CS301

Section 104

Professor Pantellis

Tutorial on using LIME

GOAL:

During our research about LIME we found how important it is that the common person understands what machine learning models are based on. We believe that the common user sees this as too complex with the inner workings of the machine learning algorithms and models. Therefore as a fellow computer science student, our goal is to explain to the common user how to understand these models and learnings.

What is LIME?

Lime to the common user is to break down Machine Learning Models into understandable bits and pieces to understand and trust how a decision is made by the computer. The common user is always filled with questions such as why should I trust that the results are accurate. One of the most common usage of this is based on image based decision making. The algorithm will look at an image and see if the image is a wolf or not, but if the computer is basing the decision on if there is snow in the background then it is a wolf it is untrustworthy. On the other hand if the computer is basing the images on the animal found in the picture and if it has the features of a wolf, such as the shape, eye color, fur color and patterns then it is more trustworthy than basing off of a background of snow. As a result, if the user understands why we picked

if the image is a wolf or not, we will have more success in gaining the trust. As a user and consumer you want to understand why the computer made the decision it made. If you do not understand , you will fail to trust it.

What does LIME stand for?

L - Local

I - Interpretable

M - Model Agnostic

E- Explanation

Using the bottom method:

Explanation means that we can explain the purpose of the code and the decision making behind the models.

Model Agnostic means no matter the data or information, it will work for anything. This is known as a “black box”

Interpretable means that humans can understand the explanation of the decision making compared to a complex box of code and such.

Local means that is predicted in a local area in the models

LIME also includes questions such as why the prediction occurred and what can cause the models to move in a certain direction. We will see later in this paper referring to wolves of what can cause a model to predict it.

Issues with LIME?

One of the biggest issues many run into while using LIME is that LIME is based on a small area in a model. Therefore when it is needed for the the entire model to be analyzed instead of a single area, it can cause the interpretation of the data to be inaccurate.

In what way can we make it so that users can understand more?

Using pick step is one of the most important ways that a user can understand why the model made the decision it made. Pick step allows for the model to show how often something occurs. It allows for the user to inspect the model and see the chosen data. It is important to note the pick step when selecting a data for that the data must not be isolated in one area. Therefore instead it must be diverse and widespread so that the user can believe this model more. The data must also be picked where it has some significance and not redundant to the user. Not only that, the data must as well be unbiased and honest. We must also ask the following questions

1) Are the explanations true to the model? [1]
2) Can these explanations be understandable and trustable? [1]
3) Are these explanations useful to the user? [1]

Other parts in which we can allow the user to trust these explanations is including data which is untrustworthy. This amount is going to be set to 25% of the data. The other 75% of the data will be trustworthy. Part of this method also includes training our model to learn from trustworthy data at the start so that it can predict easily and eventually introducing untrustworthy data so that it can start learning mistakes and push out bad information. If the users feel the data is untrustworthy we can fix this issue in a few

ways. One of the ways which is removing and adding features that the users feel like may help get more accurate results.

In an example during the article [1] we learned that when selecting your information to pick , for example the wolf and husky problem it is important to understand what it is selecting. At the start the problem began with 10 wolves in a snowy background. Eventually they added some without snow as the background. While you cannot base if something is a wolf because of the background. It can be an attribute and feature to determine if it is a wolf. If the experts understand that there are faults and can result in a failure, it is key to note we must allow the user to know about these failures or potential because it will cause them to trust the system more. Trust is easily lost if the user is not informed about these possible errors.

Conclusion:

Because the usage of LIME can help both parties of expert and non expert in a topic so that the user can understand more of why predictions occur it is most useful to the Machine Learning industry. LIME overall allows us to choose which model will fit the data the best, see how truthful the data is, figuring methods on how to make data and models which are untrustworthy into becoming trustworthy and understanding the predictions. We made sure that LIME is for any machine learning model instead of one model. This is known as a “black box”

Citation:

[1] https://arxiv.org/pdf/1602.04938.pdf