LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Machine Learning

CART 2

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

Department of Statistics – LMU Munich

MONOTONE FEATURE TRANSFORMATIONS

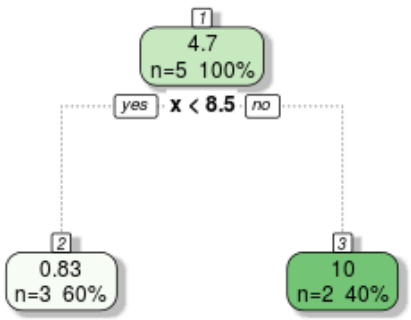
Monotone transformations of one or several features will not change the value of the impurity measure, neither the structure of the tree (just the numerical value of the split point).

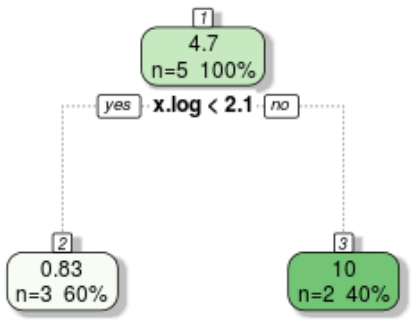
Original data

x	1	2	7.0	10	20
y	1	1	0.5	10	11

Data with log-transformed x

log(x)	0	0.7	1.9	2.3	3
y	1	1.0	0.5	10.0	11





CART: CATEGORICAL PREDICTORS

- A categorical split partitions the feature levels:

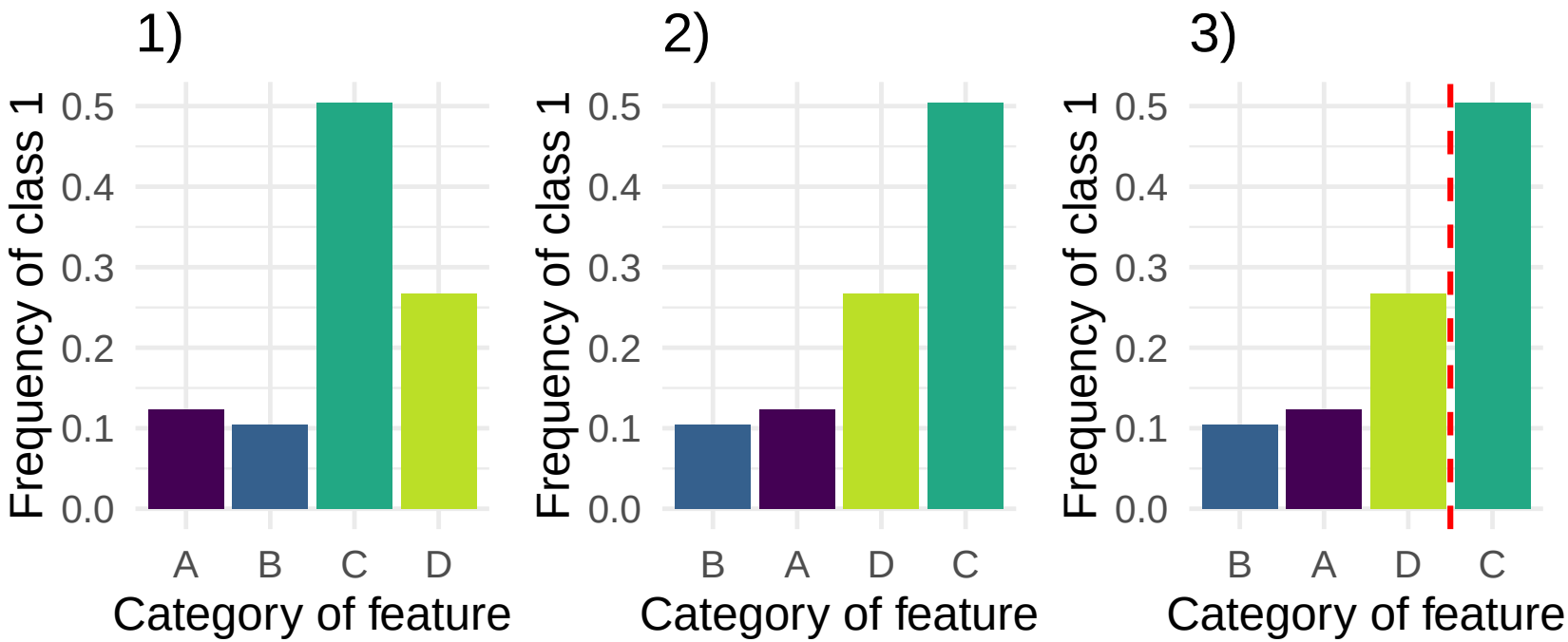
$$x_j \in \{a, c, e\} \leftarrow node \rightarrow x_j \in \{b, d\}$$

