COMS 6998-4 Fall 2017 Presenter: Wenxi Chen
November 15, 2017 Scribe: Wei Zhang

Explainable & Interpretable Models

1 Introduction

This lecture is delivered in more philosophical sense rather than technically. In the past, we have learned a series of algorithms which can dig deeper in a training dataset and generate a model based on such a dataset. However, without knowing of what knowledge is contained in the model, it is generally hard for human beings to trust the trained model and apply it in reality. Sometimes, biased model can really hurt people in real life. Thus, it is important for us to understand trained model better.

1.1 Motivation Example

Image that some researchers have trained a complex ML model to perform automatic medical treatment and recommendation. There is a question that people who actually use it might ask: can researchers explain to the users about this model? The possible answers might be that the error bound of this model is 3% or the empirical error is 3%. Although it might be a satisfactory answer for someone, this explanation might not be enough for someone who really doubts the power of a ML model. In general, this type of explanation might be too vague to be a good one. It might be helpful when we ask the trained model for a specific question.

For example, we might want to input some numbers which come from medical examinations to the model so that it can form suggestions to improve our health status. If the generated answer is "Wenxi, you should stop eating meat for three weeks." then different people might choose to trust this suggestion or not. This is basically leading to the issue of trust on a trained model.

Last but not least, after we train a Ml model, it is also frequently noticed that the knowledge that comes up with the model can also be used for further inference. The knowledge might also be used for discovering more unknown medical theory about human beings. However, the question remained is how can we use this type of knowledge?

All of these question are unclear. This lecture tries to reveal several possible solutions to answer those questions.

2 Felten's Explainability Problems[1]

It is interesting to notice that human brain is actually more complex than any algorithms. This is because human brain is usually biased based on personal past experience, which makes human brain unpredictable. To this extent, any algorithm is more predictable than human

brain since we can at least analyze an algorithm mathematically. However, this does not mean that people cannot ask for explanation of a trained model. Instead, we can categorize all such question into four types of explainability problems.

2.1 Claim of Confidentiality

The first type of explainability problem is called claim of confidentiality. This is usually resulted from an institutional or legal constraint rather than algorithm itself. For example, when Microsoft Visual Basic pops up a window showing some error messages, it is apparent that someone who wrote this code knows what might go wrong. However, due to legal issue, such explanation of what go wrong become much less obvious when it is displayed to user end. In this case, it is hard for people to understand what is happening inside except for those who actually design it. Similar phenomenon can be observed in many circumstances such as trading market.

2.2 Complexity

The second type of explainability problem is called complexity. In this case, we know everything about the model. However, the model itself might be too complex for human to understand and interpret. For example, Deep Neural Networks with non-linear activation functions such as Relu can be really hard to fully understand the model and explain it in a formal way. Thus, we will have the trouble of "big-picture understanding" in this scenario. However, it is still feasible to ask the model "What-if" question. For example, if I change an input by certain amount, how much the corresponding output would have changed? By doing this, we can at least have some sense on which factors might have changed a particular output by how much.

2.3 Unreasonableness

The third type of explainability problem is called unreasonableness. In this case, we know everything in the model, and we can even make a statistical analysis on the model. However, the predictive results do not make sense. For example, if people ask the model about where to have a haircut, the model returns to not eat meat, which does not make sense at all. Thus, in this case, it is not hard to understand the algorithm but it is difficult to understand the outcome that the model predicts.

2.4 Injustice

The fourth type of explainability problem is called injustice. In this type of question, we also know everything about the algorithm and we fully understand how it works. However, we concern about if the decision made by the model is fair, just and moral. For example, if a self-driving car has to crash on one of two persons, one homeless person and one rich person,

what should the algorithm do in this case? If a self-driving car chose to kill the homeless one, this would be considered unmoral action. Thus, the problem is not that nobody has explained how the algorithm works. Instead, the problem is that it seems impossible to explain how the algorithm is consistent with law or ethics.

