Introduction to Machine Learning

Bagging and Random Forest 1

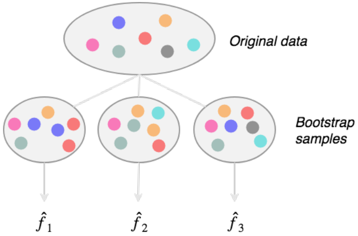
Department of Statistics – LMU Munich

BAGGING

- Bagging is based on **B**ootstrap **Agg**regation.
- Ensemble that improves instable / high variance learners by variance smoothing

Train on B **bootstrap** samples of data $\mathcal{D}$:

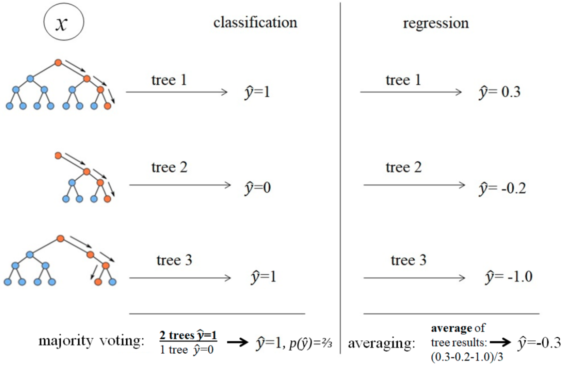
- Draw n observations with replacement
- Fit the base learner on each of the B bootstrap samples



BAGGING

Aggregate the predictions of the B estimators:

- Aggregate via averaging (regression) or majority voting (classification)
- Posterior probabilities for $\mathbf{x}$ in classification can be estimated by calculating class frequencies over the ensemble

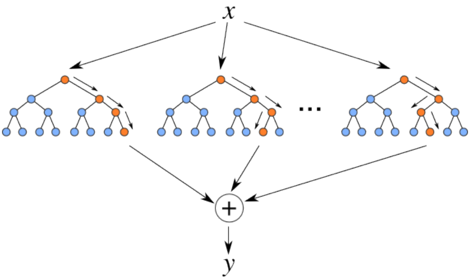


BAGGING

- Reduces variance of the predictor, but (slightly) increases its bias
- Bagging works best for unstable/high variance learners (learners where small perturbations of the training set can cause large changes in the prediction)
  - Classification and regression trees
  - Neural networks
  - Step-wise/forward/backward variable selection for regression
- For stable estimation methods bagging might degrade performance
  - k-nearest neighbor
  - discriminant analysis
  - naive Bayes
  - linear regression

RANDOM FORESTS

- Modification of bagging for trees proposed by Breiman (2001)
- Construction of bootstrapped decorrelated trees through randomized splits
- Trees are usually fully expanded, without aggressive early stopping or pruning, to increase variance

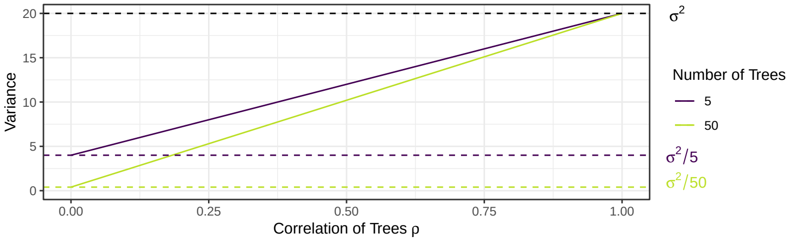


VARIANCE OF BAGGING

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 = \left(\rho + (1-\rho)\frac{1}{B}\right)\sigma^2$$

σ^2 is variance of a tree and ρ the correlation between trees

- If trees are highly correlated ($\rho \approx 1$), variance $\rightarrow \sigma^2$
- If trees are uncorrelated ($\rho \approx 0$), variance $\rightarrow \frac{\sigma^2}{B}$
- Variance can be reduced by increasing the number of trees B

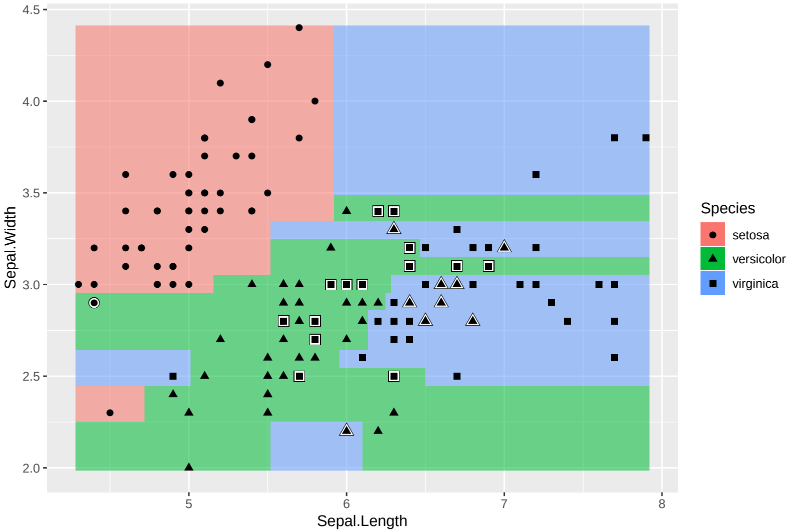


RANDOM FEATURE SAMPLING

- From our variance analysis we can see that decorrelating trees further might reduce the variance of the predictor
- Simple randomized approach:
  Instead of all p features, draw mtry $\leq p$ random split candidates.
  Recommended values:
  - Classification: $\lfloor \sqrt{p} \rfloor$
  - Regression: $\lfloor p/3 \rfloor$