- For a categorical feature with m categories, there are ca. 2^m different possible partitions of the m values into two groups ($2^{m-1} - 1$ exactly, considering symmetry and that we do not split off empty groups)
- Searching over these becomes prohibitive for larger values of m
- But for regression with squared loss and binary classification shortcuts exist.

CART: CATEGORICAL PREDICTORS

For $0 - 1$ responses, in each node:

1. Calculate the proportion of 1-outcomes for each category of the feature.
2. Sort the categories according to these proportions.
3. The feature can then be treated as if ordered



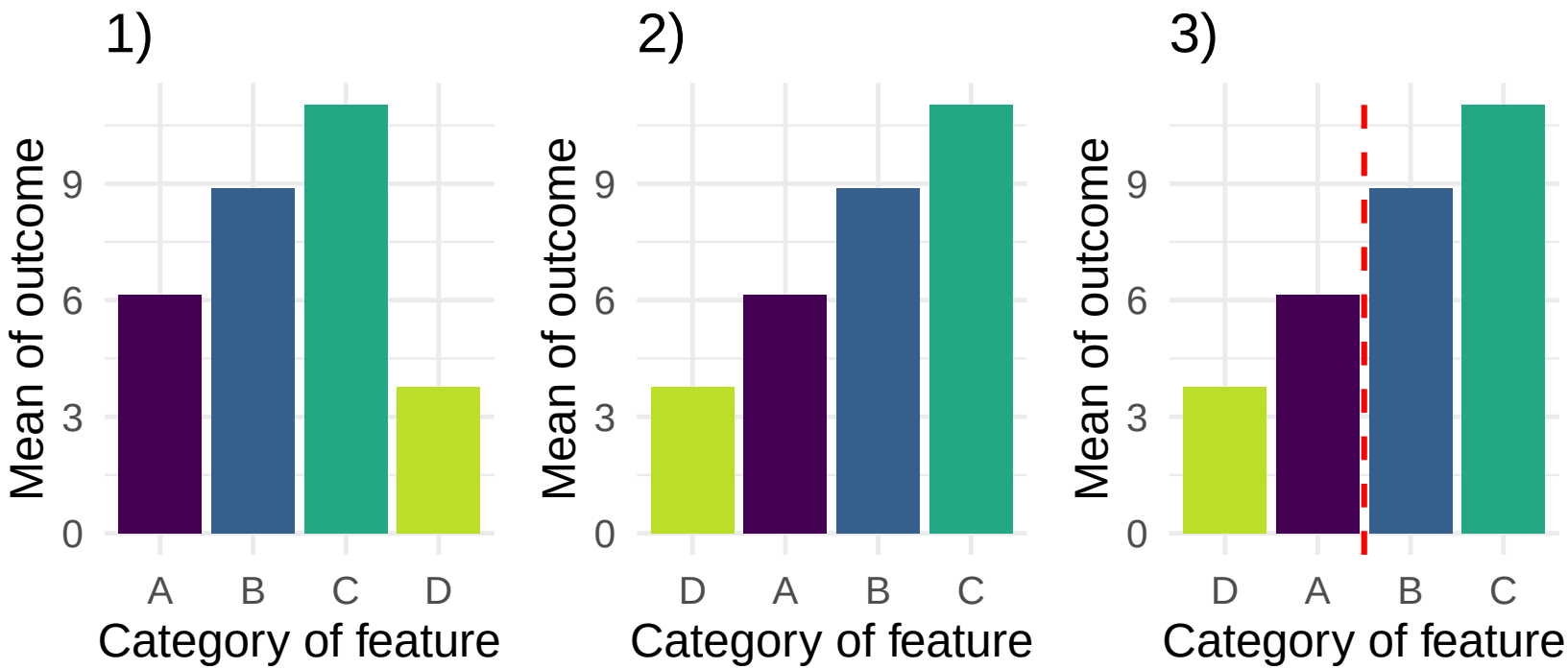
CART: CATEGORICAL PREDICTORS

- This procedure obtains the optimal split
