Interpretable Maching Learning

Chapter 01 Interpretability

Hao ZHAN

haozhan1993@gmail.com

2020.5

Table of Contents

1. Improtance of Interpretability
2. Taxonomy of Interpretability Methods
3. Scope of Interpretability

Table of Contents

1 Improtance of Interpretability

2 Taxonomy of Interpretability Methods

3 Scope of Interpretability

1.Improtance of Interpretability

What is the interpretability?

Miller (2017): Interpretability is the degree to which a human can understand the cause of a decision.

Kim et.al. (2016): Interpretability is the degree to which a human can consistently predict the model's result.

A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model.

1.Improtance of Interpretability

Why do not we just trust the model and ignore why it made a certain decision?

Doshi-Velez Kim (2017): The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.

- What is predicted?
- Why the prediction was made?

The model must also explain how it came to the prediction (the **why**), because a correct prediction only partially solves the **original problem**.

1.Improtance of Interpretability

The following reasons drive the demand for interpretability and explanations (Doshi-Velez and Kim 2017 and Miller 2017).

1.Human curiosity and learning

Humans have a mental model of their environment that is updated when something unexpected happens.

2.Find meaning in the world

The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior.

1.Improtance of Interpretability

The following reasons drive the demand for interpretability and explanations (Doshi-Velez and Kim 2017 and Miller 2017).

3.Goal of science

The model itself becomes the source of knowledge instead of the data. Interpretability makes it possible to extract this additional knowledge captured by the model.

4.Safety measures

Machine learning models take on real-world tasks that require safety measures and testing.

1.Improtance of Interpretability

The following reasons drive the demand for interpretability and explanations (Doshi-Velez and Kim 2017 and Miller 2017).

5.Detecting bias

Interpretability is a useful debugging tool for detecting bias in machine learning models. It might happen that the machine learning model you have trained for automatic approval or rejection of credit applications discriminates against a minority.

6.Social acceptance

The process of integrating machines and algorithms into our daily lives requires interpretability to increase social acceptance.

1.Improtance of Interpretability

The following reasons drive the demand for interpretability and explanations (Doshi-Velez and Kim 2017 and Miller 2017).

7.Manage social interactions

By creating a shared meaning of something, the explainer influences the actions, emotions and beliefs of the recipient of the explanation.

8.Debugged and audited

Machine learning models can only be debugged and audited when they can be interpreted.

1.Improtance of Interpretability

The following reasons drive the demand for interpretability and explanations:

- Human curiosity and learning
- Find meaning in the world
- Goal of science
- Safety measures
- Detecting bias
- Social acceptance
- Manage social interactions
- Debugged and audited

1.Improtance of Interpretability

If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

Fairness

Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups.

Privacy

Ensuring that sensitive information in the data is protected.

1.Improtance of Interpretability

If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

Reliability or Robustness

Ensuring that small changes in the input do not lead to large changes in the prediction. (ill-posedness)

Causality

Check that only causal relationships are picked up.

Trust

It is easier for humans to trust a system that explains its decisions compared to a black box.

1.Improtance of Interpretability

When we do not need interpretability:

- Interpretability is not required if the model has no significant impact.
- Interpretability is not required when the problem is well studied.
- **Interpretability might enable people or programs to manipulate the system.**

Table of Contents

1 Improtance of Interpretability

2 Taxonomy of Interpretability Methods

3 Scope of Interpretability

2.Taxonomy of Interpretability Methods

Methods for machine learning interpretability can be classified according to various criteria.

(1) Intrinsic or post hoc?

Intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models.

Post hoc interpretability refers to the application of interpretation methods after model training. Permutation feature importance is, for example, a post hoc interpretation method.

2.Taxonomy of Interpretability Methods

(2) Result of the interpretation method

Feature summary statistic: Many interpretation methods provide summary statistics for each feature.

Featuresummaryvisualization: Most of the feature summary statistics can also be visualized.

Model internals (e.g. learned weights): The interpretation of intrinsically interpretable models falls into this category.

Data point: This category includes all methods that return data points to make a model interpretable.

Intrinsically interpretable model: One solution to interpreting black box models is to approximate them with an interpretable model.

2.Taxonomy of Interpretability Methods

(3) Model-specific or model-agnostic?

Model-specific interpretation tools are limited to specific model classes.

Model-agnostic tools can be used on any machine learning model and are applied after the model has been trained (post hoc).

(4) Local or global?

Does the interpretation method explain an individual prediction or the entire model behavior? Or is the scope somewhere in between?

Table of Contents

1 Improtance of Interpretability

2 Taxonomy of Interpretability Methods

3 Scope of Interpretability

3.Scope of Interpretability

An algorithm trains a model that produces the predictions. Each step can be evaluated in terms of transparency or interpretability.

(1) Algorithm Transparency: How does the algorithm create the model?

Algorithm transparency is about how the algorithm learns a model from the data and what kind of relationships it can learn. If you use convolutional neural networks to classify images, you can explain that the algorithm learns edge detectors and filters on the lowest layers.

- Least squares method for linear models are well studied and understood. They are characterized by a high transparency.

- Deep learning approaches (pushing a gradient through a network with millions of weights) are less well understood and the inner workings are the focus of ongoing research. They are considered less transparent.

3.Scope of Interpretability

(2) Global, Holistic Model Interpretability: How does the trained model make predictions?

We could describe a model as interpretable if you can comprehend the entire model at once (Lipton, 2016).

To explain the global model output, you need the trained model, knowledge of the algorithm and the data. This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures.

Global model interpretability is very difficult to achieve in practice.

3.Scope of Interpretability

(3) Global Model Interpretability on a Modular Level: How do parts of the model affect predictions?

While global model interpretability is usually out of reach, there is a good chance of understanding at least some models on a modular level.

The interpretation of a single weight always comes with the footnote that the other input features remain at the same value, which is not the case with many real applications.

Example: Boston House Price

3.Scope of Interpretability

(4) Local Interpretability for a Single Prediction: Why did the model make a certain prediction for an instance?

We can zoom in on a single instance and examine what the model predicts for this input, and explain why.

(5) Local Interpretability for a Group of Predictions: Why did the model make specific predictions for a group of instances?

Model predictions for multiple instances can be explained either with global model interpretation methods (on a modular level) or with explanations of individual instances.

Refrence

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269.

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, Learn to Criticize! Criticism for Interpretability." Advances in Neural Information Processing Systems. 2016.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning," no. MI: 1–13. http://arxiv.org/abs/1702.08608.

Lipton, Zachary C. "The Mythos of Model Interpretability." arXiv preprint arXiv:1606.03490, 2016.

Molnar Christoph. Interpretable machine learning[M]. Lulu. com, 2019.

Thank you for your time!