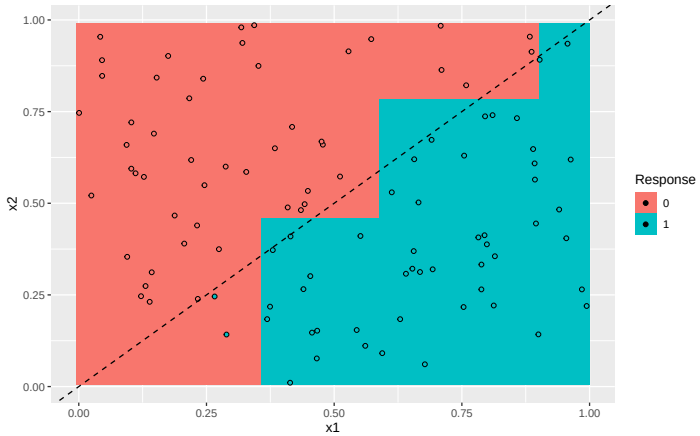
Introduction to Machine Learning

CART: Advantages & Disadvantages



Learning goals

- Understand advantages and disadvantages of CART
- Know how CART can be expressed in terms of hypothesis space, risk and optimization

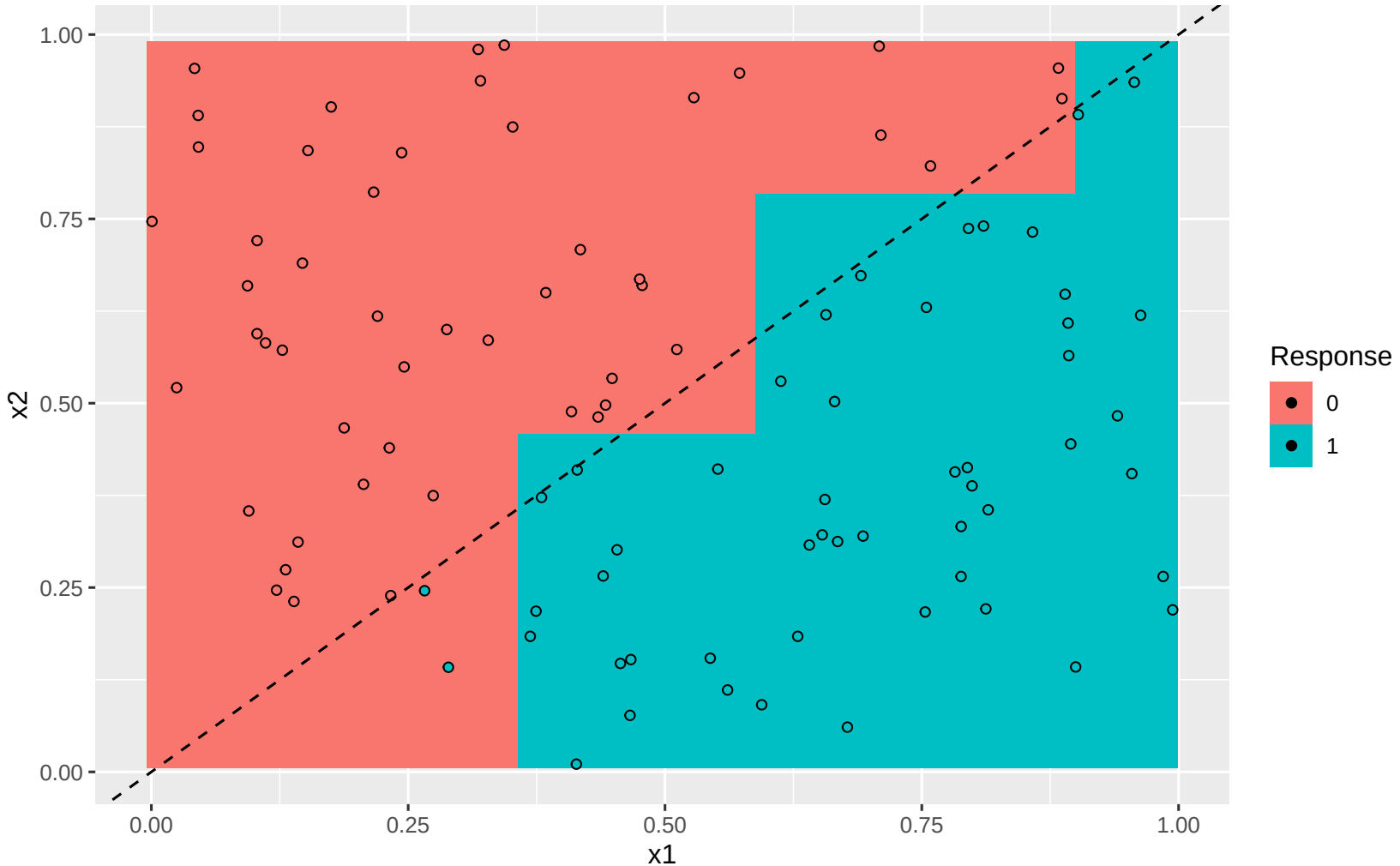
ADVANTAGES

- Fairly easy to understand, interpret and visualize.
- Not much preprocessing required:
  - automatic handling of non-numerical features
  - automatic handling of missing values via surrogate splits
  - no problems with outliers in features
  - monotone transformations of features change nothing so scaling of features is irrelevant
- Interaction effects between features are easily possible, even of higher orders
- Can model discontinuities and non-linearities (but see "disadvantages")