- This result also holds for regression trees (with squared error loss) – the categories are ordered by increasing mean of the outcome
- The proofs are not trivial and can be found here:
  - for 0-1 responses:
    - Breiman, 1984, Classification and Regression Trees.
    - Ripley, 1996, Pattern Recognition and Neural Networks.
  - for continuous responses:
    - Fisher, 1958, On grouping for maximum homogeneity.
- Such simplifications are not known for multiclass problems.

CART: CATEGORICAL PREDICTORS

For continuous responses, in each node:

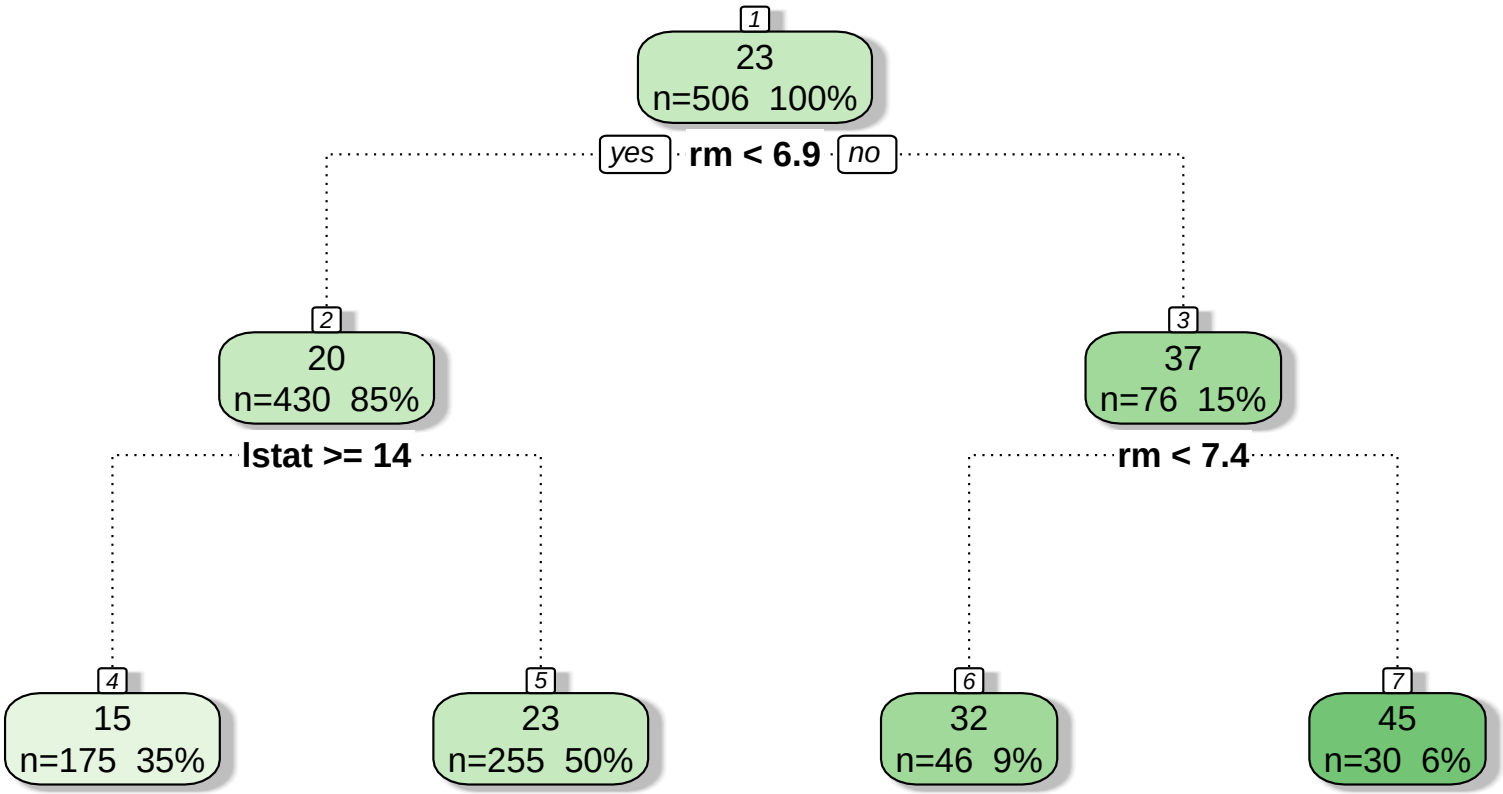
1. Calculate the mean of the outcome in each category
2. Sort the categories by increasing mean of the outcome



