Recap Lecture 11

Trees and Forests

ISLR8

Jilles Vreeken
Krikamol Muandet

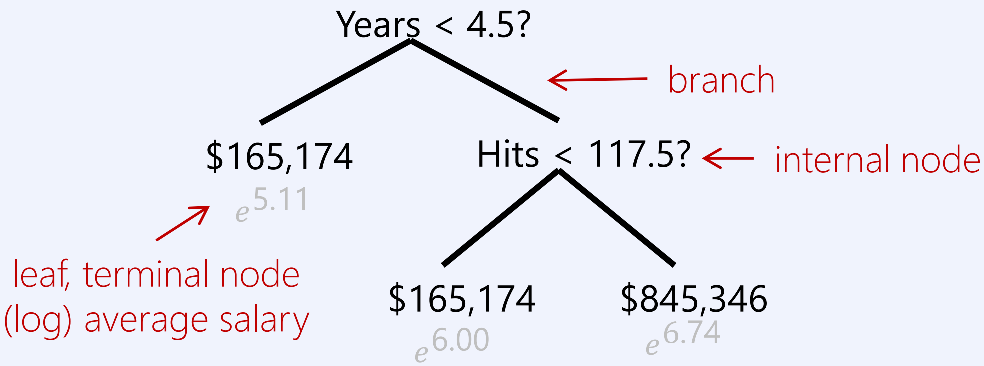
UNIVERSITÄT DES SAARLANDES

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Advantages and Disadvantages of Trees

- ➕ Trees are easy to explain to people
- ➕ Trees arguably mimic human decision-making
- ➕ Trees have a simple graphical representation and are easy to interpret, especially, if they are small
- ➕ Trees can handle qualitative predictors without the need of dummy variables
- ➕ Trees allow for systematically imputing missing values

- ➖ Trees are often not as accurate as the other models
- ➖ Trees can be very non-robust, i.e. performance can change dramatically upon small changes in the data

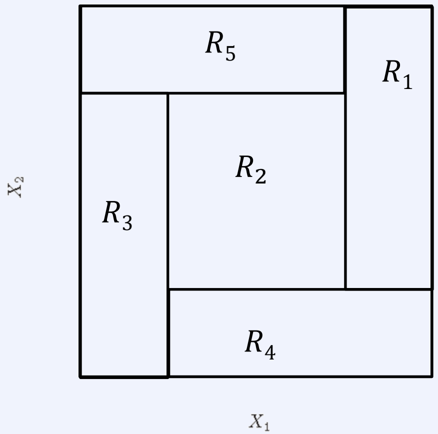


How to Build a Regression Tree

Two simple steps

1. divide predictor space (data space) into J disjoint regions $R_1, \dots, R_J$
2. build a constant model within each region – the mean value of all points in the region

In theory regions can have any shape in step 1, we will only use rectangles (cuboids)



Pruning a Tree

The criterion is formed in analogy to the lasso procedure from Ch. 6

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T|$$

penalty parameter α

subtree $T \subset T_0$ of the full tree T_0

- $|T|$ is the number of leaves of T
- $2|T| - 1$ the number of nodes in T
- α controls the tradeoff between fit and complexity
  - $\alpha = 0$ selects the full tree
  - as α increases, the tree gets smaller

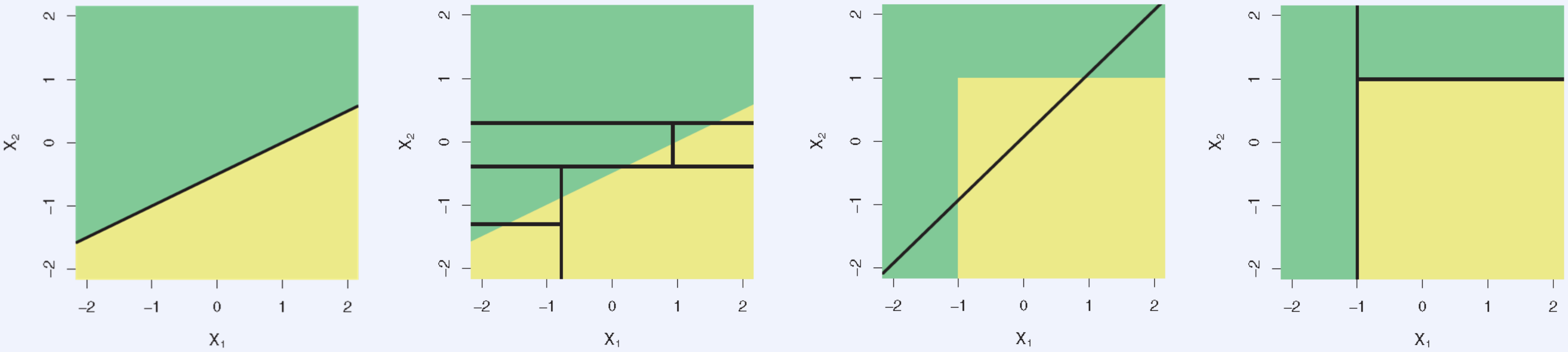
Trees vs. Linear Regression

Linear regression $f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, the world is globally linear

Trees $f(X) = \beta_0 + \sum_{m=1}^{M} c_m I(X \in R_m)$, the world is regionally constant

Which model is more suitable depends on the problem, example 2D binary classification

linear data, linear model | linear data, tree model | regional data, linear model | regional data, tree model



Ensemble Methods based on Trees

Ensemble methods calculate several models for a dataset and merge their predictions

- getting several predictions can reduce variance

Bagging

Apply the bootstrap method (Ch 6) to tree models to reduce the variance

1. generate B training datasets using the bootstrap
2. build a tree on each dataset affording the response $\hat{f}^{*b}(x)$
3. average over the response of all trees for the final prediction $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$

For regression predict the average, for classification predict the majority vote of all trees

Boosting

Boosting is a powerful ensemble technique

- here trees are not calculated independently but in sequence

Model Parameters of Boosting

- Number of trees B
- Shrinkage parameter λ
- Number of splits d per tree controls complexity

Summary

Trees: decompose the space into regions and fit a constant model in each region

- optimal tree is hard, so we reclusively split the data, greedily selecting the current best predictor

Bagging: apply the bootstrap method to tree models to reduce the variance

- because bootstrap samples have a large overlap, bagged trees are highly correlated

Random forests: apply a trick on top of bagging to decorrelate the trees

- randomly sample out of $m < p$ predictors at each split

Boosting: slowly improve the model in areas in which it does not perform well

- in each iteration fit a small and/or shrunken tree on the residuals