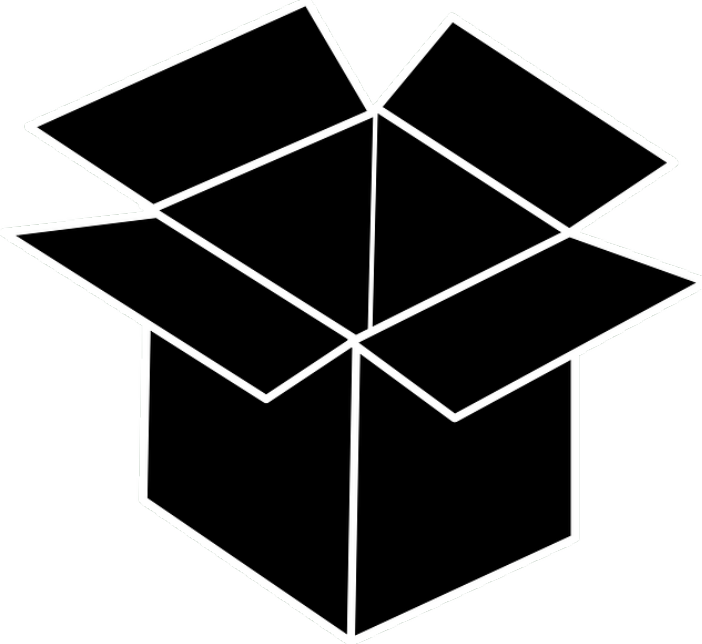
Interpretable Machine Learning

Interpretable Models



Learning goals

- Interpretable Models

INTERPRETABLE MODELS

Linear models (LM) and generalized linear models (GLM): The specification of model parameters makes LMs and GLMs intrinsically interpretable.

- **LM:**

$$\mathbb{E}_Y(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

- **GLM:**

$$g\left(\mathbb{E}_Y(y|x)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Generalized additive models (GAM): Add flexibility by replacing linear term by more general functional form, but retain interpretability by keeping the pre-specified additive predictor.

$$g\left(\mathbb{E}_Y(y|x)\right) = \beta_0 + \beta_1 h(x_1) + \cdots + \beta_p h(x_p)$$

INTERPRETABLE MODELS

Model-based boosting:

- Idea: Combine boosting with interpretable base learners (e.g., linear model with single parameter).
- Consider two linear base learners $b_j(x, \Theta)$ and $b_j(x, \Theta^\star)$ with the same type, but distinct parameter vectors Θ and $\Theta^\star$. They can be combined in a base learner of the same type:

$$b_j(x, \Theta) + b_j(x, \Theta^\star) = b_j(x, \Theta + \Theta^\star)$$

- We create a selection of interpretable base learners. In each iteration, all base learners are trained on the so-called pseudo residuals, and the one with the best fit is added to the previously computed model.
- The final model has an additive structure (equivalent to a GAM), where each component function is itself interpretable.

INTERPRETABLE MODELS

Rule-based ML:
Decision rules follow a general structure: IF the conditions are met THEN make a certain prediction. A single decision rule or a combination of several rules can be used to make predictions.

There are many ways to learn rules from data:

- OneR learns rules from a single feature. OneR is characterized by its simplicity, interpretability and its use as a benchmark.
- Sequential covering is a general procedure that iteratively learns rules and removes the data points that are covered by the new rule. This procedure is used by many rule learning algorithms.
- Bayesian Rule Lists combine pre-mined frequent patterns into a decision list using Bayesian statistics. Using pre-mined patterns is a common approach used by many rule learning algorithms.