Definitions, methods, and applications in Interpretable machine learning(Review)

Tejas Gaikwad
gaikwad.2@iitj.ac.in

Indian Institute of Technology, Jodhpur,
Rajasthan, India

Abstract

The absence of a well-formed definition of interpretability, a broad range of methods with a correspondingly broad range of outputs (e.g., visualizations, natural language, mathematical equations) have been labeled as interpretation. This has led confusions in the actual meaning of interpretation and explainability of the Machine Learning(ML) or Deep Learning(DL) models. This work has tried to provide a standard and more generalised perception of explainability and interpretability. A standard which can be followed to consider an entity as interpetable or explainable. For this, the authors have proposed a Predictive, Descriptive and Relevancy(PDR) framework that can be followed to evaluate the methods as Interpretable/ Explainable or not. By doing so, they tried to provide a common vocabulary for researchers and practitioners to use in evaluating and selecting interpretation methods.

1 Introduction

In the context of machine learning and artificial intelligence, explainability and interpretability are often used interchangeably. It is the extent to which one is able to predict what is going to happen, given a change in input or algorithmic parameters. It is like being able to look at an algorithm and can see what's happening here. Explainability, meanwhile, is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. It is basically the ability to literally explain what is happening inside a model. framework that consists of 3 necessities for evaluating and constructing interpretation. Using these terms, they have categorised a wide range of existing methods.
Interpretability is a quickly growing field in machine learning, and there have been multiple works examining various aspects of interpretations (sometimes under the heading, explainable AI). interpretability is a major topic when considering bias and fairness in ML models. There are 2 related areas that are distinct but closely related to interpretability and those are causal inference and stability. Causal inference meth- ods focus solely on extracting causal relationships from data, i.e., statements that altering one variable will cause a change in another. Whether or not these relationships are causal cannot be verified through interpretable ML techniques, as they are not designed to distinguish between causal and noncausal effects.Stability, as a generalization of robustness in statistics, is a concept that applies throughout the entire data–science life cycle, including interpretable ML. The stability principle requires that each step in the life cycle is stable with respect to appropri- ate perturbations, such as small changes in the model or data.

2 Summary

The nature of the problem plays a role in interpretability, as the relevant context and audience are essential in determining the methods to use. Choosing a domain problem helps to collect the dataset to study it as it can effect the interpretation pipeline. Based on the chosen problem and collected data, the practitioner then can constructs a predictive model. The data is then processed, cleaned, and visualised which is further followed by extracting the features, model selection and fitting the dataset into the model. Interpretability considerations often come into play in this step related to the choice between simpler, easier to interpret models and more complex, black-box models, which may fit the data better. The model's ability to fit the data is measured through predictive accuracy and Post Hoc Analysis. Having fitted a model (or models), the practitioner then analyzes it for answers to the original question. The process of analyzing the model often involves using interpretability methods to extract various (stable) forms of information from the model. The extracted information can then be analyzed and displayed using standard data analysis methods, such as scatter plots and histograms. The ability of the interpretations to properly describe what the model has learned is denoted by descriptive accuracy.
Interpretability can be bifurcated into 2 sub-domains as Model based and post-hoc based. Interpretability in the modeling stage stated as model-based interpretability. This part of interpretability is focused upon the construction of models that readily provide insight into the relationships they have learned. Model-based interpretability is best used when the underlying relationship is sufficiently simple that model-based techniques can achieve reasonable predictive accuracy or when predictive accuracy is not a concern. In contrast to model-based interpretability, which alters the model to allow for interpretation, post hoc interpretation methods take a trained model as input and extract information about what relationships the model has learned. They are most helpful when the data are especially complex, and practitioners need to train a black-box model to achieve reasonable predictive accuracy. The important requirements for PDR framework are stated as Accuracy and Relevancy. There are 2 areas where errors can arise: when approximating the underlying data relationships with a model (predictive accuracy) and when approximating what the model has learned using an interpretation method (descriptive accuracy). For an interpretation to be trustworthy, one should try to maximize both of the accuracy's. The accuracy's can further be divided into predictive and descriptive accuracy. For predictive accuracy first source of error occurs during the model stage, when an ML model is constructed. If the model learns a poor approximation of the underlying relationships in the data, any information extracted from the model is unlikely to be accurate. For descriptive accuracy the source of error occurs during the post hoc analysis stage, when interpretation methods are used to analyze a fitted model. The interpretation is said to be relevant if it provides insight for a particular audience into a chosen domain problem. Relevancy often plays a key role in determining the trade-off between predictive and descriptive accuracy. Depending on the context of the problem at hand, a practitioner may choose to focus on one over the other. For instance, when interpretability is used to audit a model's predictions, such as to enforce fair- ness, descriptive accuracy can be more important. In contrast, interpretability can also be used solely as a tool to increase the predictive accuracy of a model, for instance, through improved feature engineering.

