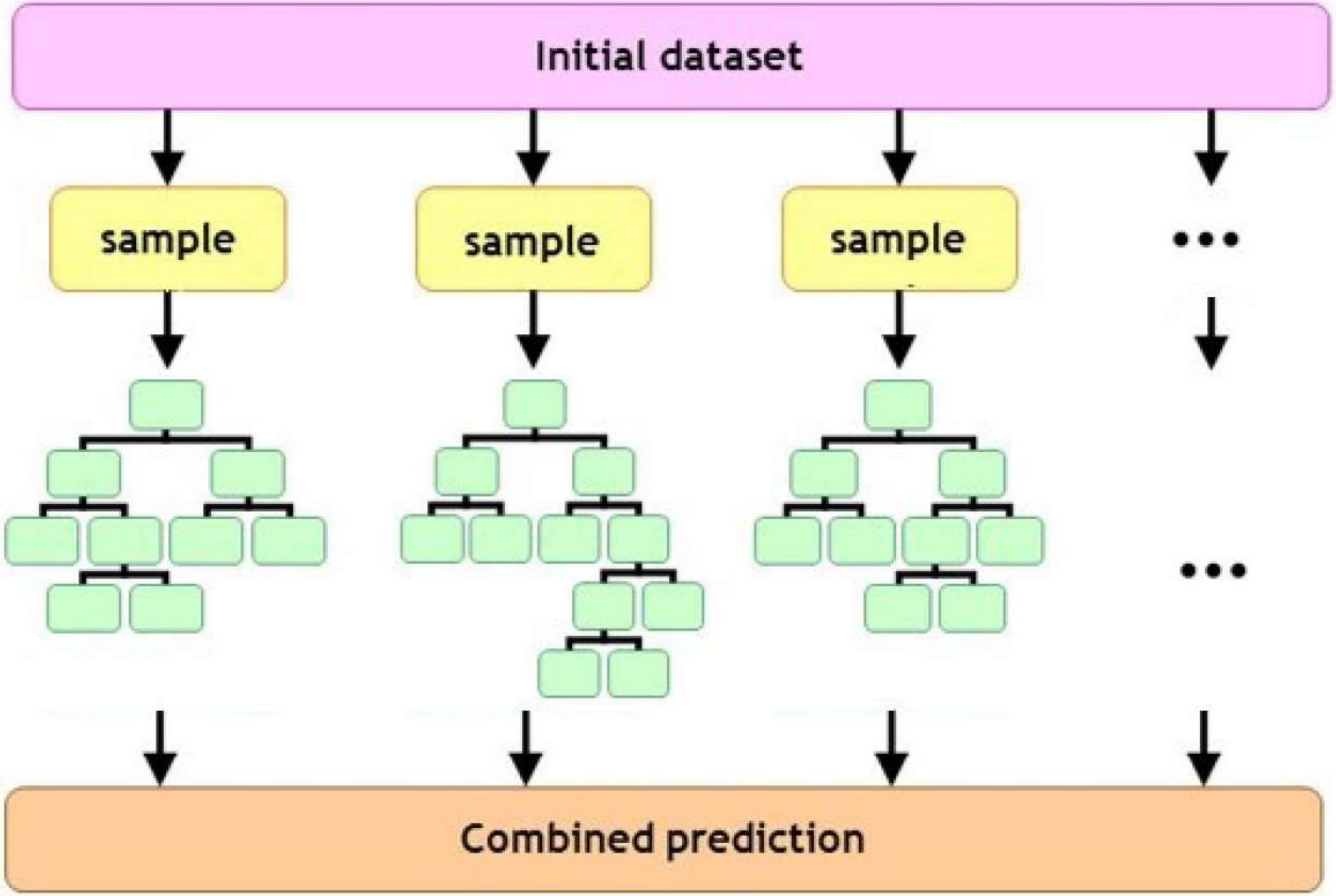
Single trees

- Good: Interpretability
- Bad: Not most accurate prediction method
- Ugly: Instability

Ensembling trees



Bagging

Draw B samples from the (initial) training dataset

- Use bootstrap (bagging) or sub sampling (subagging)
- Bootstrap yields higher inclusion frequencies for noise variables (De Bin et al., 2014)
  - Mostly disadvantageous for interpretation, less for
- Fit a tree $\hat{f}_{*b}(x)$ on each sample
- Final predictive model takes average over individual trees' predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{*b}(x)$$

- Note: Predictions of individual trees are class labels for binary outcomes.

