Can We Trust The Prediction of Machine Learning Models?

Cavit Çakır

May 2021

1. Introduction

The world has been changing rapidly with the innovations of science and technology. These technological changes are affecting almost every area of the economy, society, and culture. Especially inventions that relate to artificial intelligence have been changing people's lives and habits by facilitating and automating their tasks. Almost every simple task which does not require human operation is getting done by computers. However, machine learning allows computers to perform a particular task when machines are not explicitly programmed. This aspect of machine learning is useful in most industries from education to medicine. But, the actions of complex and black box machine learning models are not always precise as those of humans. Thus, automation with machine learning is not always trustable for critical applications like surgery, autonomous vehicles, and military defense systems. To trust a machine learning model, the model should be explainable and interpretable by humans.

2. Explaining the Learned Models

In 1959, Arthur Samuel coined the term “Machine Learning” as a field of study that gives computers the ability to learn without being explicitly programmed(Samuel, 1959). The term machine learning also can be defined as an application of artificial intelligence that allows computers to learn from experience. Nowadays, hardware and technological developments have come together to create an environment for developing and using machine learning applications. These developments made it possible to use machine learning to make people's work easier throughout their daily lives. People are surrounded by applications that use machine learning and are affected by the decisions made by artificial intelligence applications more and more. Lipton claims that “As machine learning models penetrate critical areas like medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic.” (Lipton, 2018) In the literature, there are simple and easy-to-understand machine learning models. When we change the inputs, we know how this action will affect the predicted outcome for those models, also we can make the reasoning for every prediction. As Lipton argues, the rapid growth of the fields where machine learning is applied has pushed the models to be more successful and therefore more complex. Advanced and complex models including complex deep neural networks are black boxes. Rudin and Radin explained the black box models as follows: “these black box models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions.”(Rudin and Radin, 2019). The complexity of black box models makes them able to predict better but also makes them very difficult to evaluate, observe, and trust. Due to the design of these models, in general, they only provide a probability of prediction. However, the output does not explain why the model did a particular prediction because there are a lot of parameters that combine and affect the prediction.

3. Interpreting the Learned Models

Furthermore, the data set is one of the most important determinants in terms of measurement scores. Generally, machine learning models are measured by accuracy and F1-Score, which are calculated by comparing the output of a learned model with the true values from the input data set. Very high measurement scores can be achieved for a specific data set, but if the data set is biased, not representing the production environment, or becomes outdated quickly, the extraordinary high scores are pointless. Thus, we cannot trust a model by examining measurement scores for specific data sets due to the possible memorization of unnecessary features and patterns. To trust the predictions of a learned model, the model should be interpretable, transparent, and understandable. Additionally, Lipton claims that “We want models to be not only good, but interpretable.”(Lipton, 2018) Because people need reliability and interpretability for several reasons. Machine learning model developers need to understand every feature and behavior of the learned model to improve the model’s predictions. The company which sells the machine learning-based application needs to be sure the product is working deterministically and the effects of wrong predictions are not critical. The people who are using machine learning-based applications need to be sure the decisions of the application are fair for everyone and can be easily explained. Lilian summarized the properties of an interpretable model by reviewing Lipton’s paper, “A human can repeat (“simulatability”) the computation process with a full understanding of the algorithm (“algorithmic transparency”) and every individual part of the model owns an intuitive explanation (“decomposability”).”(Lilian, 2017) In other words, the term interpretable means understanding the learned model through explaining the decisions made by the model, and finding the important features of the model. Simple models such as decision trees and regression algorithms are white box models. The explanation of these models is simple enough to be understood by humans. As suggested before, black box models do not aim to be interpretable. Humans cannot simulate, understand the algorithm, and predict the predictions of these models. Thus, there is only a prediction outcome to interpret the black box model. However, the uninterpretable architecture of black box models enables developers to add more parameters, which makes black box models powerful and successful.

To explain and understand the black box model, “Explainable Artificial Intelligence (XAI)” program is introduced by DARPA. The project aims to “produce more explainable models while maintaining a high level of learning performance (prediction accuracy); enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.”(Barredo Arrieta et al., 2020). This indicates that black box models will also become interpretable and reliable.

4. Conclusion

Rapidly advancing technology has enabled computers to become a part of life and bring comfort, laziness, and productivity. These factors pushed people to automate every single task by computers. Therefore, automation caused the demand for replacing humans with computers due to their efficiency and reliability, which could cause technological unemployment. To expand the capabilities of computers, artificial intelligence is developed with more sophisticated techniques, but the complexity arises with incomprehensible and unreliable systems. To rely on machine learning systems, models should be understandable. Thus, modern machine learning models are divided into two categories by their interpretability as “black box” and “white box” models. Besides the model design, the data set is also a big determining factor in the reliability of these models. Machine learning models can be used in any area if and only if the data set is relevant and the model is interpretable. Deep neural network models are still a black box, but maybe in the future, we will have perfectly interpretable complex models.

5. References

1. Samuel, A. L. “Some Studies in Machine Learning Using the Game of Checkers.” IBM Journal of Research and Development, vol. 3, no. 3, July 1959, pp. 210–229, 10.1147/rd.33.0210. Accessed 3 Apr. 2019.
2. Lipton, Zachary C. “The Mythos of Model Interpretability.” Communications of the ACM, vol. 61, no. 10, 26 Sept. 2018, pp. 36–43, 10.1145/3233231. Accessed 29 Mar. 2020.
3. Rudin, Cynthia, and Joanna Radin. “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson from an Explainable AI Competition.” 1.2, vol. 1, no. 2, 1 Nov. 2019, 10.1162/99608f92.5a8a3a3d. Accessed 12 Jan. 2020.
4. Weng, Lilian. “How to Explain the Prediction of a Machine Learning Model?” Lilianweng.github.io, 1 Aug. 2017, lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model.html#interpretable-models.
5. Barredo Arrieta, Alejandro, et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.” Information Fusion, vol. 58, June 2020, pp. 82–115, arxiv.org/pdf/1910.10045.pdf, 10.1016/j.inffus.2019.12.012.