Interpretable ML

Surrogate Models

Global Surrogate

General idea: Approximate "black box" model with a simpler model

1. Get predictions of black box model for a dataset (e.g., the training data)
2. Select an interpretable model type
    1. linear model, decision tree, ...
3. Train the interpretable model with the black box predictions as the outcome
4. Check performance of the surrogate model
5. Interpret the surrogate model

→ Use, e.g., a single tree to "summarize" a random forests decisions

Local Surrogate

LIME – Local interpretable model-agnostic explanations (Ribeiro et al. 2016)

- Focus on explaining individual predictions
- Assumption: Complex model is linear/ simple on a local scale
- Intuition: Fit a locally optimal model g given proximity measure π_x and complexity $\Omega(g)$

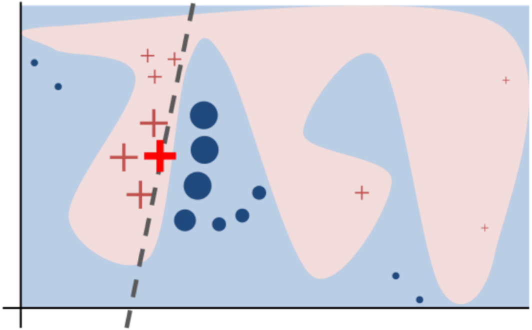
$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

1. Select instance of interest and permute observation n times
2. Predict the outcome of permuted observations with the complex model
3. Weight the new observations according to their proximity to the instance of interest
4. Train a weighted, interpretable model on the permuted data
5. Interpret the local model

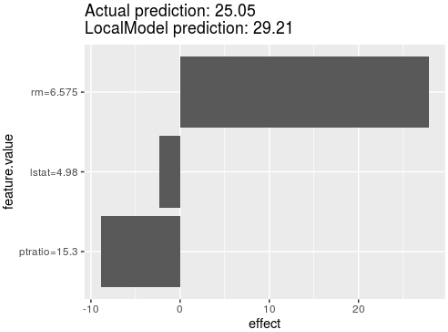
Local Surrogate

Figure: Local interpretable model-agnostic explanations (LIME)

(a) Intuition



(b) Example



More interpretable ML

- Shapley values and SHAP
  - https://link.springer.com/article/10.1007/s10115-013-0679-x
  - https://arxiv.org/abs/1705.07874
- Feature interaction (H-statistic)
  - https://arxiv.org/abs/0811.1679
- Partial dependence-based variable importance
  - https://arxiv.org/pdf/1805.04755.pdf
- Representative trees from ensembles
  - https://www.ncbi.nlm.nih.gov/pubmed/22302520