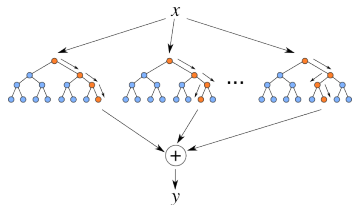


Introduction to Machine Learning

Random Forest: Introduction



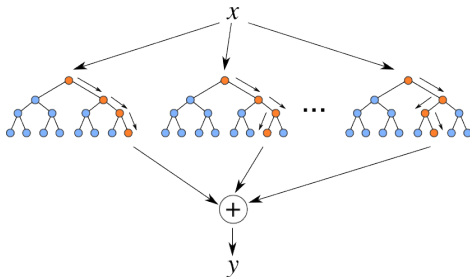
Learning goals

- Know how random forests are defined by extending the idea of bagging
- Understand that the goal is to decorrelate the trees
- Understand that the out-of-bag error is a way to obtain unbiased estimates of the generalization error during training

RANDOM FORESTS

Modification of bagging for trees proposed by Breiman (2001):

- Tree base learners on bootstrap samples of the data
- Uses **decorrelated** trees by randomizing splits (see below)
- Tree base learners are usually fully expanded, without aggressive early stopping or pruning, to **increase variance of the ensemble**



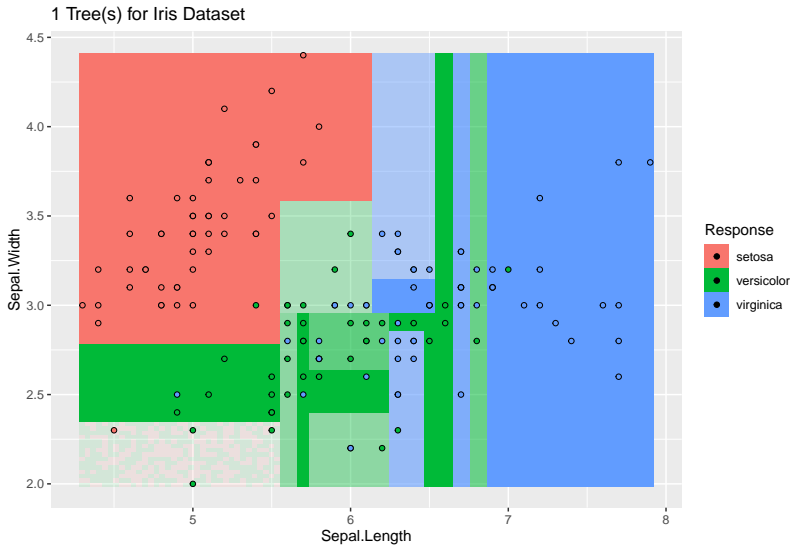
RANDOM FEATURE SAMPLING

- From our analysis of bagging risk we can see that decorrelating trees improves the ensemble
- Simple randomized approach:

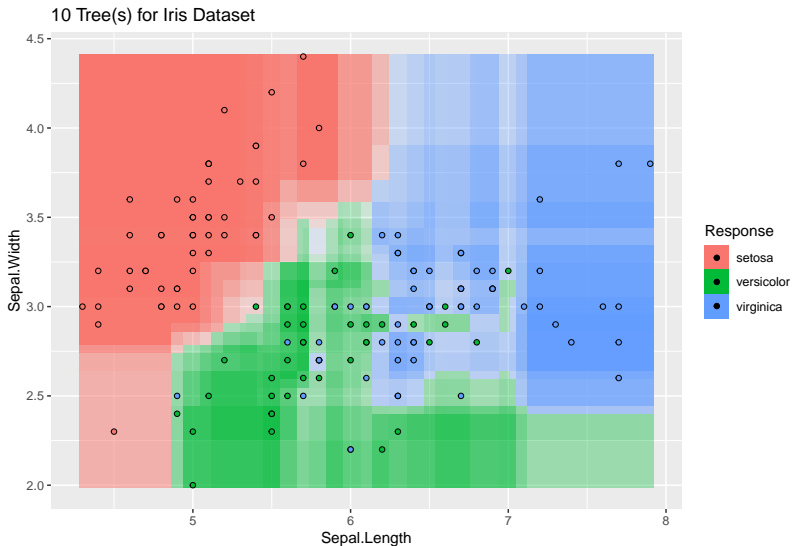
At each node of each tree, randomly draw $m_{\text{try}} \leq p$ candidate features to consider for splitting. Recommended values:

