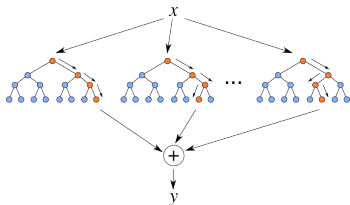


# Introduction to Machine Learning

## Random Forest: Introduction



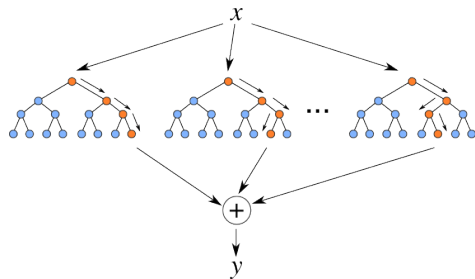
### Learning goals

- 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**

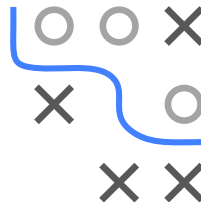
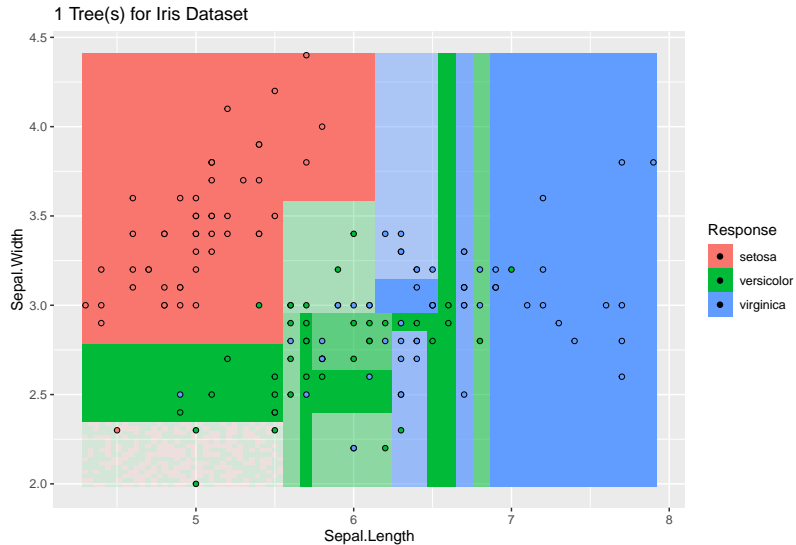


# RANDOM FEATURE SAMPLING

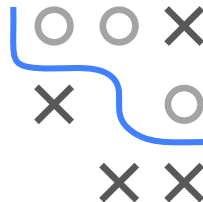
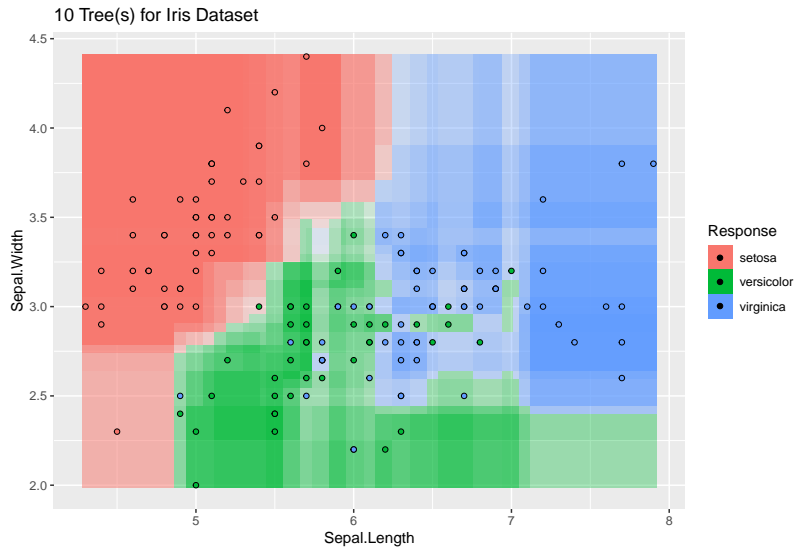
- From our analysis of bagging risk we can see that decorrelating trees improves the ensemble
- Simple randomized approach:  
At each node of each tree, randomly draw  $m_{\text{try}} \leq p$  candidate features to consider for splitting. Recommended values:
  - Classification:  $m_{\text{try}} = \lfloor \sqrt{p} \rfloor$
  - Regression:  $m_{\text{try}} = \lfloor p/3 \rfloor$



