Introduction to Machine Learning

Random Forests: Advantages and Disadvantages

compstat-lmu.github.io/lecture_i2ml

RANDOM FOREST: ADVANTAGES

- All advantages of trees also apply to RF: not much preprocessing required, missing value handling, etc.
- Easy to parallelize
- Often works well (enough)
- Integrated variable importance
- Integrated estimation of generalization performance via OOB error
- Works well on high-dimensional data
- Works well on data with irrelevant “noise” variables
- Often not much tuning necessary

RANDOM FOREST: DISADVANTAGES

- Often sub-optimal for regression
- Same extrapolation problem as for trees
- Harder to interpret than trees (but many extra tools are nowadays available for interpreting RFs)
- Implementation can be memory-hungry
- Prediction is computationally demanding for large ensembles

RANDOM FOREST: SYNOPSIS

Hypothesis Space:
Random forest models are (sums of) step functions over rectangular partitions of (subspaces of) $\mathcal{X}$.
Their maximal complexity is controlled by the number of trees in the random forest ensemble and the stopping criteria for the constituent trees.

Risk:
Like trees, random forests can use any kind of loss function for regression or classification.

Optimization:
Exhaustive search over all (randomly selected!) candidate splits in each node of each tree to minimize the empirical risk in the child nodes.
Like all bagging methods, optimization can be done in parallel over the ensemble members.