PRUNING

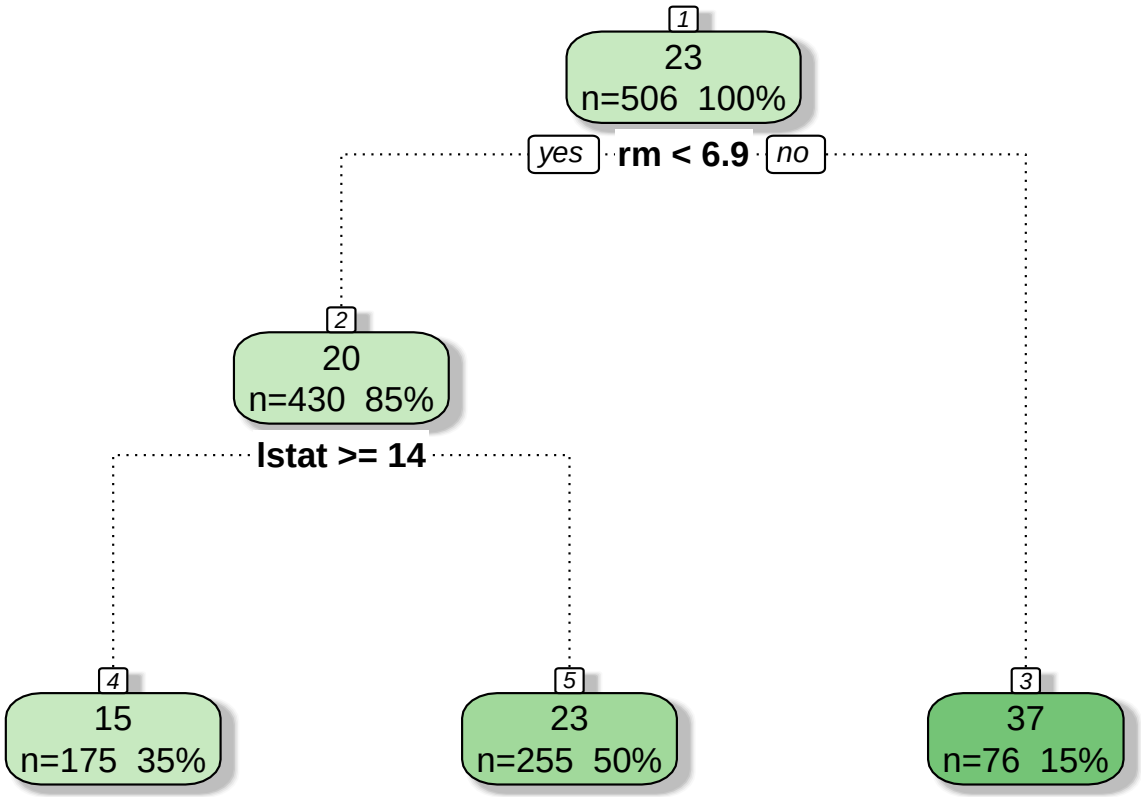
- Method to select optimal size of the tree, to avoid overfitting
- During tree growing, it is hard to tell when we should stop, as we don't know if the addition of further splits down the line will dramatically decrease error (called horizon effect).
- Aggressive (early) stopping criteria can be used, sometimes this is called pre-pruning, and you would have to tune over their settings
- Or post-pruning: Grow a deep tree, then remove parts, so that the resulting smaller tree is optimal w.r.t. cross-validated error

