Ensemble methods: Bagging, boosting, random forests

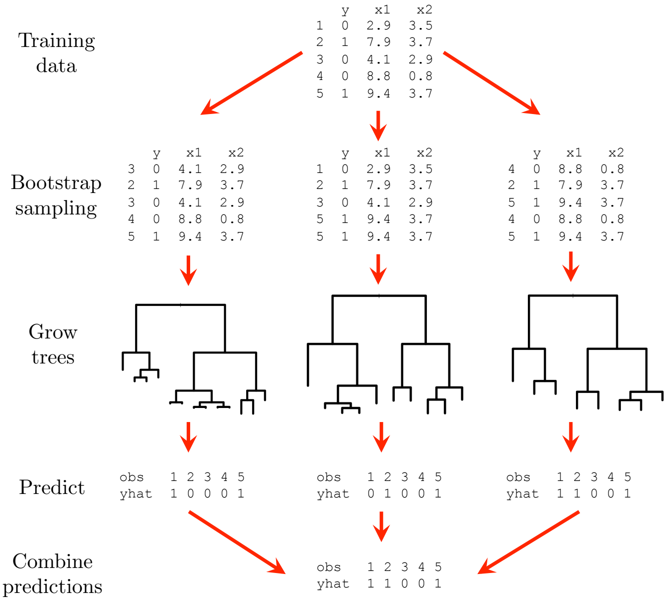
Machine Learning for Ecology workshop

SEEC - Statistics in Ecology, Environment and Conservation

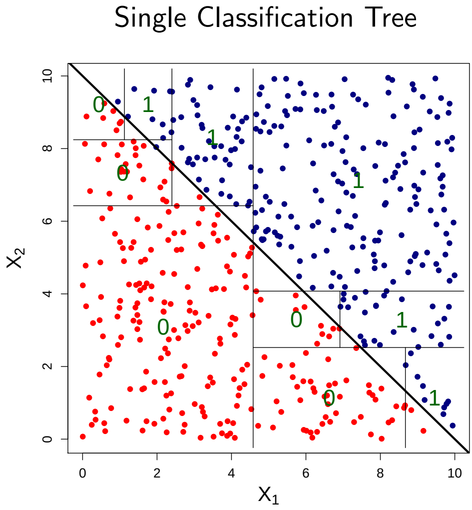
The Problem of High Variance

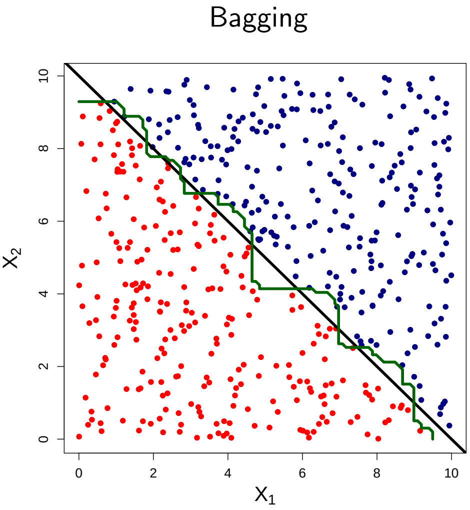
- Trees suffer from *high sampling variability*
- Small changes to sample $\Rightarrow$ Large changes in fitted tree
- Bootstrap aggregation or *bagging* is a general purpose procedure for variance reduction
- General idea: averaging reduces variance ($var(\bar{X}) = \sigma^2/n$)

Bagged regression trees

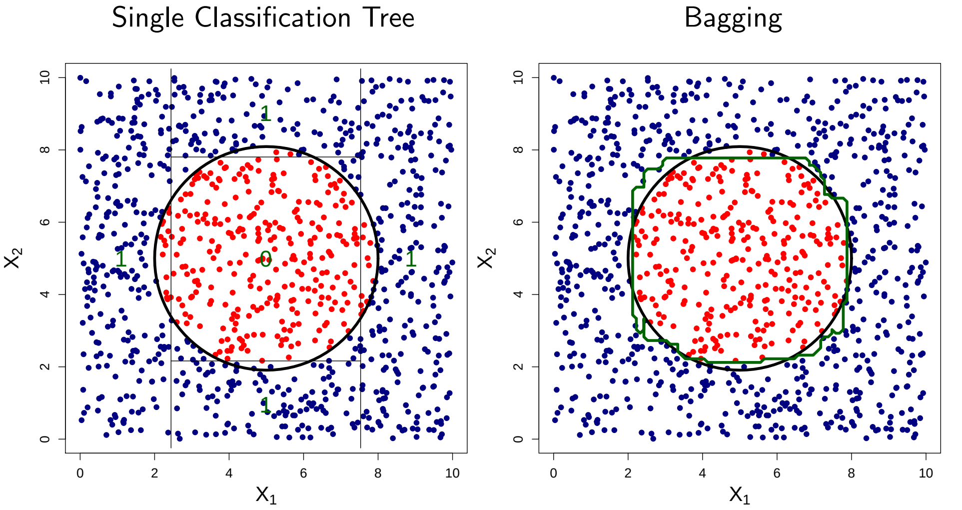


Why does bagging help?





