the question of: In practice, how should we choose the implementation method to use? Thank you.