PRUNING



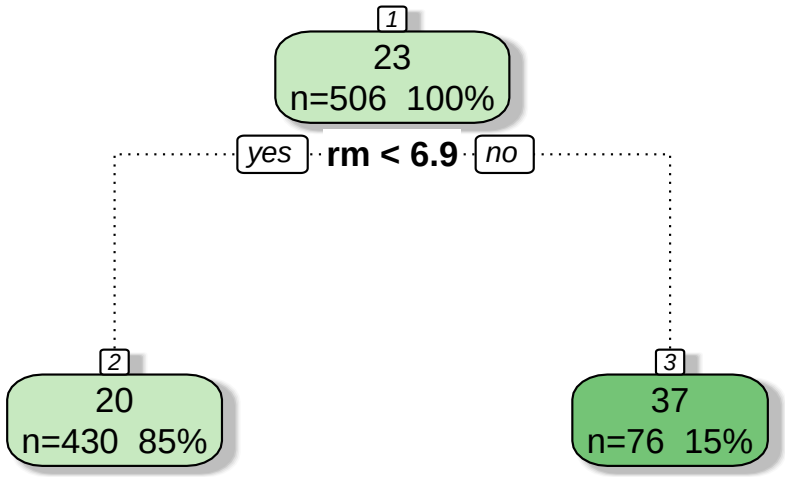
Pruning with complexity parameter = 0.0100.

PRUNING



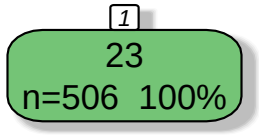
Pruning with complexity parameter = 0.0717.

PRUNING



Pruning with complexity parameter = 0.1712.

PRUNING



Pruning with complexity parameter = 0.4527.