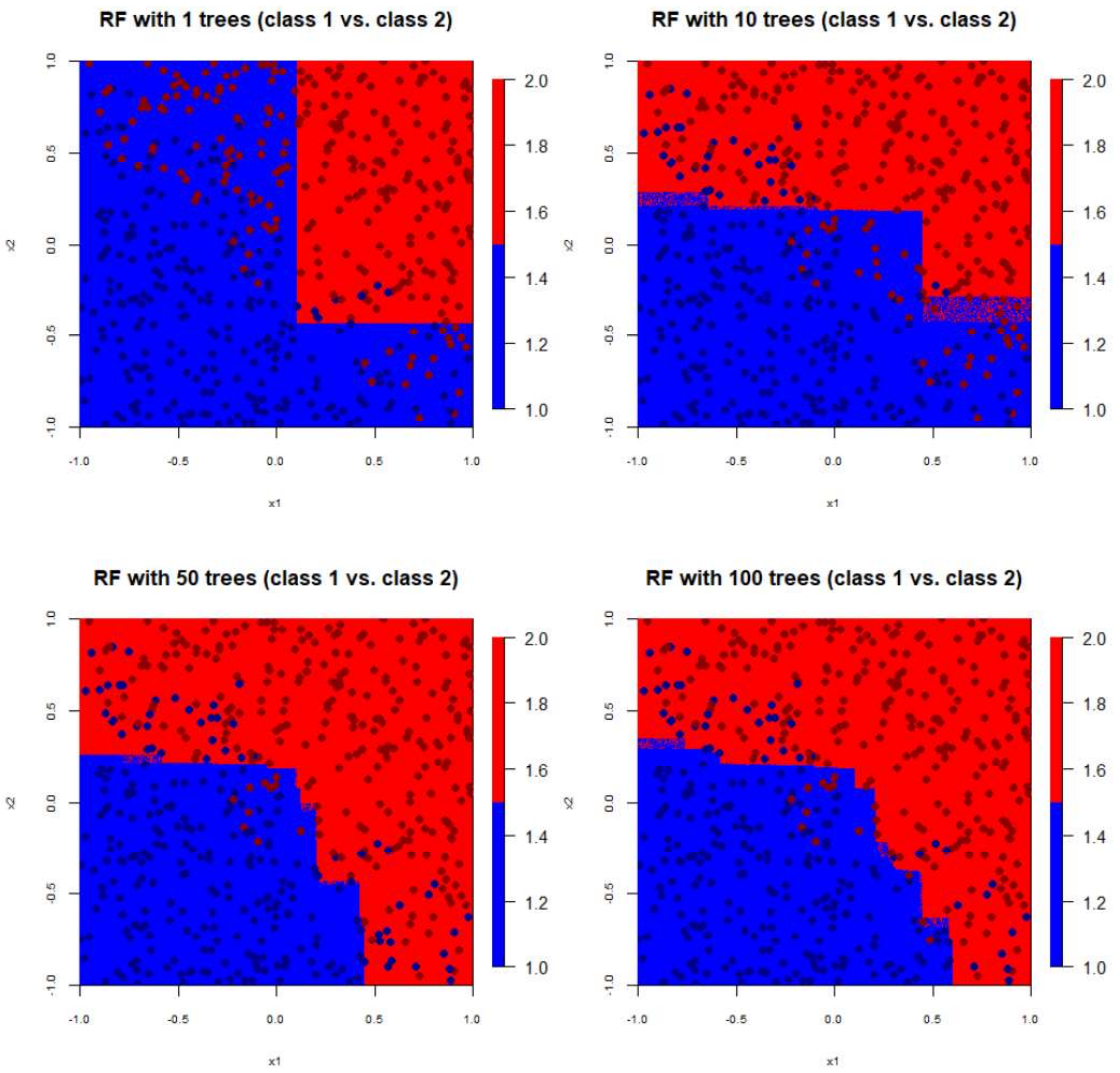
Random forest

Same as bagging, in addition:

- Select random sample of *mtry* candidate predictors for every split
- Random sampling of rows as well as columns:
- Trees become more dissimilar, thus less correlated (remember: we take advantage of *instability*!)
- Allows correlated predictors to also be selected for splitting
- Final predictive model is again:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{*b}(x)$$

Ensembling trees performs smoothing



Out-of-bag (OOB) error

Can be computed for every baselearner (trees) fitted on samples of the training data:

- Compute OOB predictions: For every training observation i, get predicted values from each tree, fitted on samples *excluding* observation i.
- Take the average (or majority vote) to obtain the OOB prediction $\hat{y}_i^{OOB}$
- Compute MSE (or other error measures) on OOB predictions as usual:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

- While OOB error is computed using training data, it provides better (less optimistic) estimate of generalization error than training error.

Interpretation: Variable importance

Importance of a predictor variable j can be computed in many (!) ways. E.g.:

- Training error: Sum over the error reductions resulting from each split involving variable j (comparable to sums of squares in GLMs)

- OOB permutation importance:

  1. Compute OOB error MSE_{OOB}
  2. For each variable j, randomly permute values of variable x_j
  3. Again generate predictions and compute MSE_{OOB}
  4. Difference between the MSE_{OOB} under 1) and 3) is the importance of variable j

Variable importances

Use with care:

- "Importance" of a variable may sound exactly like what we want to know!
- But importances merely quantify contribution of a variable to the predictions of a given fitted model
- Thus, importance of the same variable will differ between different fitted models
- Importances are not so well defined as e.g., linear regression coefficients
- Behavior under multicollinearity, higher-order interactions may not be as expected (e.g., Strobl et al., 2007, 2008; Nicodemus et al., 2010)
- Thus, use only as a rough and approximate ordering of relevance.

- Different packages may compute importances in a different manner. Always consult help files!

Interpretation: Partial dependence functions

- The effect of a predictor variable can be computed using a partial dependence function:
$$\bar{f}(x_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_j, x_{i\setminus j})$$

- where x_j indicates the predictor variable of interest, and $x_{\setminus j}$ are all the remaining variables.

- This computes *marginal* effects (cf. GAMs, where effects are additive, and conditional effects can be computed)

- Like variable importances, partial dependence plots should be interpreted with care:
  - Possible interactions of x_j with other variables are averaged over.
  - Different packages may use different computation strategy.

References

De Bin, R., Janitza, S., Sauerbrei, W., & Boulesteix, A. L. (2015). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*, *72*, 272-280.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 110.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 25.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307.