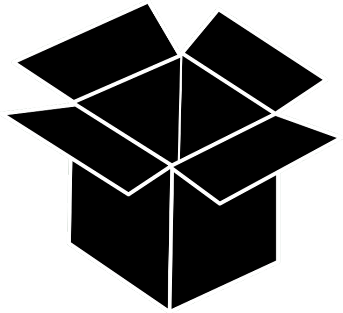
Interpretable Machine Learning

Fundamental Terms and Concepts



Learning goals

- What is interpretable machine learning (IML)
- What is Explainable Artificial Intelligence (XAI) and how does it differ from IML?
- What is the purpose of IML?
- What are the fundamental terms and concepts of IML

WHAT IS INTERPRETABLE MACHINE LEARNING

- Machine learning (ML) algorithmically trains predictive models with no or little pre-specifications or assumptions.
- Several algorithms such as decision tree learning create interpretable models. However, most algorithms create models which can be considered a black box.
- We use the term black box, although the internal workings of the model are in fact accessible, but too complex for the human mind to comprehend.
- Interpretable machine learning (IML) is an umbrella term for all models and methods that allow for some kind of interpretation.

WHAT IS EXPLAINABLE AI

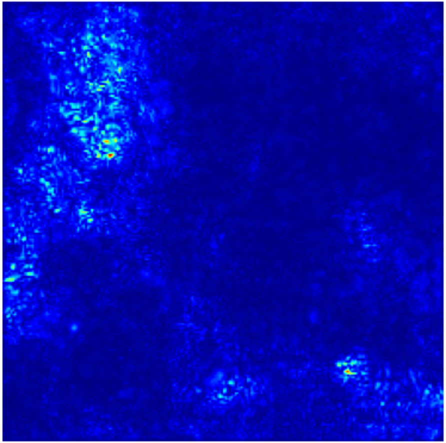
- IML is often used synonymously with Explainable AI (XAI), as ML is often used synonymously with AI. There is no unified standard for these terminologies. We find that XAI often is specifically concerned with the interpretation of neural networks, whereas IML is used as an encompassing term for everything related to model interpretability, i.e., interpretable models such as generalized additive models, model-agnostic techniques, as well as interpretations of neural networks.
- The nature of neural networks allows for powerful model-specific interpretation techniques, e.g., layer-wise relevance propagation (LRP) and saliency maps. They have in common that influence on the output layer is backpropagated layer by layer through the entire network up to the input layer.

XAI - SALIENCY MAPS

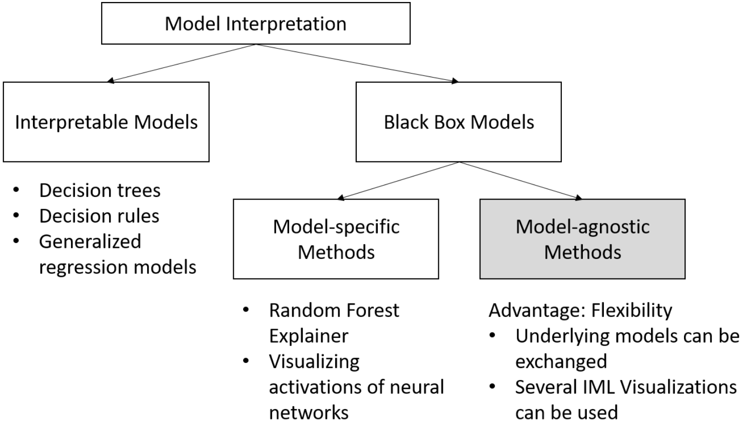
- The traditional assumption is that deep learning is best suited for non-tabular data, whereas other ML techniques such as gradient boosting or random forests are better suited for tabular data. However, recent developments seem to undermine this assumption, i.e., neural networks increasingly outperform other methods on tabular data as well.
- For visual data, i.e., pixels being represented as a matrix, deep learning has proven to deliver remarkable results. The default way to interpret neural networks on visual data is a saliency map. A saliency map is essentially a heatmap indicating pixel influence on the prediction (e.g., a classification of an image) through

XAI - SALIENCY MAPS

backpropagation:



WHAT TOOLS DO WE HAVE?