CART: MISSING FEATURE VALUES

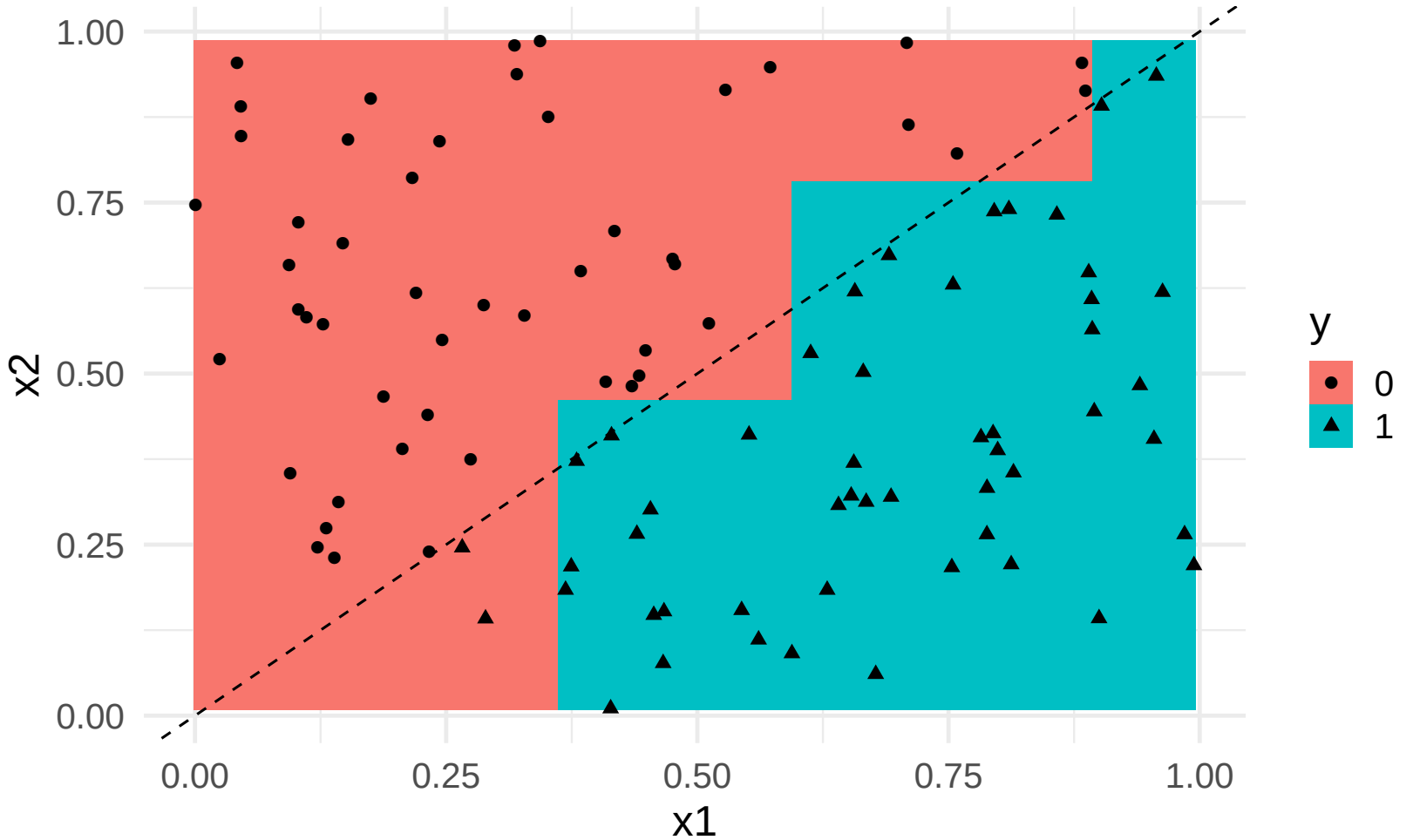
Two approaches:

1. CART often uses the so-called surrogate split principle to natively deal with missing values in features
2. When splits are evaluated, only observations for which the feature is not missing are used
3. When observations are passed down the tree (in training or prediction), and the feature value needed at a split is missing, we use a "surrogate split" to send the data down. These surrogate splits are created during training and resemble as close as possible the actual splitting rule.