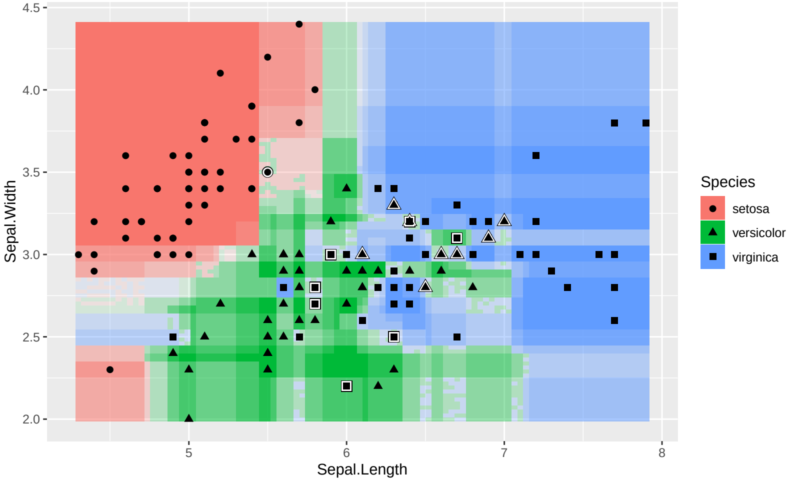
EFFECT OF ENSEMBLE SIZE

With 1 Tree on Iris



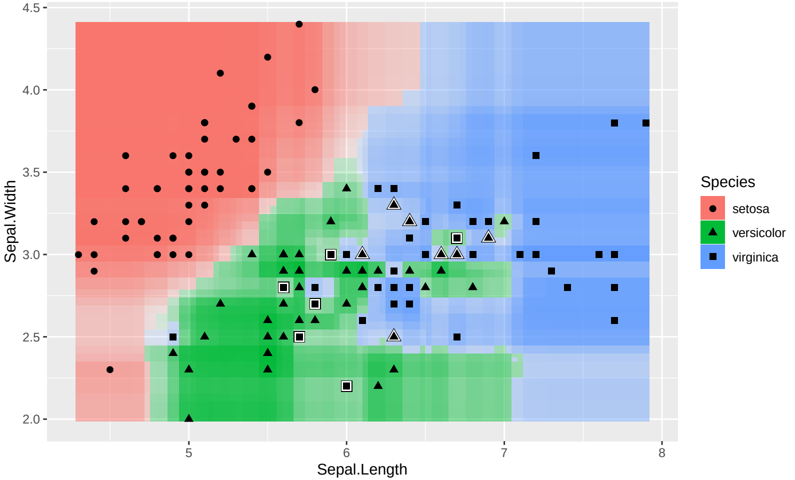
EFFECT OF ENSEMBLE SIZE

With 10 Trees on Iris



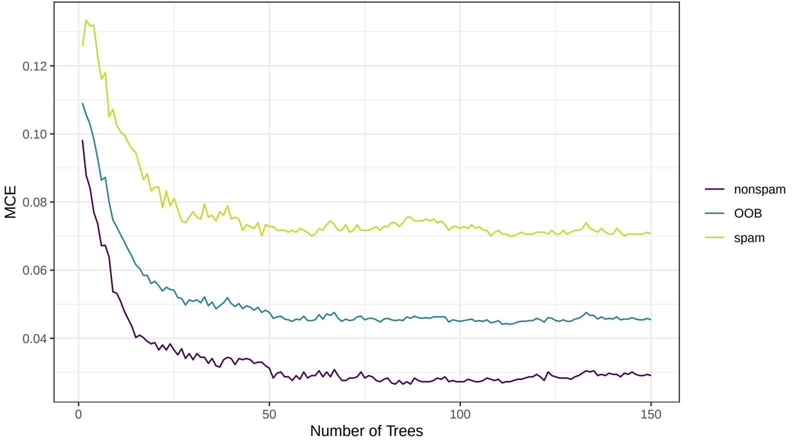
EFFECT OF ENSEMBLE SIZE

With 500 Trees on Iris

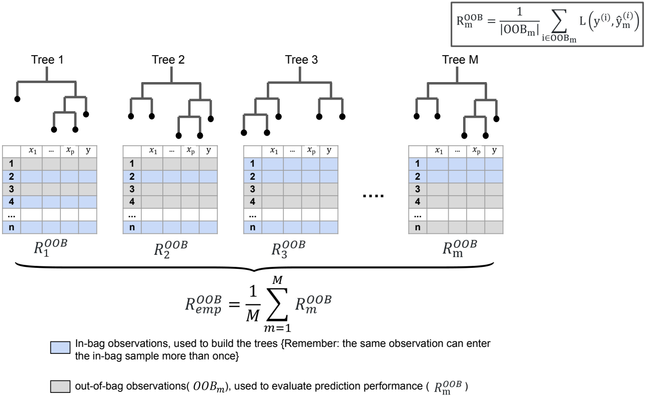


OUT-OF-BAG ERROR ESTIMATE

With the RF it is possible to obtain unbiased estimates of generalization error directly during training:



OUT-OF-BAG ERROR ESTIMATE



- OOB size: $P(\text{not drawn}) = \left(1 - \frac{1}{n}\right)^n \stackrel{n \to \infty}{\longrightarrow} \frac{1}{e} \approx 0.37$
- Predict all $\mathbf{x}$ with trees that didn't see it, average error
- Similar to 3-CV, can be used for a quick model selection