2019/11/17

Q&A for Lecture 8 - ER Explanations

Why should I trust you? Explaining the predictions of any classifier

Zhou Fang

Q1. *In section 3.3, given a perturbed sample z', how can we recover the sample in the original representation z?*

A: We have defined how to convert the original input into an interpretable representation, where recovering is its reverse process. For images, we need to recover the non-zero super pixels in the binary vector. For text, we will discard those word tokens out of the subset bag of words, then we will recover the word tokens to the origin words.

Q2. *In section 3.4, it is stated that "we can estimate the faithfulness of the explanation of Z". What is the definition of faithfulness and how can we estimate it if the underlying model is highly non-linear in the locality of the prediction?*

A: The faithfulness in this paper is defined as the Fidelity Function. In this paper, they assume the model being explained is linear in the locality. We can still estimate the Fidelity Function if the underlying model is highly non-linear in the locality of the prediction; however, the explainer may not work even if the return of Fidelity Function is low.

Q3. *In submodular pick, what is the intuitive explanation for instances that represent completely different features? Does that imply the nonlinearity in the underlying model?*

A: The idea of submodular pick is to find a subset of instances covering more interpretable components so as to show the comprehensive explanations to the users. Thus, the instances whose explanations represent completely different features reveal the different aspects (different interpretable components) of the underlying model. I believe these instances don't imply the nonlinearity of the model be explained.

Sean Singh

Q1. *In section 2, it seems that the emphasis placed on explanations and human made decisions is at odds with the idea of the autonomous systems?*
Do you feel that there are certain domains where interpretability is not important?

A: First, the explanations are also generated automatically, which won't conflict with the autonomous system. Second, the explanations are proposed to help machine learning practitioners to understand the black-box models.
If the traditional metrics meet the requirements of practical applications, the interpretability won't be such important.

Q2. *In section 5.3, what is the reason for using 25% of the features?*

A: They randomly choose 25% attributes as the untrustworthy ones to simulate trust in individual predictions.

Q3. *In section 6.3, wouldn't using explanations to engineer features result in a self-fulfilling prophecy of sorts?*

A: Indeed, this explanation may self-fulfill the classifier. However, the explanation is calculated based on the specific classifier, which means using explanations to feature engineering will improve the better classifier and deprave the worse classifier.

Mustafa Barez

Q1. *How does submodular optimization work for SP-LIME?*

A: As I have mentioned on the presentation, it is NP-hard for the pick step to find the best subset of instances. Fortunately, the coverage function satisfies the definition of submodular function. Then we could use the property of submodular function, which guarantees the approximation with the error of 1-1/e to the optimal with a greedy strategy.

Q2. *Are there any other ways besides data leakage and dataset shift that would make an evaluation go wrong?*

A: Such as the adversarial perturbation as mentioned on my presentation.

Q3. *Out of the two algorithms presented, is one more preferable than the other?*

A: I guess the two algorithms you mentioned is LIME and SP-LIME. LIME is proposed for explaining predictions while SP-LIME is for the whole model, where the latter is based on the former.

Yinbin Zhang

Q1. *For an instance x (blue circle), if the decision boundary around it is highly non-linear, how to choose the interpretable model? Do you think the interpretable model can still be local fidelity?*

A: In this paper, they assume the model being explained is linear in the locality. We can still estimate the Fidelity Function if the underlying model is highly non-linear in the locality of the prediction; however, the explainer may not work (i.e. not be local faithful) even if the return of Fidelity Function is low.

Q2. *Why the submodular pick algorithm can ensure the diversity of picked instance?*

A: Diversity is reflected by the coverage of interpretable components. The larger the coverage, the larger the diversity, when the number of instances is guaranteed to be the same.

Q3. *How to determine the number of instances we are going to pick when applying SP-LIME?*

A: This paper uses Budget B to represent users' patience. Thus, the number of instances should be decided by users themselves.

George Wu

Q1. *What do you think is the most important thing for a classifier to report to trust it?*

A: For the input instances which are close (similar) to each other, the output of them given by the classifier should be similar.

