The LIME algorithm, explaining predictions of black box models

Shadman Ahmed
shadmana@kth.se

Amir Zamli
amiraz@kth.se

Kevin Dalla Torre
kevindt@kth.se

Oguzhan Ugur
oguzhanu@kth.se

***Abstract*— While supervised machine learning is constantly increasing in popularity, it is becoming more and more important to be able to not just have good performance on test data but also be able to trust that the predictions makes sense. The LIME paper proposed a model-agnostic prediction explanation algorithm that showcases the features that a prediction bases its decision on, in an interpretable manner. We investigated this algorithm by implementing it and applying it on two black box models, one purposely overfitted and one better model, both trained on a subset of 20 Newsgroups dataset, to see if we can assess our trust in the model's predictions. Our findings show that both models predictions should not be trusted and that the overfitted model, is even less trustworthy.**

I. Introduction

Machine learning applications are becoming more pervasive. We have AI assistance on the phones, cars, and even in our entertainment systems. Creating an application with machine learning are usually done in four steps. First, we have the data which holds valuable information. Second, we insert this data to a machine learning model which is going to learn and find some underlying pattern from the data. Third, we get our prediction and decision from the model which in the fourth stage is used in the interface of an application that a human can interact with for various purposes. One important factor that is not usually shown in the pipeline of the creation of a machine learning application, is to determine if the predictions and decisions from the model are reliable.

Trustworthiness on a prediction is a matter of trusting a specific prediction that convinces a user to act or not. One great example is medical diagnoses that are predicted by a machine learning model for doctors to use. By just telling the doctor to do a treatment on a patient because the model said so with 90 percent probability, is not going to convince the doctor to proceed. Instead, if we told the doctor that the model showed a 90 percent probability because of certain symptoms from similar cases, will more likely convince the doctor to proceed with the treatment.

The authors who wrote the research paper ""Why Should I Trust You?" Explaining the Predictions of Any Classifier" [1], proposed an algorithm to make any machine learning model interpretable by explaining its predictions. The algorithm is called LIME (Local Interpretable Model-Agnostic Explanation), that explains the prediction in a faithful way by approximation and locally around the prediction. The authors extended further in the paper with a proposal of another algorithm called SP-LIME, how to convince and put trust on a model as well, and not just on its decisions. Trusting the model can inflict confidence to users that the model can make good decisions on real-world data.

In this paper, we are going to test the LIME algorithm on the same dataset that was used as an example in the paper for text classification. And evaluate how well the explanation technique works on a different text dataset.

We expect that the explainability analysis using the methods mentioned in the paper will also show us how poor black box models base their decisions on irrelevant features while better performing models base their decisions on more relevant features.

II. Method

A. Dataset

We are going to replicate part of their findings on text classification explainability on the 20 Newsgroups dataset and then see the applicability on a different set. The data consists of many text emails from 20 different categories (Christian, atheist, hockey, baseball etc) [2]. To work with the data, we are going to utilize the Sklearn library in python. It has the functionality to automatically fetching certain categories and splitting the data into train and test sets.

B. Pre-processing

To train a simple machine learning black box on a text dataset, we need to pre-process the text into a tabular data type, one convention is to use TF-IDF transformation [3]. Which is basically a frequency count of how many times a word appears in a document, but it also considers how many different documents the word appeared in.

C. LIME algorithm implementation

We started by creating two black box models using Sklearn library, one decision tree with default hyperparameters (default selected by Sklearn library), that represented our bad model. For the good and second black box model, we built a random forest with 100 trees and tuned the hyperparameters using random grid search with 2-fold cross-validation. Usually one would use 5-fold or 10-fold cross-validation, but in our case, we had limited resources, so we settled with 2-fold.

After creating the models, we needed to assess our trust in the model's predictions, and in our case the tool for trust-assessment was LIME. The algorithm works by training an explanatory model around the vicinity of a data point, in our case, an email text. This means that our explanatory model should be able to approximate the behaviour of our model in the vicinity of that data point. The explanatory model we used was a simple linear model with weights corresponding

to 6 numbers of features. We tested if the model successfully approximated the black box by selecting three random data points from the test set. And then using them with LIME, which gave us words (features) that explained the prediction and then later removed 2 of them in order to measure the importance of these features and see the effect it has on the black box model's prediction. If the LIME algorithm worked correctly, one should see that the probability shifts towards opposite class after removing the opposite feature weights, for example the probability of belonging to a positive class should increase by removing weights with negative numbers and vice versa [1].

