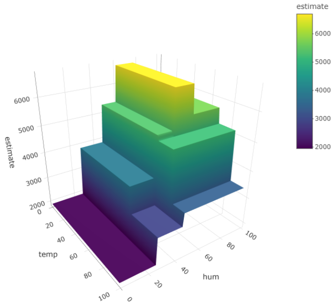
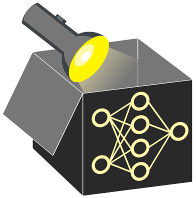
Interpretable Machine Learning

Rule-based Models

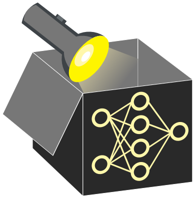


Learning goals

- Decision trees
- RuleFit
- Decision rules



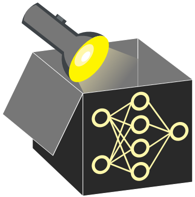
DECISION TREES ▸ Breiman et al. (1984)



Idea of decision trees: Partition data into subsets based on cut-off values in features (found by minimizing a split criterion via greedy search) and predict constant mean c_m in leaf node $\mathcal{R}_m$:

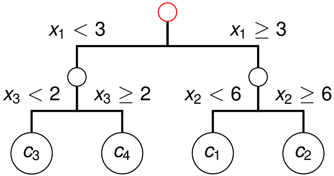
$$\hat{f}(x) = \sum_{m=1}^{M} c_m \mathbb{1}_{\{x \in \mathcal{R}_m\}}$$

DECISION TREES ▸ Breiman et al. (1984)

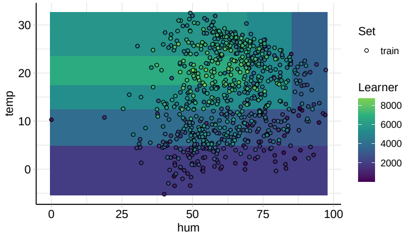


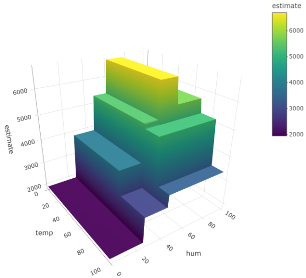
Idea of decision trees: Partition data into subsets based on cut-off values in features (found by minimizing a split criterion via greedy search) and predict constant mean c_m in leaf node $\mathcal{R}_m$:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \mathbb{1}_{\{x \in \mathcal{R}_m\}}$$

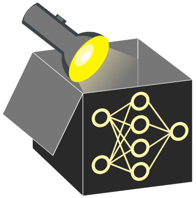


- Applicable to regression and classification
- Able to model interactions and non-linear effects
- Able to handle mixed feature spaces and missing values





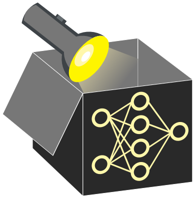
INTERPRETATION



- Directly by following the tree structure (i.e., sequence of decision rules)
- Importance of x_j: Aggregate "improvement in split criterion" over all splits where x_j was involved

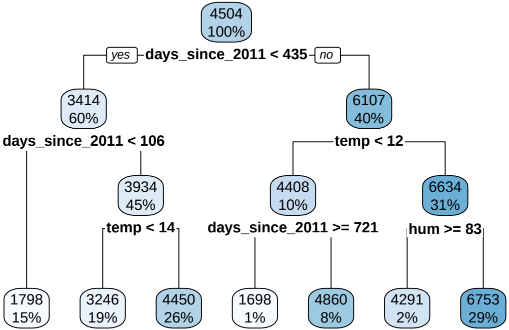
  $\leadsto$ e.g., variance for regression or Gini index for classification

