Bagging

University of Santiago de Compostela
Manuel Mucientes

Bagging

- Technique to reduce the variance of an estimated prediction function
- Works especially well for high-variance, low-bias procedures
  - Example: trees
- Given a set of n independent observations Z_1, …, Z_n, each with variance σ^2:

$$var\left(\bar{Z}\right) = \sigma^2/n$$

- Naïve approach:
  - Use many training sets
  - Average predictions of the models

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

Bagging

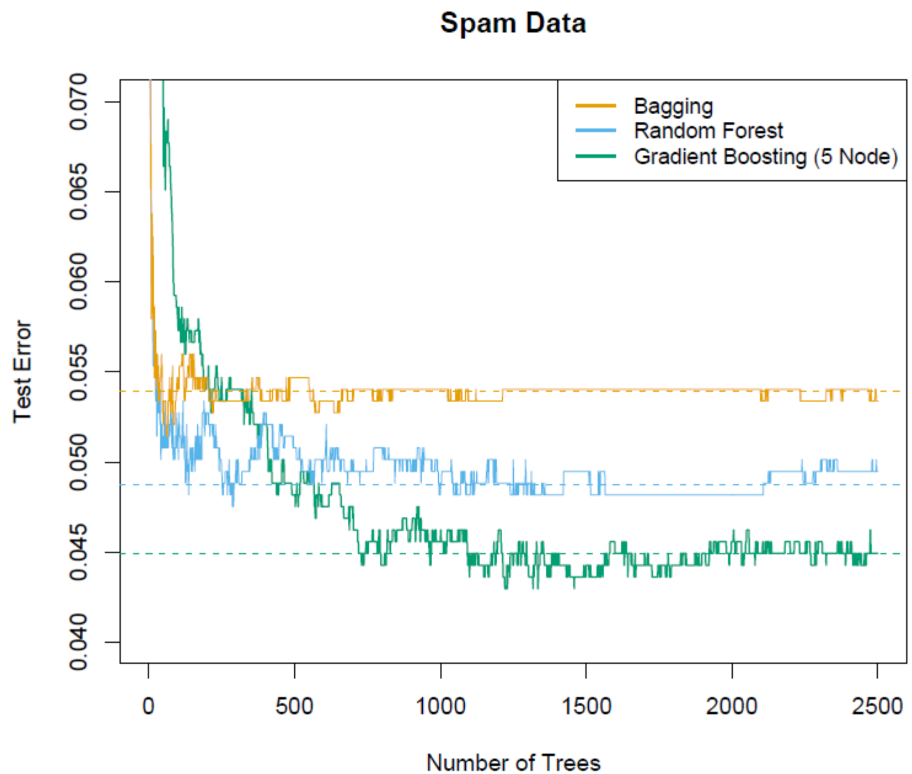
- Use bootstrap to take samples from the training set
  - Generate *B* different bootstrapped training sets
  - Reduce the variance by averaging many noisy but approximately unbiased models
- Trees are ideal candidates for bagging:
  - Capture complex information
  - If sufficiently deep, relatively low bias
- Regression: $\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$
- Classification: $\hat{G}_{\text{bag}}(x) = \arg\max_k \hat{f}_{\text{bag}}(x)$
  - f_{bag}: proportion of trees predicting each class
    - Average probabilities for each class (better)

Bagging

- Boosting: trees are grown to remove bias; not i.d.
- Bias of bagged trees is the same as individual trees
  - Trees are i.d.
  - Improve by variance reduction
- Bagging: as *B* increases, variance reduces, but up to a limit
- Variance of the average
  - i.i.d. trees: $\frac{1}{B}\sigma^2$
  - i.d. trees: $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$
    - *B* large: correlation increases

Random Forest (RF)

- Boosting appears to dominate bagging in most problems
- RF: improve variance reduction of bagging
  - Decrease correlation without increasing variance too much for single trees
  - Performance similar to boosting, but simpler to train and tune
- Radom selection of input variables as candidates for splitting
  - Typical value for *m*: $\sqrt{p}$
  - Reducing *m* will reduce the correlation between trees



Random Forest

Algorithm 15.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.

   (b) Grow a random-forest tree T_b to the bootstrapped data, by re-cursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