3 Lipton's Interpretability[2]

In this part, the lecturer transits from the definition of Felten's Explainability Problems to Lipton's Interpretability. In general, this section talks about the various desiderata of interpretability research and different properties of interpretability models. It starts from a supervised learning algorithm.

3.1 Setting

- Given labelled dataset $\mathcal{S} = \{Z_1, \ldots, Z_n\}$
- $Z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \forall 1 \leq i \leq n$
- Algorithms try to learn a mapping from the feature space $\mathcal{X}$ to the output space $\mathcal{Y}$
- We can compute scores to measure the performance of an algorithm $\mathcal{A}$ (e.g. use $err_n := \sum_{i=1}^{n} \mathbb{1}_{\{\mathcal{A}(x_i) \neq y_i\}}$)

3.2 Objectives of Interpretability Researches

In the setting described above, we can discuss the objectives of interpretability researches from five different aspect.

3.2.1 Trust

Back to the motivative example shown at the beginning, if Wenxi was advised to stay away from eating meat for three weeks, would Wenxi trust this suggestion or not? The answer might be quite subjective. Different people might have different level of trust on this results. For example, from a patient's point of view, any model which can give him the advice that experts would give can be considered as trustworthy. However, from a doctor's point of view, any model which can give him advices outside his expertise might be considered as trustworthy. Thus, it is quite vague when people say it is trustworthy. Different people might have encountered different circumstances so that the trust of a model should be viewed differently.

3.2.2 Causality

Let's take the multi-armed bandit as an example. In bandit problem, there is a causality relationship between an arm and its corresponding rewards. If a model tries to learn the casualty between the two, it might be difficult to generate a good model. It is good to point out that association does not mean causal relationships and also it is generally hard to prove causality.

3.2.3 Informativeness

One of the possible objectives could be to extract information from a trained model. Sometimes, this type of information can be valuable. For example, each branch in a decision tree could reveal some intermediate information about how a decision is made. In sparse conditional linear regression model, the segment which is learned might also contain some level of information of the model. Thus, extracted information is of the interests as well.

3.2.4 Fair and Ethical Decision-Making

This objective is aligned with Felten's injustice problem. It is unavoidable that training a model has to take fair and ethical decision into account. For example, if a model is trained for predictive policing on post-secondary education, it should be unbiased on genders. In addition, back to the example of killing the homeless person, this is also unethical, which should not be trained into a developed model. There is an interesting quote"It makes the AI ethical or unethical in the same way that large numbers of people are ethical or unethical." by James Grimmelmann.

3.2.5 Transferability

It is also an objective that a trained model should have transferability on different domains. For example, using a classifier trained on ImageNet on microorganism classificaiton should also produce a valuable and meaningful results. Another example can be adding a little noise on a cleaned image so that human cannot perceive the change. However, the model might be wrong on the prediction. For instance, adding a little white noise on a cleaned image of a panda will cause a trained classifier to make a wrong prediction.

3.3 Properties of Interpretable Models

In this part of the presentation, the presenter starts to talk about the properties of an interpretable models. The first part focuses on the question: how does the model work? The second part answers the question: what else can the model tell me? Correspondingly, they can be categorized into Transparency and Post-hoc analysis.

3.3.1 Transparencies

Transparency is to reveal the contents inside a blackbox. In genera, it can discussed from three different aspect.

3.3.1.1 Simulatability Simulatability is the property that an interpretable model can let a person repeat the procedure designed by the mode in a reasonable time to produce the correct output. It has the similar meaning as Felten's "big-picture understanding." However, this property is only available when the trained model is simple and small. Here, we can list several examples which shows the definition of small and simple.

- A deep decision tree: simple but not small
- A neural Network with 1 hidden layer: small but not simple
- LASSO: simple and small by regularizaiton
- Conditional sparse linear regression: simple and small by condition

3.3.1.2 Decomposability In Decomposability, we are less concerned about the big-picture. Rather, we focus on the components of a model such as input, parameters and calculations. For example, We might be interested in a node in a deep decision tree or we are interested in the weights of a feature in a linear regression model. An interpretable model should be capable of revealing this to a user.