ADVANTAGES

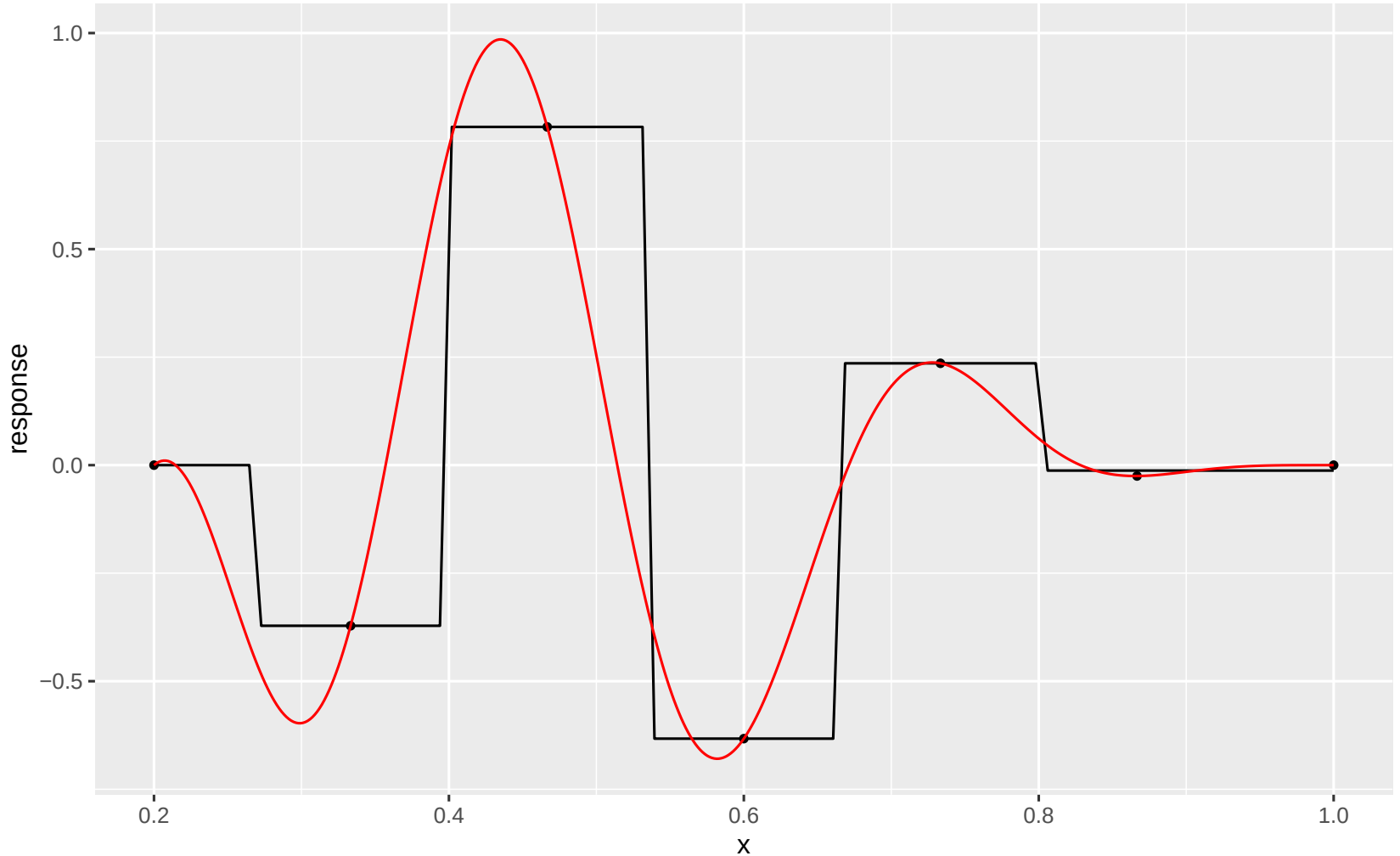
- Performs automatic feature selection
- Quite fast, scales well with larger data
- Flexibility through definition of custom split criteria or leaf-node prediction rules: clustering trees, semi-supervised trees, density estimation, etc.

DISADVANTAGE: LINEAR DEPENDENCIES



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

DISADVANTAGE: SMOOTH FUNCTIONS



Prediction functions of trees are never smooth as they are always step functions.

DISADVANTAGES

- Empirically not the best predictor: Combine with bagging (forest) or boosting!
- High instability (variance) of the trees. Small changes in the training data can lead to completely different trees. This leads to reduced trust in interpretation and is a reason why prediction errors of trees are usually not the best.
- In regression: Trees define piecewise constant functions, so 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)

CART: SYNOPSIS

Hypothesis Space:
CART models are step functions over a rectangular partition of $\mathcal{X}$.
Their maximal complexity is controlled by the stopping criteria and the pruning method.

Risk:
Trees can use any kind of loss function for regression or classification.

Optimization:
Exhaustive search over all possible splits in each node to minimize the empirical risk in the child nodes.
Most literature on CARTs based on "impurity reduction" which is mathematically equivalent to empirical risk minimization:
Gini impurity $\cong$ Brier Score loss,
entropy impurity $\cong$ Bernoulli loss,
variance impurity $\cong$ L2 loss.