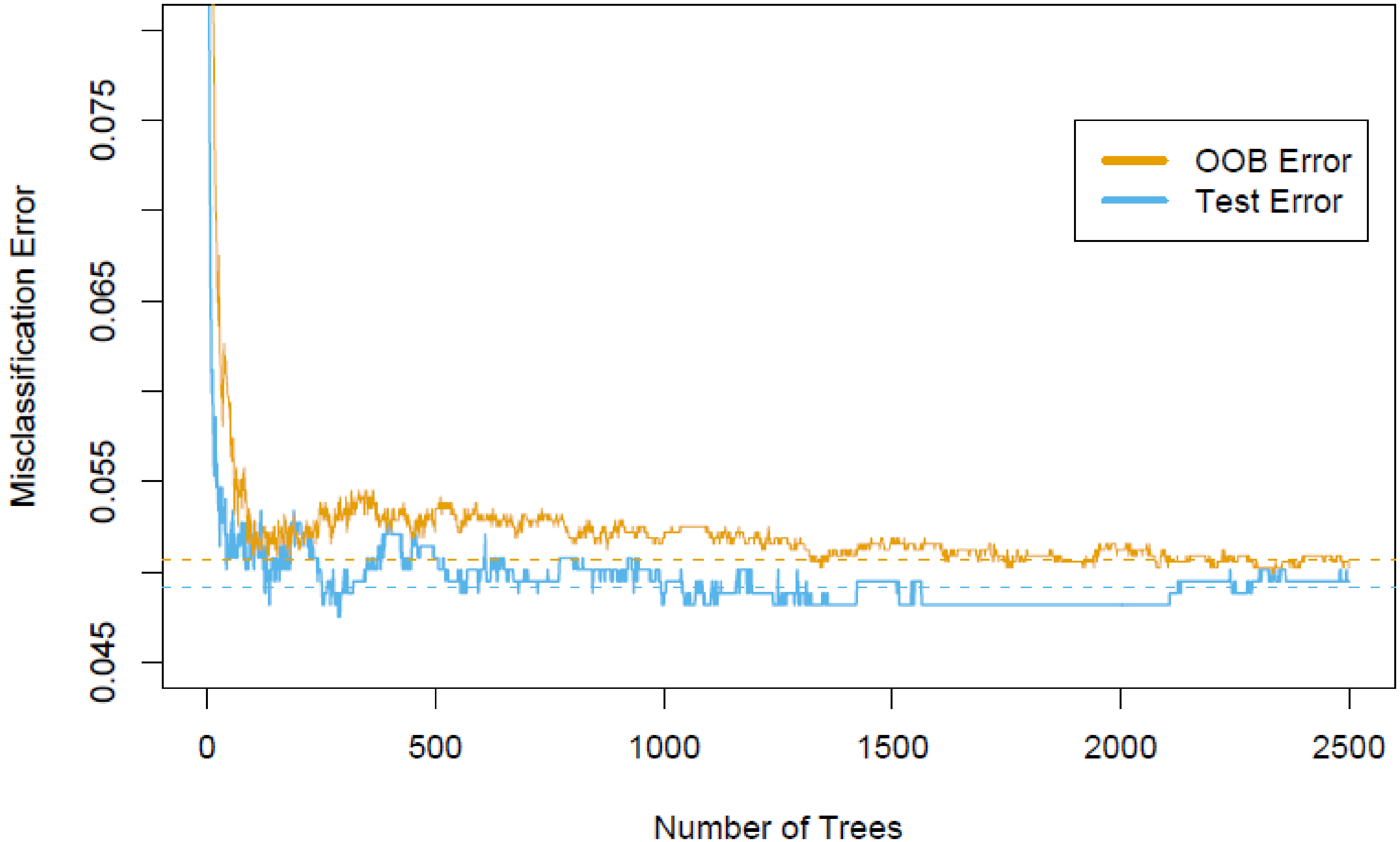
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

