

An Assessment of LIME Explanations from a Random Forest Models

Katherine Goode and Heike Hofmann

February 11, 2019

1 Introduction

2 Data

3 Methods

3.1 LIME Algorithm

Notation

- $x \in \mathbb{R}^d$: original representation of an instance being explained
 - e.g. feature vector containing word embeddings
- $x' \in \mathbb{R}^{d'}$: vector for the interpretable representation of the instance being explained
 - e.g. bag of words
- G : class of potentially interpretable models
 - e.g. linear models, decision trees, rule lists
- g : explanation model where $g : \mathbb{R}^{d'} \rightarrow \mathbb{R}$ and $g \in G$
- $\Omega(g)$: measure of complexity of g
 - e.g. depth of a tree, number of non-zeros in a linear model
- f : model that is being explained where $f : \mathbb{R}^d \rightarrow \mathbb{R}$
 - note: in classification $f(x)$ is the probability that x belongs to a certain class
- $\Pi_x(z)$: proximity measure between an instance z to x which defines a locality around x
- $\mathcal{L}(f, g, \Pi_x)$: a measure of how unfaithful g is in approximating f in the locality defined by Π_x

Step by Step Procedure

Step 1:

3.2 Application of LIME to Bullet Matching Data

3.2.1 lime R Package Perturbation Creation Methods

The LIME R package allows for the following four methods to sample the perturbations based on the distributions of the features from the training data.

- Equally Spaced Bins
- Quantile Bins

- Normal Approximation
- Kernel Density Approximation

The methods of equally spaced bins and quantile bins also allow the user to specify the number of bins. As of now, there are no recommendations or procedures provided for how to determine which method to use. By default, LIME uses four quantile bins. It was of interest to see how the explanations from LIME varied across the four sampling methods when applied to the bullet matching data. The LIME algorithm was applied to each prediction from the test data obtained from the ‘rtrees’ random forest model for each of the sampling methods. For the bin based sampling methods, the algorithm was applied for 2 to 6 bins. That is, the LIME algorithm was applied a total of 10 times to each row in the test data with a different sampling method used each time.

3.2.2 Proposed Bin Creation Methods

4 Results

4.1 LIME Package Explanations Dependent on Sampling Method

In order to assess the LIME explanations created using different sampling methods, it was of interest to compare the top three features chosen as the important predictors by lime within a case from the test data across the different sampling methods. Figure . . . is a heatmap showing the top feature chosen by lime for each of the cases in the test data and different bin based sampling methods. The rows represent the cases in the test data, and the columns represent the sampling methods. There are twenty methods included in the plot. These include the equally spaced bins and quantile bins from the lime package and the random forest score tree based bins and the same source tree based bins proposed in this paper. The colors represent the top feature chosen by lime. The rows are faceted by the test set and whether or not the observation is a match or not.

5 Discussion