DECISION TREES - EXAMPLE



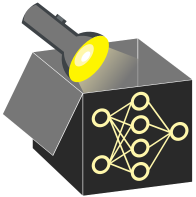
- Fit decision tree with tree depth of 3 on bike data
- E.g., mean prediction for the first 105 days since 2011 is 1798
  $\rightsquigarrow$ Applies to $\hat{=}15\%$ of the data (leftmost branch)
- $\mathrm{days\_since\_2011}$: highest feature importance (explains most of variance)

Feature	Importance
days_since_2011	79.53
temp	17.55
hum	2.92



UNBIASED RECURSIVE PARTITIONING

▸ Hothorn et al. (2006) ▸ Zeileis et al. (2008) ▸ Strobl et al. (2007)

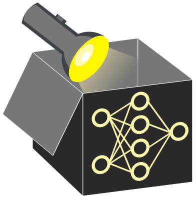


Problems with CART (Classification and Regression Trees):

1. Selection bias towards high-cardinal/continuous features
2. Does not consider significant improvements when splitting ($\rightsquigarrow$ overfitting)

UNBIASED RECURSIVE PARTITIONING

▸ Hothorn et al. (2006) ▸ Zeileis et al. (2008) ▸ Strobl et al. (2007)



Problems with CART (Classification and Regression Trees):

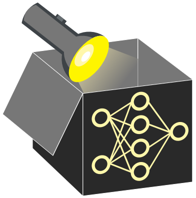
1. Selection bias towards high-cardinal/continuous features
2. Does not consider significant improvements when splitting ($\rightsquigarrow$ overfitting)

Unbiased recursive partitioning via conditional inference trees (ctree) or model-based recursive partitioning (mob):

1. Separate selection of **feature used for splitting** and **split point**
2. Hypothesis test as stopping criteria

UNBIASED RECURSIVE PARTITIONING

▸ Hothorn et al. (2006) ▸ Zeileis et al. (2008) ▸ Strobl et al. (2007)



Problems with CART (Classification and Regression Trees):

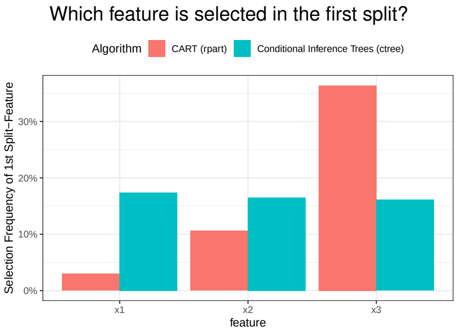
1. Selection bias towards high-cardinal/continuous features
2. Does not consider significant improvements when splitting ($\rightsquigarrow$ overfitting)

Unbiased recursive partitioning via conditional inference trees (ctree) or model-based recursive partitioning (mob):

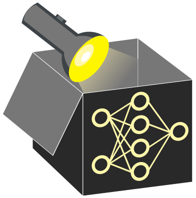
1. Separate selection of **feature used for splitting** and **split point**
2. Hypothesis test as stopping criteria

Example (selection bias):
Simulate data ($n = 200$) with $Y \sim N(0, 1)$ and 3 features of different cardinality independent from Y (repeat 500 times):

- $X_1 \sim Binom(n, \frac{1}{2})$
- $X_2 \sim M(n, (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$
- $X_3 \sim M(n, (\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}))$



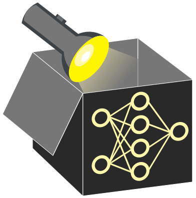
UNBIASED RECURSIVE PARTITIONING



Differences to CART:

- Two-step approach (1. find most significant split feature, 2. find best split point)
- Parametric model (e.g. LM instead of constant) can be fitted in leave nodes
- Significance of split (p-value) given in each node
- ctree and mob differ in hypothesis test used for selecting the split feature (independence test vs. fluctuation test) and how to find the best split point