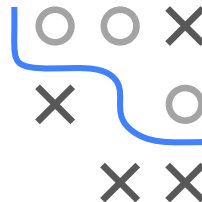
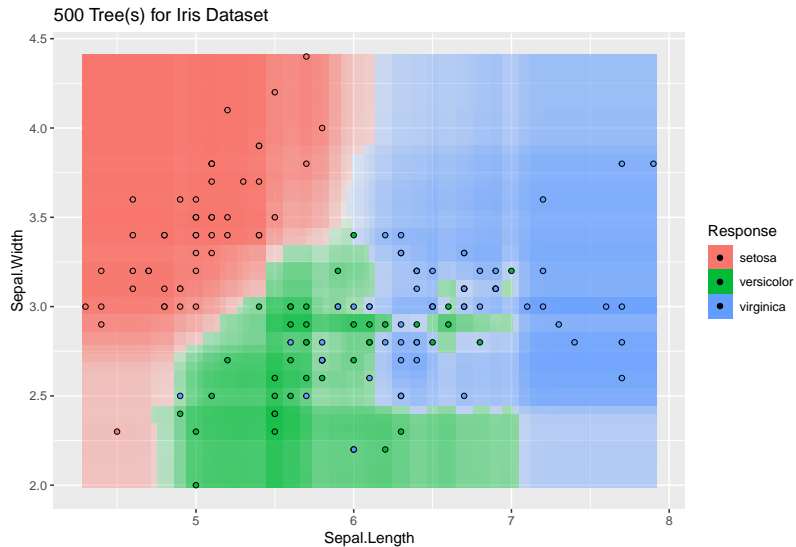
# EFFECT OF ENSEMBLE SIZE



# EFFECT OF ENSEMBLE SIZE / 2

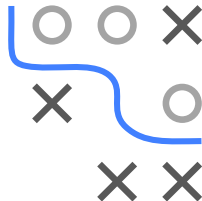
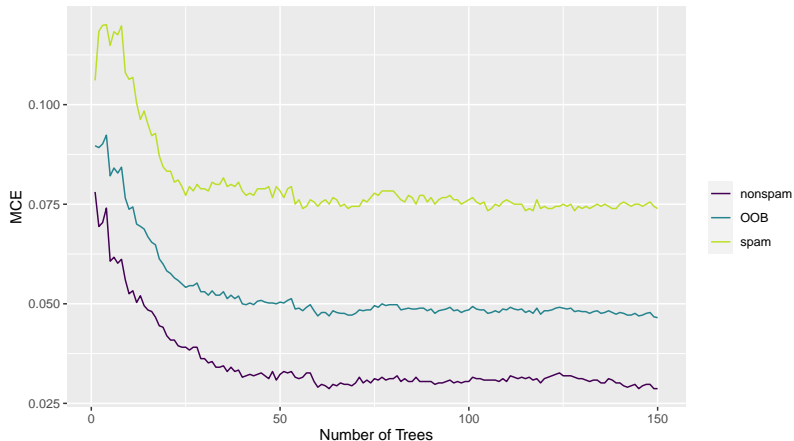


# EFFECT OF ENSEMBLE SIZE / 3

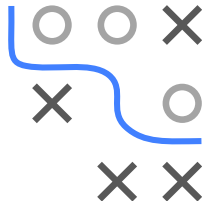
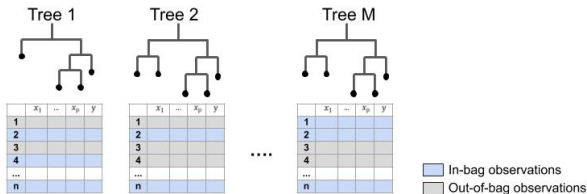


# OUT-OF-BAG ERROR ESTIMATE

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 / 2



- For an estimation of the generalization error, we exploit the fact that the  $i$ -th observation acts as unseen test point for all trees in which it is OOB.
- Let  $\text{OOB}^{[m]}$  denote the index set  $\left\{ i \in \{1, \dots, n\} \mid (\mathbf{x}^{(i)}, y^{(i)}) \text{ is OOB for } b^{[m]} \right\}$ .
- The number of trees for which the  $i$ -th observation is OOB is then given by  $S_{\text{OOB}}^{(i)} = \sum_{m=1}^M \mathbb{I}(i \in \text{OOB}^{[m]})$ .
- We can compute the average over predictions  $\hat{y}^{(i)[m]}$  from trees  $b^{[m]}$  that have observation  $i$  in their OOB data to obtain an ensemble prediction.
- The average loss of these ensemble OOB predictions over all  $n$  observations yields an estimate for the generalization error.



## OUT-OF-BAG ERROR ESTIMATE / 3

- Compute the ensemble OOB prediction for each observation:

$$\hat{y}_{\text{OOB}}^{(i)} = \begin{cases} \frac{1}{S_{\text{OOB}}^{(i)}} \sum_{m=1}^M \mathbb{I}(i \in \text{OOB}^{[m]}) \cdot \hat{y}^{(i)[m]} & \text{in regression,} \\ \arg \max_{k \in \{1, \dots, g\}} \frac{\sum_{m=1}^M \mathbb{I}(i \in \text{OOB}^{[m]}) \cdot \mathbb{I}(\hat{h}^{(i)[m]} = k)}{S_{\text{OOB}}^{(i)}} & \text{in classification.} \end{cases}$$

- Then, take the average of the resulting point-wise losses to estimate the OOB error of the forest:

$$\widehat{\text{err}}_{\text{OOB}} = \frac{1}{n} \sum_{i=1}^n L(y^{(i)}, \hat{y}_{\text{OOB}}^{(i)})$$

- Note that the use of class labels commands the use of 0-1 loss in classification (alternative formulations for other losses are possible).
- OOB size:  $\mathbb{P}(i \in \text{OOB}^{[m]}) = (1 - \frac{1}{n})^n \xrightarrow{n \rightarrow \infty} \frac{1}{e} \approx 0.37$  for  $i \in \{1, \dots, n\}$ .
- Similar to 3-CV, can be used for a quick model selection.

