Writing assignment: Machine learning in the real world

Interpretability and explanations

Yahui Wu
Chalmers University of Technology
MPMOB
yahuiw@chalmers.se

Tianshuo Xiao
Chalmers University of Technology
MPMOB
tianshuo@chalmers.se

In recent years, with the development of machine learning, many machine learning models can make very good predictions, but they do not explain very well how they make them, for example many deep neural networks now do not have a way to fully understand the model's decisions in a sense that they are made from a human perspective.

We know that the current model, AlphaGo, which can beat the world champion Go champion and get close to perfect scores in graphical recognition speech recognition [1]. However, we are always wary of these predictions because we do not fully understand what their predictions are based on and don't know when it will be wrong. This is why almost all models are now unable to be deployed in key performance-demanding areas such as transport, healthcare, law, finance, etc. We find that these areas are still not fully confident in the predictive power of the models.

Some of the drawbacks of poorly interpretable machine learning have created resistance to the development of autonomous driving. Dean Pomerleau was testing the self-driving feature of a Hummer military vehicle when the Hummer approached a bridge and suddenly swerved to one side. He was able to avoid the crash only by quickly grabbing the steering wheel and regaining control. Pomerleau tried to analyze the causes of this accident. Instead of storing what it learned in a neat block of digital memory, he found that the system propagated the information in a way that was extremely difficult to decipher [2]. NVIDIA was unlike autonomous cars designed by Google, Tesla, or General Motors, which teach themselves to drive by observing humans' behaviors, rather than following a single instruction provided by an engineer or programmer. This algorithm is disconcerting because it is not entirely clear how the car makes its decisions. Information from vehicle sensors goes directly into a huge network of artificial neurons, which process the data through the network and then deliver operational information to the car. But it could be very difficult for this car to find out the cause when there was a traffic accident [3].

The issue of interpretability exists in other areas as well. We can apply deep learning to a hospital database of patient records. This dataset contains hundreds of patient variables, and training this data can predict when people are suffering from what diseases. But for certain diseases, such as mental illnesses like schizophrenia, these diseases are very difficult to predict and diagnose. But with Deep Patient for prediction, it would ideally give doctors the reason for their prediction and give the right medication for treatment, but doctors do not know how they work, so there's no guarantee that the medication prescribed is reasonable. This issue is also present in the insurance field. For example, when you are asked to apply for family insurance and you are rejected, the reason

given by the insurance company is that your community is at high-risk area or that you have poor credit. But when processed through machine learning, the system collects your personal information but rejects your request. When you want to ask the reason for the rejection, while machine learning is a prediction given by a large amount of data training and it is a black box, which cannot give a specific reason [4].

Interpretability is a relatively serious issue, which can cause confusion to users. To deploy the machine learning model, we built and make sure it has practical value and we would like to make the model interpretable. The interpretability of a machine learning model is important, and it is not sufficed to only obtain a well predicted output from the model, we also want a good explanation of the outcome. Although the definition of interpretability is lack somehow, there is a growing body of literature proposes purportedly interpretable algorithms and some aspects to depict the interpretability of a model. Firstly, interpretability could be regarded as a prerequisite of trust. Does the ML model seem to be reliable for users can reflect the interpretability to some extent. However, it is hard to define trust and the definition could be subjective. Secondly, researchers would like to find the causality of a given outcome by a model rather than only finding relationships between the input and output superficially [5]. A ML model with good interpretability could let researchers infer causal relationships from observational data. Thirdly, a well-designed ML learning model should perform transferability when the environment where it was deployed is unstable. Humans will not make a completely different choice when there is a subtle change of the factor that affects their behavior but ML models will do. While the machine-learning objective might be to reduce error, the real-world purpose is to provide useful information and such an ability is called informativeness. The last point is that the ML model should make fair and ethical decisions and such a demand requires good interpretability of a model.

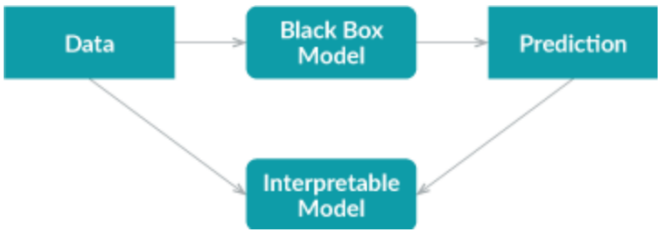
Interpretability has two techniques and model properties - transparency and post hoc interpretability. Transparency is considered at the level of the entire model (simulatability), at the level of individual components such as parameters (decomposability), and at the level of the training algorithm (algorithmic transparency). An interpretable model should be a simple model which is easy to be fully understood by users [6]. Simulatability may admit two subtypes: one based on the size of the model, and another based on the computation required to perform inference. To make a model more interpretable, each part of the model should be explained and understood by users, and it's called decomposability if a model can achieve that. The last notion of transparency is algorithmic transparency. No matter what data is given as an input, the algorithms should converge to a unique solution.