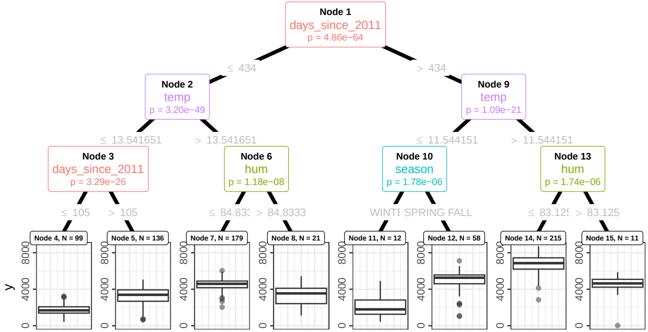
UNBIASED RECURSIVE PARTITIONING



Differences to CART:

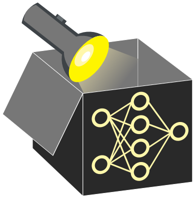
- Two-step approach (1. find most significant split feature, 2. find best split point)
- Parametric model (e.g. LM instead of constant) can be fitted in leave nodes
- Significance of split (p-value) given in each node
- ctree and mob differ in hypothesis test used for selecting the split feature (independence test vs. fluctuation test) and how to find the best split point

Example (ctree): Bike data (constant model in final nodes)



Train error (MSE):
758,844.0 (ctree)
742,244.4 (mob)

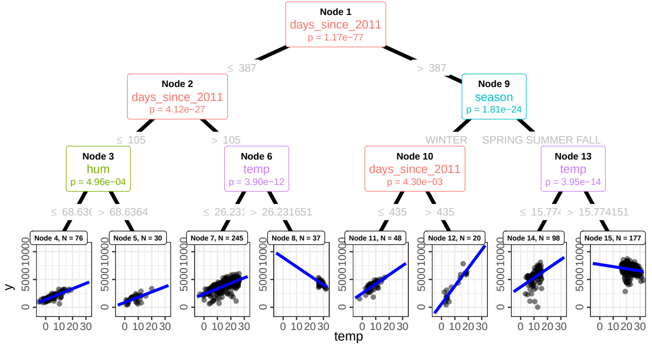
UNBIASED RECURSIVE PARTITIONING



Differences to CART:

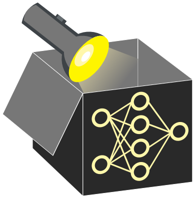
- Two-step approach (1. find most significant split feature, 2. find best split point)
- Parametric model (e.g. LM instead of constant) can be fitted in leave nodes
- Significance of split (p-value) given in each node
- ctree and mob differ in hypothesis test used for selecting the split feature (independence test vs. fluctuation test) and how to find the best split point

Example (mob): Bike data (linear model with temp in final nodes)



Train error (MSE):
758,844.0 (ctree)
742,244.4 (mob)

OTHER RULE-BASED MODELS

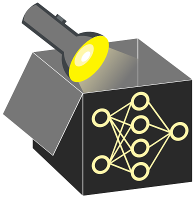


Decision Rules ▸ Holte 1993

- (Chaining of) simple “if – then” statements
  $\rightsquigarrow$ very intuitive and easy-to-interpret
- Most methods work only for classification and categorical features

IF `size=small` THEN `value=low`
IF `size=medium` THEN `value=medium`
IF `size=big` THEN `value=high`

OTHER RULE-BASED MODELS



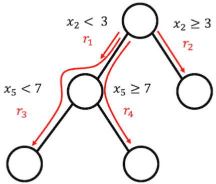
Decision Rules ▸ Holte 1993

- (Chaining of) simple "if – then" statements
  $\rightsquigarrow$ very intuitive and easy-to-interpret
- Most methods work only for classification and categorical features

IF `size=small` THEN `value=low`
IF `size=medium` THEN `value=medium`
IF `size=big` THEN `value=high`

RuleFit ▸ Friedman and Popescu 2008

- Combination of LM and decision trees
- Uses (many) decision trees to extract important decision rules r_1, r_2, r_3, r_4 which are used as features in a (regularized) LM
- Allows for feature interactions and non-linearities



▸ Molnar 2022