Consider the trust on black box models

Kasimir Aula
kasimir.aula@helsinki.fi
Department of Computer science
University of Helsinki
Helsinki, Finland

Abstract

Opaque models are becoming more and more common in machine learning applications due to their high performance on given tasks. These so called black box models are however problematic since, while producing outputs their reasoning is not clear to humans. This report goes through the importance of acknowledging the properties of black box models and the issues of blindly trusting them.

1 Introduction

Machine learning (ML) algorithms can be used to create accurate models from real life phenomena. Due to the learning phase of the algorithms and the often high number of available data, ML models usually achieve higher accuracy rates than more traditional statistical models. While improving the model's ability to produce accurate predictions by increasing its complexity, we might face an unwanted side-effect of losing understanding either from the model, the algorithm's local functionality or the limitations of the training data. As stated in [3] these ML models are learned from a large amount of data that contain many sorts of biases. These unknown biases are then inherited by the model, which may lead to unfair or simply wrong decisions, such as in the Amazon recruiting tool where men were favoured over women, and unqualified candidates were recommended [1].

These not-well-understood models carry a name of *black box models*. The name black box indicates that there is some uncertainty in the model that is not directly shown in the predictions of the model or cannot be directly estimated with some evaluation metric. In other words, something for which interpretation is needed.

2 Interpretability

To interpret in general, loosely means to be able to explain in understandable terms. However when talking of interpretation in ML, the concept of ability to explain in understandable terms, becomes rather vague. Does it apply to features, parameters, training algorithm or to the model, and does it simply mean a thorough mechanical understanding or something else? Modeling consists of many components that should, according to the loose definition, be explainable in order to be interpretable. But as it is suggested in [4], interpretation in ML is more than a monolithic concept; it covers many different perspectives from trusting a model's decisions in general to questioning the fairness of them. To keep in line with the vague definition in [3] it is stated that an explanation is an "interface" between humans and a ML decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans.

The need to interpret models indicates that there is some information that is not directly shown, when comparing a model's predictions against true values. Thus interpretation is used to explain the behaviour of the model in general or in particular occasions, and moreover to gain trust that the model performs well in a given task. The performance isn't only limited to the accuracy of the model, as a model is often required to take into account values that we as humans consider important, such as fair and ethical decision making.

Local vs global interpretation

When explanations are sought to model behaviour, they can be dealt to two levels. A *local interpretation* aims to provide an explanation on why a particular prediction was made. This is also known as post-hoc interpretation in the literature. Post-hoc interpretations means include natural language explanations, learned lower level visualizations of representations or models, and explanations by example.

For example in air quality sensor calibration, if the model predicts a high pollutant concentration value, when such a value is never actually observed, the explanation can be searched by investigating the other features at that time. Maybe they provide an explanation such as: during the timestamp of prediction, the temperature and humidity were very high, where as both pressure and wind speed were very low, and thus the Random Forest used for calibration clearly thought that it was a similar kind of phenomenon that occurred in the past, when the pollution levels have been varying during similar environmental conditions. It is a local explanation that a human might find adequate for the situation, but it doesn't explain why the model has learned such a dependency.

When the scope of explaining observations is broadened, and the aim is to explain thoroughly the whole logic of a model and follow the entire reasoning leading to all the different possible outcomes, it is considered as *global interpretation*. A single decision tree is an example from a model, where global interpretation is possible. After fitting the model, the classification reasoning can be followed easily. On the other hand global interpretation is reduced with decision tree ensemble models, such as Random Forest or Boosting. Some

methods such as computing feature importances help to understand the overall functionality of the mentioned ensemble methods, but they do not however yield a thorough interpretability.

3 Black box models

In the previous section, interpretation as a concept was discussed. The need of an interpretable model becomes relevant, when there is an unacceptable cost of wrong predictions. This holds in many real-life applications, if they are wanted to be useful.

It's hard to find an unanimous information about the models that are considered interpretable. In [3] a list of recognized interpretable models is shown. Not surprisingly linear regression appears on that list, since the algorithmic transparency is clear and in that sense the model is globally interpretable. However some go even as far as questioning the interpretability of linear regression, through its sensitivity across high dimensional feature selection which is considered to weaken the interpretability of the model [4]. Thus black box models can be thought to be black boxes due to different reasons. However in general, linear regressions and decision trees are considered to be interpretable both locally and globally.

Opaqueness of the Learning Algorithm

The learning algorithm plays an important role on how well can a model be interpreted. Where linear models and decision trees are based on algorithms that are easy to follow, deep neural networks (DNN) or tree ensemble methods are not. For example the learning algorithm[1] of DNNs are sensitive to different initializations, they are not guaranteed to converge to the global optima, they might suffer from technical problems such as gradient exploding or vanishing problems and even after the convergence the results are not easy to interpret.

Often local interpretations might be easy to do, but at least equally often they are not. The most important features in an image or text may be found by looking at saliency maps. Sometimes in images, DNNs focus on clear objects: faces, cats, people, trees etc. and sometimes they focus on some rather irrelevant background information, which can be an indication that the algorithm is drawing lines through spurious correlation between training images [3]. The same problem of spurious correlation might follow for example the calibration of air quality sensors through too similar training data. Another downside of seeking local interpretation using saliency maps is their sensitivity to single value changes [4]. A change of a single value might change the saliency map to a very different one, thus questioning the interpretability of the decision making.

Other concerns might be related to some technical decision. For example a change of activation function from ReLU to hyperbolic-tangent can completely vary a model's ability to predict high concentration values in an air quality sensor calibration model. The reasons for this kind of behaviour are hard to explain. Although most of the learning algorithms are compiled from simple pieces, the combination of operations done in training or predicting phase makes them hard to understand. Often it's difficult for a human to say what is going to be the output of a model, when given a certain input. Hence the learning algorithms can be considered to be out of the limits of human interpretation.