Model-Based Interpretability
The practitioner constructs an ML model from the collected data. We define model-based interpretability as the construction of models that readily provide insight into the relation-

ships they have learned. Different model-based inter- pretability methods provide different ways of increasing descriptive accuracy by constructing models which are easier to understand, sometimes resulting in lower predictive accuracy. The main challenge of model-based interpretability is to come up with models that are simple enough to be easily understood by the audience, while maintaining high predictive accuracy. In selecting a model to solve a domain problem, the practitioner must consider the entirety of the PDR framework. The first desideratum to consider is predictive accuracy. If the constructed model does not accurately represent the underlying problem, any subsequent analysis will be suspect. Second, the main purpose of model-based interpretation methods is to increase descriptive accuracy. Finally, the relevancy of a model's output must be considered and is determined by the context of the problem, data, and audience. Type of Model-Based Interpretability are Sparsity, Simultability, Modularity, Domain based feature engineering and Model-Based Feature Engineering. Model sparsity is often useful for high-dimensional problems, where the goal is to identify key fea- tures for further analysis. For instance, sparsity penalties have been incorporated into random forests to identify a sparse subset of important features. A model is said to be simulatable if a human (for whom the interpretation is intended) is able to internally simulate and reason about its entire decision-making process (i.e., how a trained model produces an output for an arbitrary input). This is a very strong constraint to place on a model and can generally be done only when the number of features is low and the underlying relationship is simple. Decision trees are often cited as a simulatable model, due to their hierarchical decision-making process. We define an ML model to be modular if a meaningful portion(s) of its prediction-making process can be interpreted independently. Probabilistic models can enforce modularity by specifying a conditional independence structure which makes it easier to reason about different parts of a model independently. While the type of model is important in producing a useful interpretation, so are the features that are used as inputs to the model. Having more informative features makes the relationship that needs to be learned by the model simpler, allowing one to use other model based interpretability methods. Moreover, when the features have more meaning to a particular audience, they become easier to interpret. In many individual domains, expert knowledge can be useful in constructing feature sets that are useful for building predictive models. The particular algorithms used to extract features are generally domain specific, relying both on the practitioner's existing domain expertise and on insights drawn from the data through exploratory data analysis. There are a variety of automatic approaches for constructing interpretable features. Two examples are unsupervised learning and dimensionality reduc- tion. Unsupervised methods, such as clustering, matrix factorization, and dictionary learning, aim to process unlabeled data and output a description of their structure. These structures often shed insight into relationships contained within the data and can be useful in building predictive models. Dimensional- ity reduction focuses on finding a representation of the data which is lower dimensional than the original data.

Post-HOC Interpretability

At this stage, the practitioner analyzes a trained model to provide insights into the learned relationships. This is particularly challenging when the model's parameters do not clearly show what relationships the model has learned. To aid in this process, a variety of post hoc interpretability methods have been developed to provide insight into what a trained model has learned, without changing the underlying model. These methods are particularly important for settings where the collected data are high dimensional and complex, such as with image data.Interpretation methods must deal with the challenge that individual features are

not semantically meaningful, making the problem more challenging than on datasets with more meaningful features. Once the information has been extracted from the fitted model, it can be analyzed using standard, exploratory data analysis techniques, such as scatter plots and histograms. When conducting post hoc analysis, the model has already been trained, so its predictive accuracy is fixed. Thus, under the PDR framework, a researcher must consider only descrip- tive accuracy and relevancy (relative to a particular audi- ence). Improving on each of these criteria are areas of active research. Most widely useful post hoc interpretation methods fall into 2 main categories: prediction-level and dataset-level interpretations, which are sometimes referred to as local and global interpretations, respectively. Prediction-level interpretation methods focus on explaining individual predictions made by models, such as what features and/or interactions led to the particular prediction. Dataset-level approaches focus on the global relationships the model has learned, such as what visual patterns are associated with a predicted response. For Prediction-Level Interpretation Prediction-level approaches are A. Measuring Interpretation Desiderata. Currently, there is no useful when a practitioner is interested in understanding how individual predictions are made by a model.

3 Problems with PDR Framework

Measuring and quantifying the descriptive accuracy is a challenging task also no standard evaluation protocol for this. Model based interpretability also fails to achieve higher predictive accuracy because of which it becomes a necessity to abandon the model based interpretability. The tools required for feature engineering are needed to be developed as they can provide the information of features which are needed to be choosen.