CS 634 1J2 Data Mining, Spring 2021

Team:
Parvathy Neelakanta Sarma
Rajalakshmi Vikram
Roma Dungarwal

Summary of the findings with LIME

LIME is an interpretability surrogate model which can be used on any black-box model (model-agnostic) and provides interpretability for a single observation prediction (local).
LIME works in the following manner

1. First, you choose the single prediction which you would like to be explained.
2. LIME creates permutations of your data at this instance and collects the black-box model results
3. It then gives weights to the new samples based on how closely they match the data of the original prediction.
4. A new, less complex, interpretable model is trained on the data variations created using the weights attached to each variation.
5. Finally, the prediction can be explained by this local interpretable.

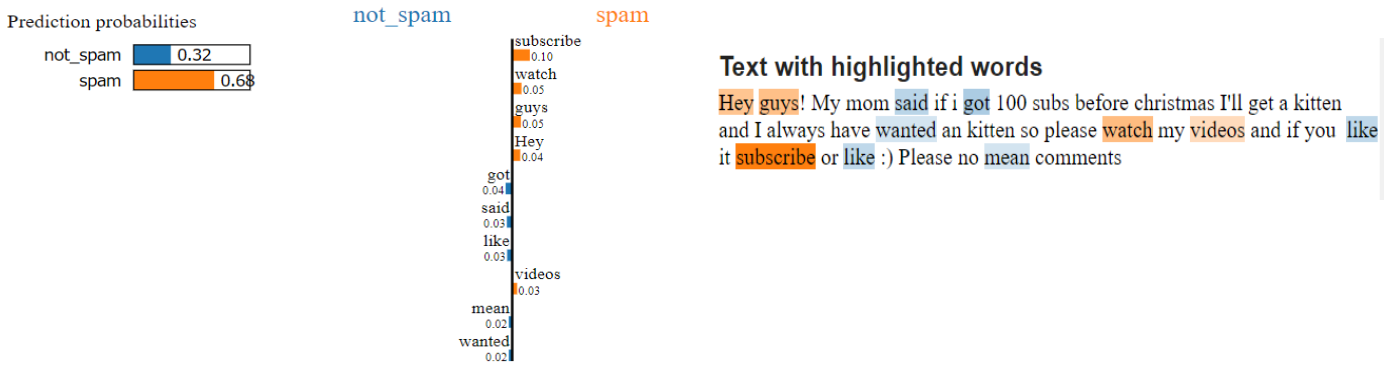
In this example, we classify YouTube comments as spam or not-spam.: The black box model is a *Logistic Regression*-trained on the document word matrix. Each comment is one document (= one row), and each column is the number of occurrences of a given the word. Let us look at the two statements of this dataset and the corresponding classes (1 for spam, 0 for non-spam analysis):

Here are the **two instances**(one accurately classified and one misclassified) by our model with their estimated local weights found by the LIME algorithm:

Example for accurately Classified - Idx: 13

Lime Interpretation

The classifier got this example right (it predicted Spam). The explanation is presented below as a list of weighted features.



In the given instance, visualization of the original document as shown above, with the words in the explanations highlighted. The most terms that affect the classifier are all in the content.

[('subscribe', 0.09638932149124665), ('watch', 0.050744077023384515), ('guys', 0.04751839232368645), ('Hey', 0.04396108060362811), ('got', -0.035949847941526045), ('said', -0.02930286066882812), ('like', -0.02774335820298928), ('videos', 0.026891635731732127), ('mean', -0.01886528332930796), ('wanted', -0.01872949509325923)]
These weighted features are a linear model, which approximates the behavior of the Logistic Regression classifier in the vicinity of the test example. Roughly, if we remove *'subscribe' and 'watch'* from the content, the prediction should move towards the opposite class (not_spam) by about the sum of the weights for both features.

Original prediction: 0.3890290502437696
Prediction removing some features: 0.454946817013648
Difference: 0.06591776676987837

The words that explain the model around this document seem very arbitrary - not much to do with either not_spam or spam. These are words that appear in the content, which makes distinguishing between the classes much easier.

Example for Inaccurately Classified - Idx: 51

Understanding why content is inaccurately classified

As we look at the original text (spam email) wrongly classified as (not_spam), it shows that negation is the problem. ‘Katy’, 'perry' and 'love' are the powerful feature that determines the label. However, the email is spam. To improve the performance of our predictive model, we need to consider negation.

The email is categorized as not_spam as it has words like ‘Katy’, 'perry', and 'love.'

Summary

As you can see in the above examples, LIME-The explainer helps the user understand what's happening behind the black-box model. With the given set of emails(dataset), the Logistic Regression model was able to predict whether it is NOT_SPAM or SPAM, and the LIME explained why. The LIME calculates the weight of the individual words locally in the given data based on the probability and classifies them based on the terms that carried the highest weightage. It is evident that if the words with the highest weightage are removed, the prediction probability varies, making a difference to become the other class.