Opaqueness of Other Factors

Besides understanding the learning algorithms, models possess other information that makes their reasoning unexplainable. Many of these factors might reflect the incompleteness of the problem formulation as suggested in [4] and [2]. For example, can it be confirmed that the hyper parameters of a model have been tuned properly, and what is the actual effect of tuning the hyper parameters? Or has the effect of feature or data selection been done thoroughly to ensure that the created machine learning model isn't just an inadvertently over-simplified and truncated version of the real-life problem? In the case of calibrating air quality sensors there might be important factors that are neglected in the model, such as frequent appearance of heavy source of emissions, or the effect of the direction of the wind. Respectively, the first might bias the data selected for either training or testing period, and the latter could help the model better depict the effect of wind speed. One possible approach to cover these mentioned problems is to focus on having diverse enough training and execute the evaluation in different environments.

4 Problems of using black box models

As discussed black box models are hard to interpret, and some possible challenges in them have been named in the previous chapters. However so far it hasn't been considered whether interpretation is always required or could some benefit from using black box models regardless of their little understanding? In fact it is often acknowledged that there are no risks in using a black box model, when there is no real-life cost on false predictions produced by the model [3][4][2]. Not only for black box models, but for more transparent yet highly complex models that take into account a large number of parameters, the verification step before real-life deployment is crucial to have been done exhaustively. Therefore another (yet already a subjective) option, is to trust a black box model that goes through heavy verification procedures in various real-life applications to ensure trust in the model's decision making [2].

A false prediction plays no role in a service that no-one trusts, or for example in an advertisement engine. However most of the times this is not the case. Models in general are created to extract information on situation, where computers

[1] often some version of back-propagation

can perform decision making somehow better than humans. For example the task of image recognition, where DNNs are vastly favoured, models try to recognize objects from images. Humans are not perfect in image recognition either, so it would be wrong to expect such behaviour from a model. However if a model makes errors in cases where a human would not, it questions the fact whether the task is suitable for a machine or should a human be used in the decision making. So as stated in chapter 2 we require trust towards the model to actually benefit from it since without trust, black box models can carry ethical or actual safety risks.

Despite their high accuracy in a field of computer vision, DNNs are known to be easily fooled as it is possible to alter the pixels of an image imperceptibly (for a human) so that the network misclassifies the image to another class with an extremely high accuracy. Or to input an image of specific kind of noise that makes no sense to a human, yet DNNs might recognize objects with extremely high confidence [5]. After knowing that state-of-the-art networks can be fooled with techniques invisible to a human eye, it diminishes the trustworthiness of such a system.

This is why it is emphasized that making machine learning technologies more transparent would improve the understanding of all reasoning or decisions made by the model, and hence raise the trustworthiness of them [4]. Also from the scientific point of view, the black box models need better explanations not only to have better trust and acceptance of results, but also from the fundamental perspective of scientific discovery and the progress of research.

5 Ways to explain black box models

Without going into further details, or even individually naming techniques, a high hierarchical concept called "Open the black box problems taxonomy" has been suggested to make black box models more transparent [3]. The opening of a black box model consists of two main problems: *black box explanation* and *transparent model design*. The black box explanation is further divided into three subproblems: *model explanation*, *outcome explanation* and *model inspection*. These techniques provide some abstract tools and different approaches for explaining black box models.

In model explanation problem, the intention is to find an interpretable model that mimics the unknown black box model, thus being an approximation of it. A decision tree mimicking the decisions of a neural net would be one scenario of this kind. The outcome explanation seeks to find local explanations for single predictions (corresponds to local interpretation), and to reveal the logic of the model with local-first explanations. In model inspection a textual or visual representation of the models predictions is created, while altering the model's inputs to understand some properties of the model. All three are concepts that different techniques use to provide an explanation to better understand the behaviour

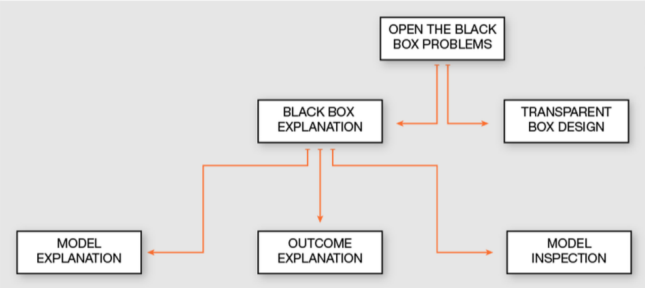

Figure 1. Open the black box problems taxonomy introduced in [3].

of a model. Transparent box design intends to directly create a model that is either locally or globally interpretable.

Conclusion

Our capability to explain the decisions behind a supervised learning algorithm, depends on the model's interpretability. The more complex the model or the the modelling environment, the harder it is to provide an explanation. When looking for explanations, global interpretability is sought, but sometimes even a local interpretation might be satisfying. However not even local interpretation of black box models is always possible, not to mention global interpretation. The urge for stronger interpretation of models implies that more careful understanding of the model's functionality is desired in order to gain more trust on their decision making. We want to trust the models that are considered today as black box models, but as there are no clear indicators of all the subjects that affect the learning process, it makes them sometimes dubious. To better trust the systems using black box models and hence better benefit from those, the capriciousness of black box models is desired to be revealed.

References

[1] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, 2018.

[2] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

[3] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, August 2018.

[4] Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, June 2018.

[5] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.