Why does bagging help?



Cross-validation for bagging: Out-of-Bag Error

	y	x1	x2
1	1.1	2.9	3.5
2	1.7	7.9	3.7
3	2.3	4.1	2.9
4	1.8	8.8	0.8

Bootstrap samples

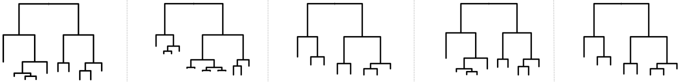
	y	x1	x2
2	1.7	7.9	3.7
2	1.7	7.9	3.7
3	2.3	4.1	2.9
1	1.1	2.9	3.5

	y	x1	x2
4	1.8	8.8	0.8
2	1.7	7.9	3.7
4	1.8	8.8	0.8
4	1.8	8.8	0.8

	y	x1	x2
1	1.1	2.9	3.5
3	2.3	4.1	2.9
3	2.3	4.1	2.9
1	1.1	2.9	3.5

	y	x1	x2
1	1.1	2.9	3.5
2	1.7	7.9	3.7
1	1.1	2.9	3.5
2	1.7	7.9	3.7

	y	x1	x2
4	1.8	8.8	0.8
2	1.7	7.9	3.7
2	1.7	7.9	3.7
4	1.8	8.8	0.8



OOB predictions

	yhat1
1	
2	
3	
4	2.0

	yhat2
1	1.3
2	
3	2.1
4	

	yhat3
1	
2	1.8
3	
4	1.7

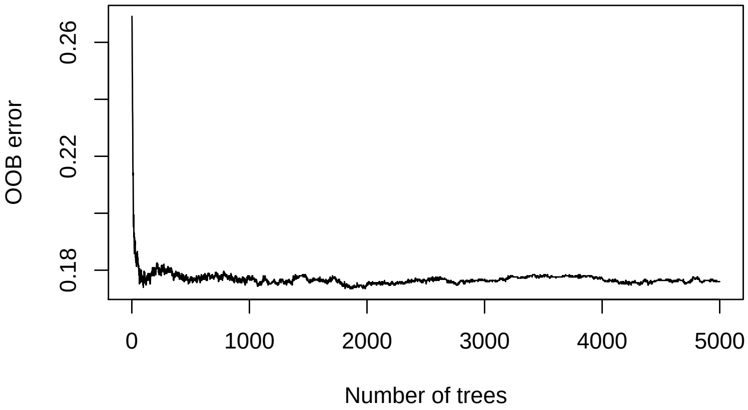
	yhat4
1	
2	
3	1.9
4	2.8

	yhat5
1	1.0
2	
3	2.3
4	

	yhat	e2
1	1.2	0.01
2	1.8	0.01
3	2.1	0.04
4	2.2	0.16

$\boxed{\text{OOB error} = 0.22}$

Out-of-Bag Error Estimation



Variable Importance

- The key advantage of a decision tree is ease of interpretation
- When we bag a large number of trees, it is no longer possible to represent the model with a single tree
- Variable importance: for each tree, record improvement in splitting criterion due to each predictor, and average over all trees
- More on this later

Random forests

- A small tweak that *decorrelates* the trees produced by bagging
- Each time a split is considered, *a random sample of* $m < p$ *predictors* are chosen as split candidates
- Bagging is a special case with $m = p$

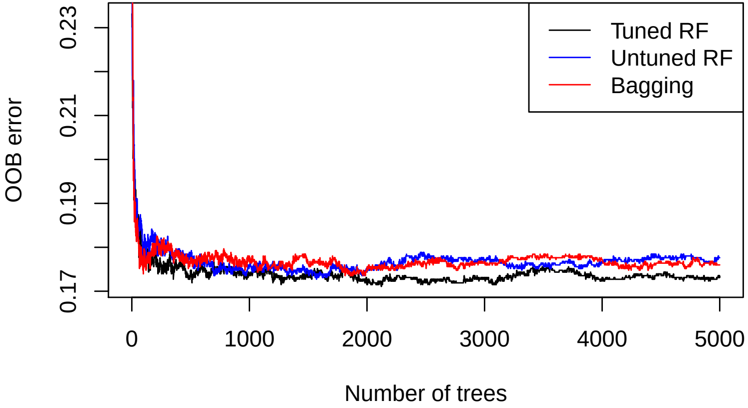
Why random forests work

- Bagging reduces the sampling variability of predictions by averaging over many trees
- Correlated trees are bad for variance reduction

$$\mathsf{Var}[\bar{Z}] = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2} \sum_{i \neq j} \rho_{ij}$$