Model Interpretation

Interpretable Models

- Decision trees
- Decision rules
- Generalized regression models

Black Box Models

Model-specific Methods

- Random Forest Explainer
- Visualizing activations of neural networks

Model-agnostic Methods

Advantage: Flexibility

- Underlying models can be exchanged
- Several IML Visualizations can be used

⇒ We will focus on model-agnostic interpretability!

INTRINSIC AND MODEL-AGNOSTIC INTERPRETATION

- Intrinsically interpretable models:
  - Examples are linear models and decision trees.
  - They are interpretable because of their simple structures, e.g., a weighted combination of feature values or a tree structure.
  - However, they are difficult to interpret with many features or complex interaction terms.

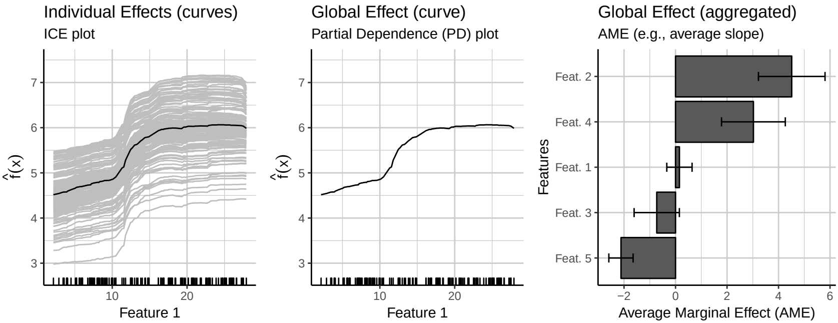
- Model-agnostic interpretation methods:
  - They are applied after training (post-hoc).
  - They also work for more complex black box models.
  - They can also be applied to intrinsically interpretable models, e.g. feature importance for decision trees.

MODEL-AGNOSTIC INTERPRETABILITY

- Model-agnostic interpretability methods work for **any** kind of machine learning model.
- Explanation type is not tied to the underlying model type.
- Often, only access to data and fitted predictor is required. No further knowledge about the model itself is necessary.
- We usually distinguish between **feature effect** and **feature importance** methods.

FEATURE EFFECTS VS. FEATURE IMPORTANCE

Feature effects indicate the direction and magnitude of a change in predicted outcome due to changes in feature values.

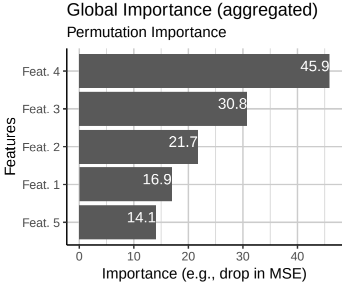


- Methods include: Partial Dependence Plots, Individual Conditional Expectation, Accumulated Local Effects (ALE)
- Pendant in linear models: Regression coefficient $\hat{\theta}_j$

FEATURE EFFECTS VS. FEATURE IMPORTANCE

Feature importance methods rank features by how much they contribute to the predictive performance or prediction variance of the model.

- Methods include: Permutation Feature Importance, Functional Anova
- Analog in linear models: Absolute t-statistic $\left|\frac{\hat{\theta}_j}{SE(\hat{\theta}_j)}\right|$



GLOBAL AND LOCAL INTERPRETABILITY

Global interpretability methods explain the expected model behavior for the entire feature space by considering all available observations (or representative subsets). For example:

- Permutation Feature Importance
- Partial Dependence Plot
- Functional Anova
- ...

Local interpretability methods explain single predictions or a group of similar observations. For example:

- Individual Conditional Expectation (ICE) Plots
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley Values
- ...

FIXED MODEL VS. REFITS

- Most methods presented in this lecture analyze a fixed, trained model (e.g., permutation feature importance).
- Some methods require refitting the model (e.g., PIMP).
- Trained model ⇒ Model is the object of analysis.
- Refitting ⇒ Learning process is the object of analysis.
- The advantage of refitting is that it includes information about the variability in the learning process.