Q2. *Do explanations just benefit the user from an understandability standpoint?*

A: It also benefit the machine learning researcher as they could use augmented dataset to improve the classifiers.

Q3. *Could you clarify how instances are chosen to be explained?*

A: The instances are chosen by the step - submodular pick - which selects a new instance from the set of all the instances each literation.

Bowen Zhang

Q1: *How can be data leakage be distinguished? Would it be done manually or can it be identified by some learning methods?*

A: LIME can be used to detect data leakage as the weights of the attributes (including index attributes) are calculated by the model. We can see the importance of the irrelevant features to show whether the model has the risk of data leakage.

Q2: *Would keep the global fidelity with local fidelity together for the fidelity measurement represent better how good the model is?*

A: This model is designed to measure the local fidelity of the model being explained.

Q3: *How will the result of LIME guide us to change the training process?*

A: We could augment the training set with the bad instances detected by LIME to improve the robustness if the classifiers.

Interpreting Deep Learning Models for Entity Resolution - An Experience Report Using LIME

Zhou Fang

Q1. *What will happen to the weights learned by the surrogate when labels C(Ti) do not span uniformly c and other classes?*

A: For this problem, which is a two-category one, the unbalanced distribution of categories you mentioned will generate too many positive or negative samples, where it will affect the weights learned by the surrogate as too many positive instances will make most weights contributes to positive class. It is also the reason why this paper proposed LIME_COPY methods.

Q2. *Does LIME_COPY provide more advantages compared with LIME_DROP? Or LIME_COPY just complements LIME_DROP?*

A: LIME_COPY just complements LIME_DROP to generate more negative instances.

Q3. *In section 3, what is the benefit of processing the input word embedding sequence by both forward RNN and backwards RNN?*

A: It is not the main focus of this paper so I consult Yibin Zhang. He said it will capture the patterns of symmetric information.

Sean Singh

Q1. *In section 1, what are the limitations of using the F-measure described?*

A: The F-measure is calculated on the labeled set, which might be different from the real-world data. Also the F-measure is not enough to make people trust a model.

Q2. *In section 2, how well would the methods of this paper extend to multi-class classification as opposed to binary classification?*

A: The explainer is a linear model, which could be extended to multi-class classification.

Q3. *In section 5, when they mention using explanation to produce suggestions, what does suggestions mean in this context?*

A: The sensitivity of this model to both the training set and test set; how can we detect and overcome the problem.

Mustafa Barez

Q1. *For figure 5, what is the significance of high-levels of ale?*

A: It actually shows the sensitivity of this model to some specific words as ale.

Q2. *How does the hybrid solution work in section 3 of the paper?*

A: It is not the focus of this paper. You could refer to previous papers or consult Yibin Zhang.

Q3. *What problems arise through Figure 2 of the itunes-amazon dataset?*

A: Some attributes which are important for humans is not important for the classifier, where time is such significant, which means this model might be sensitive to time attribute.

Yinbin Zhang

Q1. *When we build the representative text sequence Tu,v in Mojito, a prefix is added to each token. So, it actually modifies the original attribute values. How does this additional prefix affect the result of LIME?*

A: The prefix is just used to construct the input text sequence for the classifier.

Q2. *In Figure 3, LIME_COPY and LIME_DROP are performed on both hybrid and RNN ER models. Base on the results, which ER model do you think is more trustable?*

A: Both of them are sensitive to time attribute and give less preference to the attributes which are important for human to classify the entity pairs.

Q3. *In Figure 2, knowing that the time attribute contributes too much during ER, which is unreasonable. How do you fine-tune the models or training data such that make it more reasonable?*

A: Under the same time attribute, try to change other attributes to generate a new training set, and improve the importance of other attributes.

George Wu

Q1: *Is Mojito just explaining Deep Learning models for Entity Resolution?*

A: They not only applied the Mojito methodology to evaluate different DL models of the DeepMatcher analysis for Entity Resolution, but also demonstrated the importance of developing solutions for the interpretability of DL models in the ER context.