ADVANTAGES

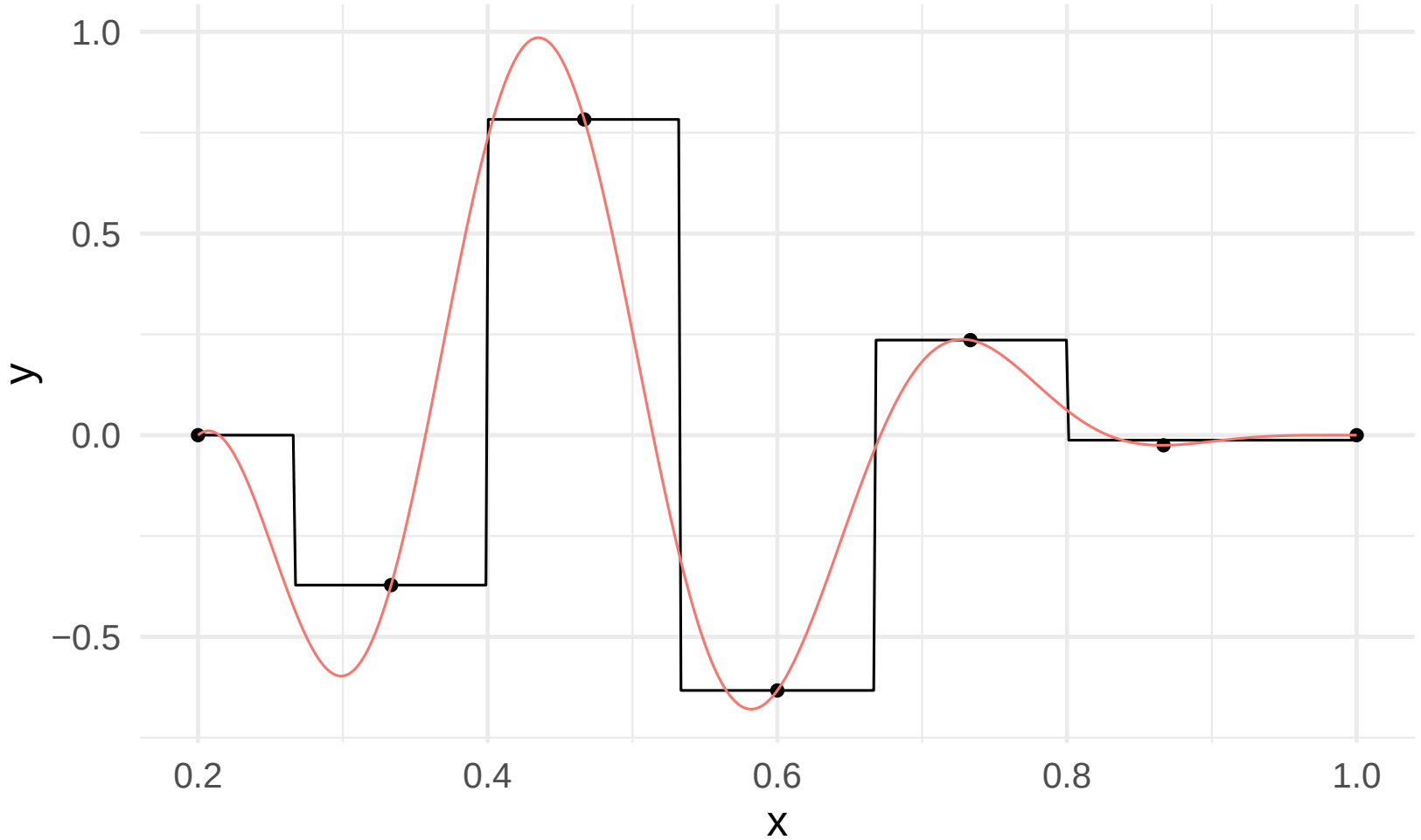
- Model is easy to comprehend, graphical representation
- Not much preprocessing required: a) native handling of categorical features and missing values b) noproblems with outliers in features c) Monotone transformations of features change nothing, so no feature scaling
- Interaction effects between features are easily possible
- Works for (some) non-linear functions, but see "disadvantages"
- Inherent feature selection
- Quite fast, scales well with larger data
- Trees are flexible by creating a custom split criterion and leaf-node prediction rule (clustering trees, semi-supervised trees, density estimation, etc.)

DISADVANTAGE: LINEAR DEPENDENCIES



Linear dependencies must be modeled over several splits. Logistic regression would model this easily.

DISADVANTAGE: SMOOTH FUNCTIONS



Prediction function isn't smooth because a step function is fitted.

DISADVANTAGES

- Really not the best predictor: Combine with bagging (forest) or boosting!
- High instability (variance) of the trees. Small changes in the data could lead to very different trees. This leads to a) less trust in interpretability b) is a reason why prediction error of trees is usually not best.
- In regression, due to fitting piecewise constant models, trees often do not extrapolate well

FURTHER TREE METHODOLOGIES

- AID (Sonquist and Morgan, 1964)
- CHAID (Kass, 1980)
- CART (Breiman et al., 1984)
- C4.5 (Quinlan, 1993)
- Unbiased Recursive Partitioning (Hothorn et al., 2006)