However, collinearity might occur in features. For example, the first variable is supposed to record the number of years taking the pill. The second variable records the number of years of having this disease. Thus, as we can recognize, these two variables are highly correlated. A general way to achieve a meaningful decomposition is to force the mode to learn a monotonic function. For example, we can let a model capture the relationships between car milage and used car price from which the parameters might have some meaningful indication.

3.3.1.3 Algorithmic Transparency Given that we understand the learning algorithm enough, then Algorithmic Transparency can give us a theoretically bound on error rate or be proved of convergence. So far in the class, the algorithms that we learned are all having algorithmic transparency. However, on the other hand, Neural networks and deep learning models will fail in this property since many non-linear activations or non-convex loss functions are involved in this case, which makes it impossible to be bounded or proved convergence.

3.3.2 Post-hoc Interpretability

The property of Post-hoc Interpretability focuses on the information that we can extract from a trained model. This type of information might not fully show how the model works but should reveal some useful information for people. In particular, there are four common approaches which can achieve this.

3.3.2.1 Text Explanations The intuition is that when you watch a soccer game, players play the game and interpreters explain the game. Similarly, a duo model has two parts, decision-making model and interpreter model. Decision-making model will be responsible for what to play next. On the other hand, interpreter model will be served as explaining how the game is played.

3.3.2.2 Visualization We can also use visualization technique to qualitatively understand the model better. In the presentation, the presenter provides two interesting examples whose links are shown below.

- visualization of CNN features. (https://distill.pub/2017/feature-visualization/)
- fnding structures in datapoints or embeddings using t-SNE. (https://distill.pub/2016/misread-tsne/)

3.3.2.3 Local Explanations The Local Explanations can give people some chances on finding intuition of local dependecies when the decomposability is impossible to achieve in a model. For example, in Q-network of learning Ateri game, the Saliency map can reveal the intermediate states of a running model. This can render some intuition of what the model can see and how it takes what it sees into account. Another example can be from CharCNN where the random weights in the model can reveal some useful information about how the model extracts information layer by layer.

3.3.2.4 Explanation by Example A trained model can also be explained by showing some examples. For example, in a word2vec model, the relationship of Paris to France can be learned so that an example of Italy and Rome can be a demonstrative example. For example, the model can be explained in word2vec model by showing that "France - Paris + Italy = Roma". This can tell us that the model has learned the relationship between Italy and Roma same as that between France and Paris.

Remarks:

- Property of Post-hoc approaches can overlap in some cases. For example, CharRNN can be both viewed through visualization and local explanations as described above. Image captioning can be another good example. On one hand, it can be understood via the texts that it generates because those texts tell more information about detector or classifier. On the other hand, the descriptions produced by image captioning model can describe different objects in one image. For each object, we have some generated descriptions for it. This can be viewed as local explanations.
- Back to Felton's "What-if" question, given that we cannot repeat the process of a trained model in a reasonable amount of time and work, how can we ask "What-if"

question? In short, a possible solution to this question can be using the decomposability on input features, depending if the input features are meaningful. An alternative way is to use post-hoc interpretability approaches.

4 Conclusion

There are several things that we should take away about the explainability of a model.

- Linear model might not be easy to interpret.
- Using post-hoc interpretability approaches such as visualization, deep learning model can possibly be interpretable.
- We should be clear about what we are trying to solve and what approaches we are using while talking about the explainability of a model.

5 Further Dicussion

About deep learning model, there are generally two difficulties to realize the algorithmic transparency of it.

- When VC Dimension of a deep learning model is larger than the number of parameters, this is hard to know insights of the model.
- Theoretical proof of convergence is still missing in the majority of deep learning models.

References

[1] Ed Felten. What does it mean to ask for an explainable algorithm? 2017.

[2] Zachary Chase Lipton. The mythos of model interpretability. *CoRR, abs/1606.03490*, 2016.