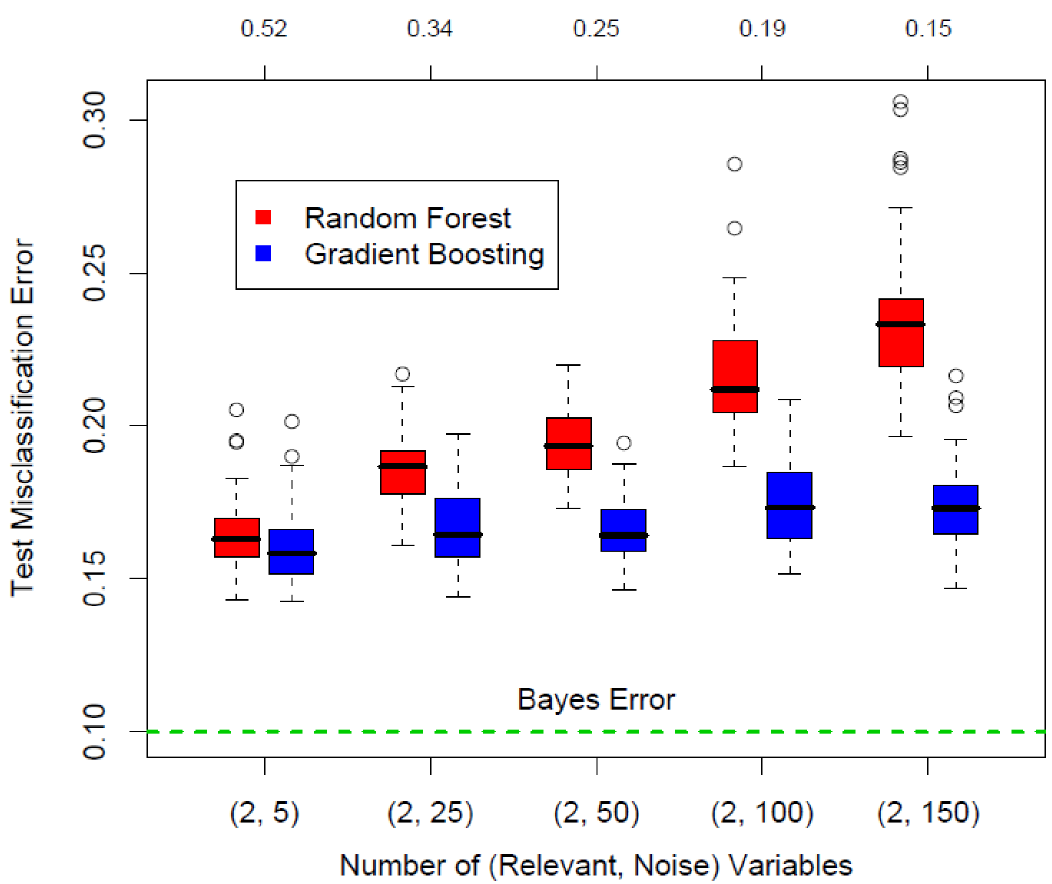
Out of Bag (OOB) error

- OOB samples: for each observation $z_i = (x_i, y_i)$, construct output by averaging trees in which z_i did not appear
- OOB error almost identical to N-fold cross-validation
- With OOB, RF can be fit in one sequence



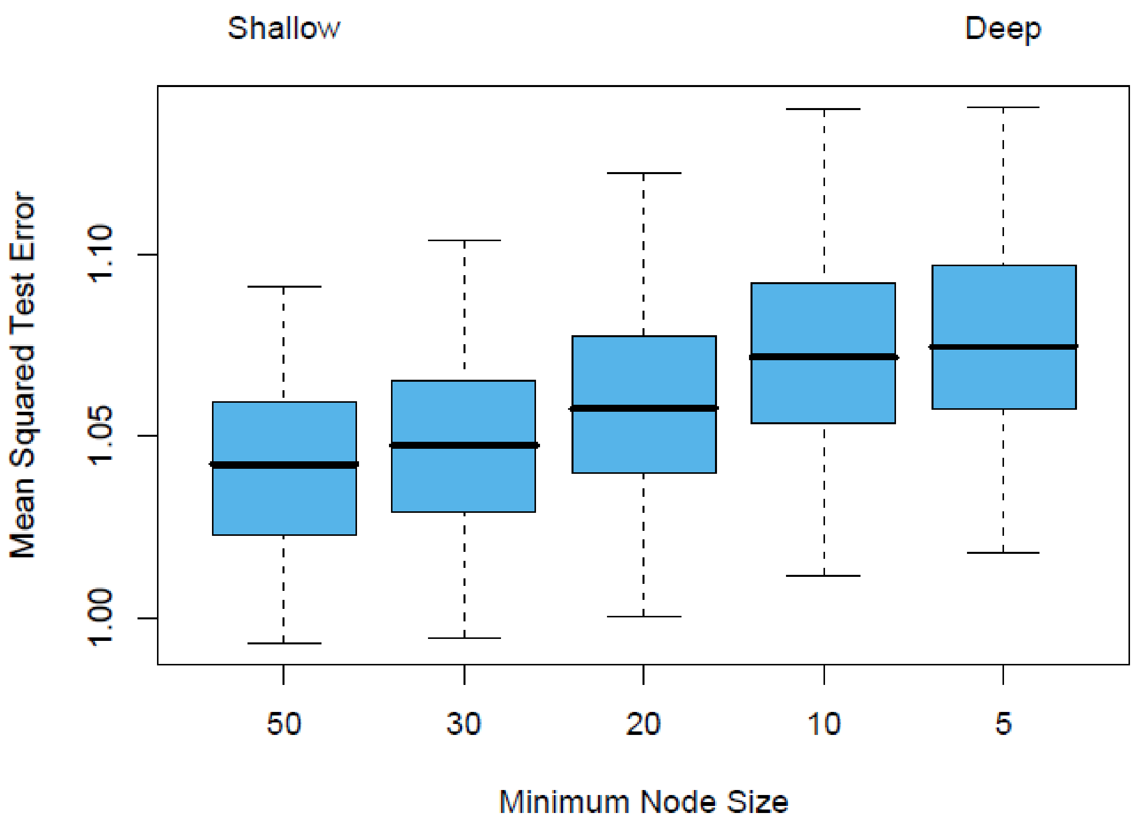
RF and overfitting

- Number of variables high, but fraction of relevant variables small: RF performs poorly with small *m*
- If number of relevant variables increases, RF is robust to an increase in number of noise variables
- Simulated example:
  - m= $\sqrt{p}$
    - At top, prob. that a relevant variable is chosen at any split
    - (6, 100) gives 0.46 vs. (2, 100) gives 0.19



RF and overfitting

- RF can overfit for large *B*
- Small gains in performance by controlling the depths of the individual trees in RF
  - Full-grown trees seldom cost much
  - One less tuning parameter
- Example: low increase in error for deeper trees



Bibliography

- T. Hastie, R. Tibshirani, y J. Friedman, The elements of statistical learning. Springer, 2009.
  - Chapter 15

- G. James, D. Witten, T. Hastie, y R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer, 2021.
  - Chapter 8, Sec. 8.2