Ashween S., Katie P., Batool A.

Professor Pantelis

Data Mining

April 4, 2021

"Why Should I Trust You?"

Explaining the Predictions of Any Classifier

The article "Why I Should Trust You" explains whether predictions and models should be trusted based on explanations. Both of these factors will determine if a human being will use the machine or not; which depends on how well a human being understands the behavior of the model. Machine learning is used for multiple purposes, thus trusting the predictions becomes a necessity "when the model is used for decision making." Furthermore, a model needs to be inspected before it can be deployed in the real world. However, the inspection might not produce a desired outcome if the dataset is large, therefore it's crucial to determine the various instances to evaluate. The article uses LIME (Local Interpretable Model-Agnostic Explanations), "a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner," to support their explanations and findings "by learning an interpretable model locally around the prediction."

As mentioned above, an explanation is a key factor in determining either a model should be accepted or not. Hence a detailed description that "provides qualitative understanding of the relationship between the instance's components and the model's predictions" will allow a person to use and trust the model more effectively. There is a possibility for an assessment to go wrong and mess up the whole process, however, through an explanation it's possible to turn an unreliable model into a reliable one. LIME, which identifies an interpretable model, comes in handy in these situations. Although a model might be too complex to be explained, LIME makes sure “an explanation is locally faithful” which can explain the difference between features and interpretable representation. An “interpretable explanation needs to use a representation that is understandable to humans,” despite the existing features of a model.

Although an explanation can enhance the understanding of a model, enabling humans to trust and deploy it, it is however not sufficient to “evaluate and assess trust in the model as a whole.” It requires a lot of time and patience to understand a model with a large data set, on that account, specific instances must be picked to inspect. Redundancy must be avoided while explaining a model, although it seems inevitable with large data sets. The article uses a toy example, represented by rows and columns, to illustrate what to pick to prevent redundancy. Once a certain row/column is chosen, the rest loses its value. This method is known as a submodular pick algorithm, which “computes the total importance of the features that appear in at least one instance in a set.”

The article continues the evaluation of an explanation through introducing simulated user experiments. This experiment addresses the questions such as if the explanations are true to the model, can the explanations assist users in developing trust in their predictions, and are the explanations applicable for assessing the model as a whole. To get an adequate understanding of these questions, the article compares LIME with parzen. Parzen is a method “that approximates the black box classifier globally with Parzen windows, and explains individual predictions by taking the gradient of the prediction probability function.” The comparison of these two methods shows that parzen is less likely to produce a faithful model, whereas LIME explanation has over 90% chance of a faithful model. The question nonetheless, still remains whether this prediction should be trusted or not. To answer this question, another experiment (Figure 8) was conducted where two classifiers were chosen and compared to different models. Out of four different methods, LIME gave the best prediction. The rest either mistrusted the prediction or they trusted way too many predictions. Hence, it proved that LIME is "helpful in assessing trust in individual predictions."

LIME has demonstrated its ability to predict faithful models, now it needs to show if the model can be trusted or not. To prove this theory, an experiment was conducted "where a human has to decide between two competing models with similar accuracy on validation data." The purpose of this experiment was to determine if a user would be able to spot the preferred classifier based on the explanations of an instance from the validation set. The classifiers with the least doubtful predictions were selected and were compared with the most reliable test set. This trial enabled the situation where the models don't only regenerate informative features but also initiates false correlations. Models with 0.1% of accuracy were labelled as trustworthy components. Many other experiments were completed in a similar manner to figure out whether LIME is a better method compared to parzen, RP-LIME, and the outcomes of these tests showed that LIME "explanations are good indicators of generalization, which [is] validate[d] with human experiments."

After showing the model is trustworthy with 0.1% accuracy, it is time to evaluate LIME and SP-LIME using human subjects in the following settings: "(1) Can users choose which of two classifiers generalizes better (§ 6.2), (2) based on the explanations, can users perform feature engineering to improve the model (§ 6.3), and (3) are users able to identify and describe classifier irregularities by looking at explanations (§ 6.4)." The experiment used the original 20 newsgroups dataset, and a new religion dataset for evaluation. Human subjects were then recruited, not for their technical knowledge but for their knowledge on religion. It was shown that the submodular pick (SP) greatly improves the user's ability to select the best classifier when compared to a random pick (RP). LIME outperformed greedy in both cases. Another experiment was performed with human subjects unfamiliar with feature engineering, and it was clear that the crowd workers can improve the model by removing features they deem unimportant for the task. This also demonstrated that machine learning knowledge is not required. The final experiment aimed to show if human subjects can discern when to say a classified cannot be trusted, the experiment involved distinguishing between photos of Wolves and Eskimo Dogs (huskies). The classifier showed instances of Husky's in snowy backgrounds, classified as wolves. Once the subjects were shown the explanations, almost all identified the correct insight and the trust in the classified dropped substantially.