D. Performance metrics

We measured the performance of the black box model using ROC-AUC as it is robust to different class distributions. And will measure the explainability by looking at few instances from the test data and see which features the models take into account in order to make "correct" predictions.

III. RESULTS

After the LIME algorithm was implemented, we needed to see if the explanation method worked for the 20 Newsgroups dataset by looking at the weighted features from the linear model.

A. Evaluation of the LIME Algorithm

For the first example, we used the dataset containing emails labeled as atheists and Christians. We used a random forest as our black box to do the binary classification task. The RF model had the following AUC scores with default parameter setting and 500 number of trees: $AUC_{train} = 1.0, AUC_{test} = 0.897$.

```
From: jenk@microsoft.com (Jen Kilmer)
Subject: Re: Homosexuality issues in Christianity
Organization: Microsoft Corporation
Lines: 27

In article <May.11.02.36.59.1993.28108@athos.rutgers.edu>
dps@nasa.kodak.com writes:
>In article 15441@geneva.rutgers.edu, loisc@microsoft.com (Lois
Christiansen) writes:

>|>he can, especially homosexuality.  Let's reach the homosexuals for
Christ.
>|>Let's not try to change them, just need to bring them to Christ.  If He
>|>doesn't want them to be gay, He can change that.  [....]

>don't hate the people.  I don't.  I don't hate my kids when they do
>wrong either.  But I tell them what is right, and if they lie or don't
>admit they are wrong, or just don't make an effort to improve or
>repent, they get punished.  I think this is quite appropriate.

Note the difference here. One is saying, if *Christ* disagrees with
a Christian being gay, *Christ* can change that.

The other is saying, if *I* think being gay is wrong, that a Christian
cannot be gay, *I* need to tell them to change.


As Lois said, and as before her Paul wrote to the believers in Rome,
WHO ARE YOU TO JUDGE ANOTHER'S SERVANT?

-jen

--

#include <stdisclaimer>  //  jenk@microsoft.com  // msdos testing
```

Figure 1, Data point from the test set with the label Christian.

The first unseen data we used in our LIME algorithm, to check the prediction of the model locally, can be seen in fig 1. We received the following weights from the first linear model.

TABLE I.

Explanatory features	***'from'***	***'christ'***	***'try'***	***'com'***	***'article'***	***'writes'***
Corresponding LM weights	0.052	0.037	-0.018	-0.021	-0.034	-0.035

Table 1, Feature weights from a linear model using figure 1 datapoint.

In table 1, one can see which features the model relied on for its prediction according to the LIME analysis and each feature coefficient magnitudes reflects the importance of the features. Positive weights mean that the feature increase the probability of the topic being Christian and negative weights means that the feature increase the probability of being atheist. Later we checked if the probability shifted to the opposite class after removing the important features. The prediction for the datapoint depicted on figure 1 was as following, Atheist: 0.208, Christian: 0.792.

TABLE II.

	Removal of ***'from', 'christ'***	Removal of ***'article', 'writes'***
Probabilities after removing features	Atheist: 0.284 Christina: 0.716	Atheist: 0.170 Christian: 0.830

Table 2: Probabilities after removing important features from table 1.

According to the authors of [1], this removal should move the predictions towards the opposite class by the sum of all weights corresponding to the removed features, we can see this behavior above. This means our explanatory model successfully learned how our black box model behaved in the vicinity of the data point, in our case the datapoint fig 1.

We also tested with another dataset from the 20 Newsgroups. This one contained data with hockey and baseball as the labels. We conducted the same procedure and got our new random forest model with the AUC scores: $AUC_{train} = 1.0, AUC_{test} = 0.92$. And used the following datapoint for the LIME Algorithm, see fig 2.

```
From: bpenrose@morgan.ucs.mun.ca (Brian Penrose)
Subject: Re: Trivia question
Organization: Memorial University of Newfoundland
Lines: 19

In article <1993Apr23.102811.623@sei.cmu.edu> caj@sei.cmu.edu (Carol
Jarosz) writes:
>
>While watching the Penguins/Devils game last night, I saw the "slash" that
>Barrasso took on the neck.  This brought to mind the goaltender who had
his
>jugular vein cut by a skate.  I think he was a Sabre, but I'm not
positive.
>Does anyone remember/know his name?  What has happened to him since?  What
>about the player whose skate cut the goalie?  Name?  Info?  Has this ever
>happened before in a hockey game?
>
>Thanks,
>
>Carol
>Go Pens!

His name is Clint Malarchuk.  I'm not sure what he does now but I've heard
he's an extra in slasher films.

--Brian
```

Figure 2, Data point from the test set with the label hockey.