- Classification: $m_{\text{try}} = \lfloor \sqrt{p} \rfloor$
- Regression: $m_{\text{try}} = \lfloor p/3 \rfloor$

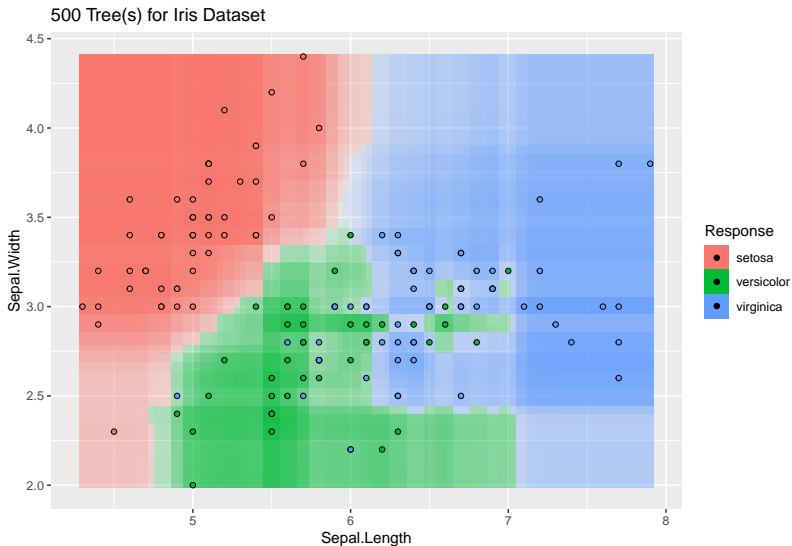
EFFECT OF ENSEMBLE SIZE



EFFECT OF ENSEMBLE SIZE

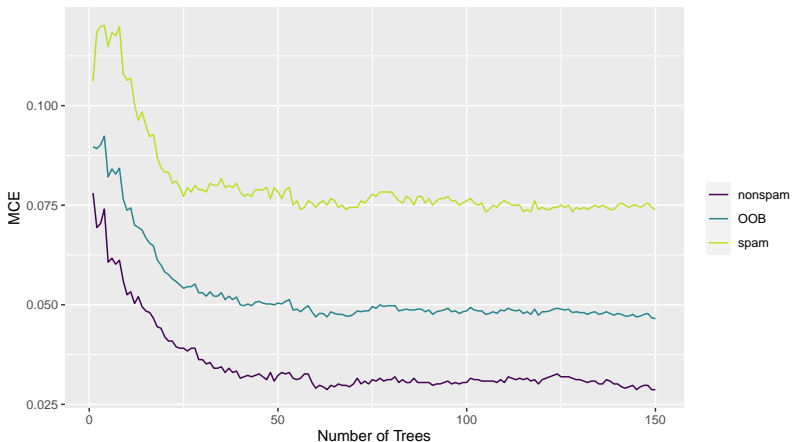


EFFECT OF ENSEMBLE SIZE

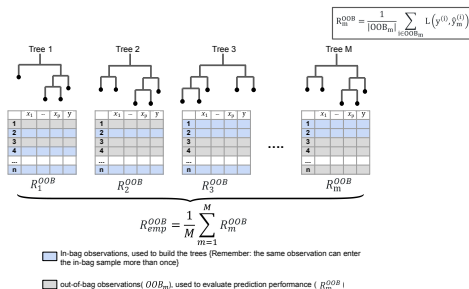


OUT-OF-BAG ERROR ESTIMATE

With the RF it is possible to obtain unbiased estimates of the generalization error directly during training, based on the out-of-bag observations for each tree:



OUT-OF-BAG ERROR ESTIMATE



- OOB size: $P(\text{not drawn}) = \left(1 - \frac{1}{n}\right)^n \xrightarrow{n \rightarrow \infty} \frac{1}{e} \approx 0.37$
- Predict all observations with trees that didn't use them for training and compute average loss of these predictions
- Similar to 3-CV, can be used for a quick model selection