Bowen Zhang

Q1: *Recall the paper "Deep Learning for Entity Matching: A Design Space Exploration" we read in last lecture, could breaking the class matching into textual matching a good idea for weight the matching and get the explanations?*

A: The different of explanations between entity matching and text matching is that the interpretable components for entity matching are attributes, while it is tokens for text matching. Consequently, they should be two different problems.

Q2: *How could Mojito deal with the problem of data leakage as mentioned in the first paper?*

A: The data leakage is a problem of classifiers. Mojito could calculate the importance of each attribute and detect the data leakage of the irrelevant attributes.

Q3: *How can we use Mojito to make our learning more interpretable?*

A: The apply of Mojito into deep learning has show the interpretability of these black-box algorithms.

On the Robustness of Interpretability Methods

Zhou Fang

Q1. *Does local stability imply global robustness?*

A: This paper believes the local continuity (stability) could be used to represent the robustness of the model.

Q2. *In section 2, why is the continuous notion of local stability not suitable for models with discrete inputs or those where adversarial perturbations are overly restrictive?*

A: These models have discrete inputs so that they cannot meet the definition.

Q3. *In section 4, what is the possible explanation for the experiment result that model-agnostic perturbation-based methods are more prone to instability than their gradient-based counterparts?*

A: As the single point-wise explanation is perhaps too optimistic.

Sean Singh

Q1. *In section 1, what is your opinion on the statement that "understanding a complex model with a single point-wise explanation is perhaps too optimistic"?*

A: The methods with perturbation approaches, such as LIME, use some specific instances to understand a model, which might be too optimistic.

Q2. *In section 2, how should one interpret the word reasonable at the bottom of page 67?*

A: The choice of different notion of robustness should depend on real application.

Q3. *Would you agree that there is a trade-off between complexity (and likely performance as well) and interpretability?*

A: For some models, it is true, such as linear models, decision tree, while some complex models are designed to be hard to interpret, such as neural networks.

Mustafa Barez

Q1. *How would you define robustness?*

A: In this paper, they define robustness as the local Lipschitz continuity.

Q2. *How do LIME and SHAP's explanations for linear models differ?*

A: This paper didn't show the robustness of LIME and SHAP's explanations for linear models. They show the interpretability of explanations of random forest, logistic regression and neural networks.

Q3. *Why does local stability not work for models with discrete input?*

A: These models have discrete inputs so that they cannot meet the definition.

Yinbin Zhang

Q1. *If the pattern of a problem is highly non-linear, which means similarly inputs should map to very different output. Is it reasonable to apply such robustness analysis on the problem?*

A: I guess not. It runs against the assumption made by this paper.

Q2. *Suppose x0 is the vicinity of x, the small |f(x) – f(x0)| represents the robustness. However, why we maximize L hat in equation (1)?*

A: The maximal [f(x) – f(x0)]/(x-x0) is the robustness.

Q3. *For a machine learning model, is it possible that it's robust around some decision boundary but not robust around other? In such case, how to measure the overall robustness of the model?*

A: Yes, it is. This paper do not try to measure the global robustness. They believe the local continuity (stability) could be used to represent the robustness of the model.

George Wu

Q1: *Is this robustness given necessarily true? If I'm training parity, 1&2 are "close", but they should give different inputs.*

A: The robustness can be used to show the sensitivity of this model to some specific instances.

Q2: *Could robustness be determined by the fact that multiple models can have the same explanation?*

A: If you could define the metric to measure whether two models should have the same explanation, then yes.

Bowen Zhang

Q1: *How will all the locally Lipschitz be summarized as the robustness?*

A: This paper believes the local continuity (stability) could be used to represent the robustness of the model.

Q2: *How will the robustness method guide the learned model to be more robust?*

A: We could augment the training set with the perturbation data (i.e. the data instances for which the model is sensitive).

Q3: *Would it be more informative if the robustness is calculated as the information gain of the model/explanation from the different input?*

A: As I have introduced in the class, some proposed the adversarial perturbation to represent the robustness of the models, where the adversarial perturbation can be seen as the gain.