This paper argued trust is crucial for effective human interaction with machine learning systems, and proposed LIME to explain the predictions of any model in an interpretable manner. Their experiments demonstrated that explanations are useful for a variety of models in trust-related tasks in the text and image domains. Expert and non-expert users were used in deciding between models, assessing trust, improving untrustworthy models, and gaining insights into predictions. The article also mentioned related work where tools complementary to LIME are used, and future work, such as, decision trees. They mention one issue that was not mentioned was how to perform the pick step for images. They would also like to investigate the potential uses in speech, video, and medical domains, as well as recommendation systems.

Task 4 - LIME Explainer VS Confusion Matrix Classifier 2 page summary of findings

The term machine learning was coined back in the late 1950's by an IBMer in the field of computer gaming and artificial intelligence. However, theorems have existed since the 1700's to predict pattern recognition. Fast forward to the present, organizations are using machine learning on a variety of data collected to build predictions and to create efficient processes. The various areas using machine learning, may range from advertising, fraud detection, spam email recognition, investing, lending, etc. For organizations to continue to grow and for workers to be skilled in today's labor market, schools are creating programs for Machine Learning and Data Science. However, with the growth of machine learning, how can you trust the classifiers prediction? The LIME method was built to understand the prediction, and to explain what the machine learning classifier is doing. In this report we will discuss using the Naive Bayes classifier for multinomial models with a confusion matrix using sklearn, and the LIME method as an explainer for the data set.

A confusion matrix is used to identify and describe the key performance of a classifier. Confusion matrices present direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives. The True Positive is where the data is predicted yes and True Negatives the data is predicted no. On the other hand, False Positives is considered as Type I error, since the data is predicted as yes, but the data output should be no. Whereas, False Negatives data is considered as no, but the output data is yes, this is also known as Type II error. In Task 3 of the project in "Why Should I Trust You", we first took the approach of using a Naive Bayes classifier for multinomial models, and then applied a confusion matrix using sklearn, to the Tubespam dataset. A multinomial Naive Bayes classifier is suitable for classification with discrete features, such as, word counts for text classification. The confusion matrix is used since classification accuracy alone may be misleading for unequal numbers of observations for each class, if more than two classes are in a dataset. Calculating a confusion matrix allows for better insight if your model is right, or to provide the types of errors being made.

After choosing a Naive Bayes classifier for multinomial models with a confusion matrix in Task 3 against the Tubespam dataset, for Task 4 we applied the LIME explainer on the classifier. LIME is known as local interpretable model-agnostic explanations, which is used to explain individual predictions of black box machine learning models. The LIME explainer, as discussed further in detail in the tutorial Task 1 of the Project, is an important tool to explain what the machine learning classifier is used to perform. As mentioned previously, the technique is model agnostic, meaning LIME may be applied to any machine learning model. In LIME variations of the data are generated differently. For instance, taking the original text, new texts are created randomly by removing words from the original text. Then the dataset is represented with binary features for each word. That same feature is divided between “1” or “0”, where 1 is presenting the corresponding word and “0” if the text has been removed.

Both Task 3 and Task 4, used the YouTube Spam Dataset from UCI Machine Learning Repository Dataset site, which contains a collection of public set of comments for spam research. The public comments were collected from the top ten most viewed videos on the collection period. The comments were manually labeled as spam or legitimate. Spam was coded with a "1" and legitimate comments with a "0". The precision, accuracy, weight average, and macro average values are different for each classification in Task 3 and Task 4. For instance, taking Katty

Perry's repository dataset as a sample data, where both the index equals to one hundred. In Task 3, the confusion matrix demonstrates “0” precision of eighty-three percent and “1” precision of eighty-one percent. Nonetheless, the LME Tabular implementation, which illustrates “0” precision of eighty seven percent and “1” being at ninety percent. Also, there is a difference of 0.7 percent in accuracy marco average and weighted average between confusion matrix and LIME explainer. LIME explainer is providing more accuracy in features than the Naive Bayes classifier for multinomial models with a confusion matrix.