- In practice bagged trees often are correlated (one strong predictor)
- Default is $m \approx \sqrt{p}$ for classification and $m \approx p/3$ for regression trees

Out-of-Bag Error Estimation



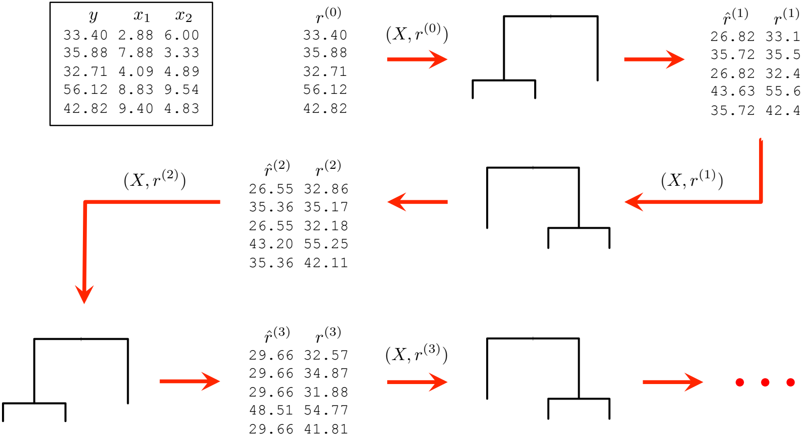
Boosting

- Bagging and RFs: each tree is grown independently of all other trees
- Boosting: grows trees *sequentially* using information from previously trees

Boosting

- First, grow a regression tree with a small number of splits, d
- The residuals of this tree are then treated as the response variable and used to grow another tree
- And so on…
- Yields a sequence of B trees, each accounting for some variation not explained by the previous trees

Boosting



Boosting algorithm with
$d = 2$ and $\lambda = 0.01$

$$\hat{y} = \lambda \sum_{b=1}^{B} \hat{r}^{(b)}$$

Why boosting works

- **Slow learning**: methods that learn slowly tend to perform well
- Small trees give slow improvement (small d)
- Slow learning by down-weighting contribution of each tree

$$\hat{y} = \lambda \sum_{b=1}^{B} \hat{r}^{(b)}$$

- Slow learning $\Rightarrow$ many trees

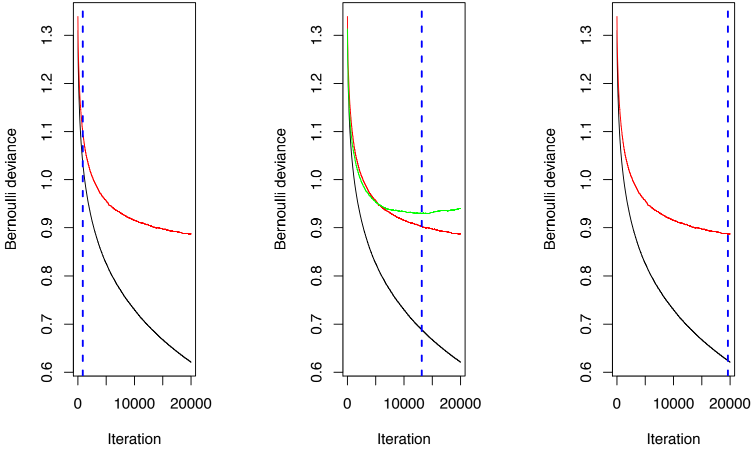
Tuning Parameters

1. The number of trees, B
   Unlike bagging and random forests, boosting can lead to overfitting if B is too large (why?). We use cross-validation to select B.

2. The shrinkage parameter or learning rate, λ
   Typical values are 0.01 or 0.001 (default in R package `gbm`). Smaller values require more trees.

3. Number of splits in each tree, d
   d is also called the *interaction depth* of the boosted model, since d splits can involve at most d variables. Often $d = 1$ works well.

Selecting number of trees B



Summary

- Bagging to reduce variance
- Random forests to reduce correlation (and hence variance)
- Boosting and the power of slow learning
- Out-of-bag but still need test sample if tuning
- **Next:** Variable importance and interpretation