Muhammad Hassan Nishat

Topic 2

PHL377

Paper 1

The use of machine learning algorithms and the implementation of algorithmic decisions has become more and more common within the last decade. As the predictive power of machine learning algorithms has grown, so has our use of decisions brought about by these algorithms. Although these algorithms have proven to be extremely useful, it is dangerous to accept them as perfect. Machine learning algorithms are black box algorithms, meaning that we cannot be sure as to why it came to its result. This is dangerous as if we become too reliant on these algorithms, we may be blind to any hidden biases or discrimination that come as a result of the use of black box algorithms. That is why there has been a call for regulation on using these kinds of algorithms. For example, in 2016 the European Union's General Data Protection Regulation was adopted. This regulation called for the ability to explain automated decisions. This attempt at regulating algorithmic decision making has also been seen in Canada with the proposed bill C-27, which would require organizations that use decisions found by automated systems to be able to explain why these decisions were made when requested. This does seem like a step in the right direction, however, there is still debate on what is classified as explaining why an algorithm came to its decision. I believe that to fully explain why an algorithm came to a decision, we must be able to explain the variables taken into account, the strength associated with the variables used, and why those strengths were given to those variables. To do this, I will provide a description of how machine learning algorithms work, and I will describe the difference between algorithm interpretability and explainability.

Firstly, it is important to have a basic understanding of how machine learning algorithms work. There are many different types of algorithms used so for the purpose of this paper I will use a simple backpropagation network as a reference. A machine learning algorithm is typically made up of multiple layers, an input layer, a number of hidden layers, and an output layer. Input taken from the input layer is passed on to the hidden layers, in which some processes are done, leading to the output being sent to the output layer. Layers consist of nodes, which are connected to each other within and across layers. A connection between two nodes can be assigned a weight, which will inform the path taken by the input throughout the network. Backpropagation is a technique in which the weights between nodes are altered until the desired outputs are found. Techniques such as backpropagation are what make black box algorithms as complex algorithms can have thousands of layers made up of thousands of nodes, meaning that we cannot know why the combination of weights found are what they are.

Now that I have established an idea of how machine learning algorithms work, it is important to distinguish between two ways that decision making models can be defined. Each decision-making model has a level of interpretability and explainability. Interpretability refers to the ability of determining the cause and effect that lead to a decision. Linear models, for example, tend to be highly interpretable as they are a direct relationship between at least two variables. For example, the decision to eat food can be represented as a linear model with two variables representing hunger and amount of food eaten. As hunger increases, so does the amount of food eaten. This is a highly interpretable model as we can see the cause of being hungry creates the effect of eating more food. Interpretability is important to understanding why a decision is made but is rarely seen in machine learning algorithms due to how they are trained to make their decisions. Low interpretability is not an issue in low risk scenarios such as any

music or movie recommendation software that may be build into Spotify or Netflix, it becomes an issue however in high risk scenarios such as predicting medical information. If an algorithm is trained on bias data, it could make decisions based on the wrong information. For example, if an algorithm designed to assign probability of having some disease is trained on a majority male dataset, it may assign being male as the cause of having said disease. If this algorithm has low interpretability, it would be difficult to detect this bias. Explainability is a very different concept to interpretability and can be applied to machine learning models. An explainable machine learning model takes in a black box model and attempts to construct a separate interpretable function, such as a linear model, that would approximate the machine learning model as best as it can with the given data. Now with this new function, the decision taken by the algorithm can be explained. For example, take a machine learning algorithm created to predict if a patient has a certain disease. This algorithm would be put into an explainable model that would output an interpretable, linear model. Now if a patient is misdiagnosed and asks for an explanation, the linear model can be used to explain to the patient why the decision was made. The interpretable model provided does not aim to replicate the relationship modeled by the machine learning algorithm but rather replicates the function of the algorithm. The key difference between interpretability and explainability is that an interpretable model is itself a white box model, whereas an explainable model takes in a black box model and attempts to replicate it with an interpretable white box model. An explainable model cannot be seen as a complete explanation of a machine learning algorithm as it merely provides a separate interpretable model that has the same outcome as the machine learning algorithm but does not explain the inner function of the machine learning algorithm itself.

One example of an explainable algorithm is the Local Interpretable Model-Agnostic Explanations algorithm, also known as LIME. The goal of LIME is to replicate individual predictions made by a black box model. A key factor of LIME is that it focuses on local replication rather than global replication, meaning that it can only replicate one prediction at a time. Given a machine learning model, LIME tests the model with variations of the data from the particular prediction in question and creates a new dataset including the input data used and corresponding predictions for each input. A new interpretable model such as a decision tree is trained on this newly created dataset. This new model can be used to explain the initial prediction made by the machine learning model (Babic, Cohen. 2023, p8-10).

I still believe, however, that there is still another level of understanding needed to claim that the decisions of a machine learning model can be explained. I feel that even if a replicated machine learning model is interpretable through LIME, there still may be confusion as to why the original weights in the machine learning model were set to what they were. Going back to the example of predicting if a patient has a certain disease. Even if a misdiagnosed patient was told why the decision was made based on the interpretable replication of the model, the patient may still have questions about why the weights were set that way as those weights still impact the replicated model. This is one thing that we are still unable to answer, and I feel is something that needs to be answered to say we truly understand how machine learning algorithms come to decisions. If we do not know why the weights were set based on the data the algorithm was fed, we cannot know what to change about the data to avoid the problem in the future.

Another question that arises from this discussion is about whether I believe explainability should be a legal requirement of algorithmic decisions. As I don’t believe that explainable algorithms such as LIME can give us a complete understanding of how machine learning

algorithms come to decisions, especially how they determine their weights, I don't believe that explainability should be a legal requirement in all cases. The only cases in which I believe it should be a legal requirement is in the case that black box algorithmic decision making is used in high risk situations such as medical care. In these situations I feel that patients should have a right to some level of explanation behind the decisions being made and that although methods such as LIME may not give all of the possible information, I believe they give as much as they can from replication.

In conclusion, I believe that an effective explanation of an algorithmic decision is one that is fully interpretable, meaning that the cause and effect can be determined throughout each part of the models process. For a machine learning model, this would mean we can even determine why the weights were set the way they were. This does not seem to be possible at the moment due to the nature of how machine learning models are trained through methods such as backpropagation. Explainable models such as LIME do provide thorough explanations of why a model came to a decision but still lack complete interpretability as they only simplify the original model. As a result of this, I do not believe that explainable models should be legally required unless algorithmic decisions are being used in high risk scenarios as at the moment, explainable models may be the best method of finding a cause behind the decision making of machine learning models.

Works Cited

Babic, Cohen. The Algorithmic Explainability “Bait and Switch”, 2023