Unlike transparency requires a precise and clear explanation of a ML model, post hoc interpretability represents a distinct approach to extracting information from learned models that does not elucidate precisely how a model works. There are several methods to make a model more interpretable in terms of hoc interpretability. Besides a model to make predictions, text explanations method could use another model to generate an explanation of the prediction. Another common approach to generating post hoc interpretations is to render visualizations in the hope of determining qualitatively what a model has learned. Local explanations focus on explaining what a neural network depends on locally instead the whole mapping. Explanation by example is to find which

other examples are most similar with respect to the model [7].

Interpretability is important to make the model's behavior understandable and consistent with the way people think. There are ways to solve the interpretability problem. The Partial Dependence Plot (PDP) was invented more than a decade ago to show the marginal effect of one or two features on the predicted outcome of a machine learning model. The feature importance shows which variables have the greatest impact on the prediction, while the Partial Dependence Plot shows how the features affect the model prediction. Such models are fitted on real unmodified real data. We can change the value of a feature multiple times to produce a series of predictions, and thus interpret the prediction results [8]. A similar approach to PDP is Individual Conditional Expectation (ICE), which displays the case for each instance. ICE can help us explain how the model's predictions change when a particular feature is changed. PDP plots the average case while ICE is the case for each instance, so we can consider using them together [9].

Another method is through Global Surrogate, which approximates the predictions of the black-box model by training an interpretable model. First, we use the trained black-box model to make predictions on the dataset, and then we train the interpretable model on that dataset and predictions. The trained interpretable model approximates the original model, and all we need to do is to interpret the model [10]. However, this method also has some disadvantages. The interpretable model to approximate the black box model introduces additional errors. Besides, the proxy model is trained only based on the predictions of the black-box model and not the real results, so the global proxy model can only explain the black-box model and not the data.



A more accurate method at this stage is through Shapley Value (SHAP). We can explain the prediction by assuming that each feature value of the instance is a "player" in the game. The contribution of each player is measured by adding and removing players from all subsets of the remaining players [11]. A player's Shapley Value is the weighted sum of all of his contributions, and the Shapley Value is additive and locally accurate. If we add up the Shapley Value of all features and add the base value, the predicted average, you will get the exact predicted value, which combines the above methods.

In conclusion, many of today's deep neural networks do not have a way to fully understand the model's decisions in a way that makes sense from a human perspective. Although machine learning predictions are highly accurate, yet we always have a tinge of caution about these predictions, which is that we don't know when it will be wrong because we don't fully understand what they are predicting based on. Especially in the fields of transportation, healthcare, law, finance, etc., we find

that these fields still cannot fully trust the predictive capabilities of the models. With some solutions, we can gradually interpret the predictions made by the models.

The necessity of interpretability is that the theory behind it is already human-centered, and the response is how we can achieve human trust in the models by interpreting them, so as to create more secure and reliable applications, and thus promote the progress of the whole AI industry.

[1]. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and Tree Search. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961

[2]. Castelvecchi, D. (2016). Can we open the black box of ai? Nature, 538(7623), 20–23. https://doi.org/10.1038/538020a

[3]. Knight, W. (2020, April 2). The dark secret at the heart of ai. MIT Technology Review. Retrieved March 12, 2023, from https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

[4]. Thompson, C. (2016, October 27). Sure, A.I. is powerful-but can we make it accountable? Wired. Retrieved March 12, 2023, from https://www.wired.com/2016/10/understanding-artificial-intelligence-decisions/

[5]. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

[6]. Delcaillau, D., Ly, A., Papp, A., & Vermet, F. (2022). Model transparency and interpretability: Survey and application to the insurance industry. European Actuarial Journal, 12(2), 443–484. https://doi.org/10.1007/s13385-022-00328-y

[7]. Molnar, C., Casalicchio, G., & Bischl, B. (2020). Quantifying model complexity via functional decomposition for better post-hoc interpretability. Machine Learning and Knowledge Discovery in Databases, 193–204. https://doi.org/10.1007/978-3-030-43823-4_17

[8]. Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., & Bischl, B. (2021, September 3). Relating the partial dependence plot and permutation feature importance to the data generating process. arXiv.org. Retrieved March 12, 2023, from https://arxiv.org/abs/2109.01433

[9]. Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics, 24(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

[10]. Müller, J., Shoemaker, C. A., & Piché, R. (2013). So-mi: A surrogate model algorithm for computationally expensive nonlinear mixed-integer black-box global optimization problems. Computers & Operations Research, 40(5), 1383–1400. https://doi.org/10.1016/j.cor.2012.08.022

[11]. Ma, S., & Tourani, R. (2020, August 12). Predictive and causal implications of using Shapley value for Model Interpretation. arXiv.org. Retrieved March 12, 2023, from https://arxiv.org/abs/2008.05052