TABLE III.

Explanatory features	***'hockey'***	***'devils'***	***'penguins'***	***'pens'***	***'ca'***	***'cmu'***
Corresponding LM weights	0.097	0.071	0.062	0.059	0.053	0.042

Table 3, Feature weights from a linear model using figure 2 datapoint.

The prediction for datapoint 2 was as following, baseball: 0.278, hockey: 0.722.

TABLE IV.

	Removal of '***hockey***', '***devils***'
Probability after removal of features	baseball: 0.284, hockey:0.716

Table 4: Probabilities after removing important features from table 3.

In table 4, one can clearly see how the probability of belonging to the hockey class, decreased after having removed the explanatory features explaining hockey for the datapoint.

B. Explainability analysis of model predictions

After training two black box models we assessed their performances on the whole training set and test set from the data containing the two labels of baseball and hockey, which resulted in the following AUC scores:

Black box overfitted model (BBO):

$$AUC_{train}: 1.0,\ AUC_{test}: 0.87.$$

Black box "good" model (BBG):

$$AUC_{train}: 0.99,\ AUC_{test}: 0.93.$$

Next step was to increase our trust in the two black box models predictions using LIME. We randomly selected three data points from the test dataset, all of which are emails with topic label hockey. We then compared which features that our implementation of LIME deemed was most explanatory for the prediction, we then assessed how well these features fit the topic label using our own judgment.

The results are displayed in table 5 for each randomly selected datapoint. The table shows the features, the models relied on while making the predictions, according to the LIME analysis. We only included the weight features that explained hockey from the predictions since the datapoints had the label hockey.

TABLE V.

First datapoint		Second datapoint		Third datapoint	
BBO	BBG	BBO	BBG	BBO	BBG
'ca'	***'playoffs'***	***'wc'***	***'playoffs'***	***'hockey'***	***'hockey'***
'montreal'	***'wings'***	***'hockey'***	***'hockey'***	*'ca'*	***'pittsburgh's'***
'wings'	***'quebec'***	***'playoffs'***	*'telecom'*	***'pittsburgh's'***	***'leafs'***
'koharskiworst'	***'toronto'***	*'gp'*	*'contact'*	*'go'*	***'buffalo'***
	'ca'	*'h'*	***'wc'***	*'and'*	*'ca'*
	'detroit'		***'sweden'***		*'cmu'*

Table 5, Explanatory features from 3 random datapoints, explaining of belonging to the hockey class.

Both models correctly classified all three data-points (emails) as hockey, which is correct, then the goal was to increase our confidence on the trustworthiness of the predictions. Using our knowledge of sports and a couple of short google searches we marked (bold) which features we deem to be "good" and therefore be features that would increase our trust in the prediction.

We can see that for the first datapoint, only two out of four features are marked as good indicators for the sport hockey, while the non-overfitted model (BBG) had five out of six, based on this information solely, we would trust the second model more. We can see this pattern on the second and the third data point as well, where it also seemed that the non-overfitted model based its decision on more relevant features compared to the overfitted black box model (BBO).

IV. Discussion

Our goal was to increase our trust in two model predictions and compare which features they use to predict the topics. For this part, we looked at emails about the topics baseball and hockey. We constructed two black box models, one bad and one better to see if we can see if the quality of the models is reflected in our LIME analysis.

The findings reflected that the overfitted black box model based its prediction on a larger number of bad features as compared to the better model. But nonetheless, both models always had at least one out of the total features that the LIME assessed as explanatory features, but we assess as bad features to base decisions on. This means that our trust decreased in both models. But more so for the bad/overfitted model.

It was interesting to see how we did not need to use the same data representation for the LIME explanatory model, which in our case is a linear ridge regression model, we simply used a binary multi hot encoding of all the words from the selected instance to train the explanatory model.

Further work in this study would be to continue to replicate the workings in the paper, to see how we can handle image data and how to trust a whole model, and not just its predictions. Because in our case we just picked three random data-points to assess the predictions on using LIME. For this to be a robust trust assessment we would need a better way or larger sample size in order to be able to gather enough data for any kind of trust assessment.

REFERENCES

[1] Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. CoRR abs/1602.04938 (2016)

[2] 5.6.2. The 20 newsgroups text dataset — scikit-learn 0.19.2 documentation, URL: https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html. Last updated: 23 december 2018

[3] Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining, URL: http://www.tfidf.com/, Last updated: 31 mars 2018