

# RECSM Summer School:

## Machine Learning for Social Sciences

Session 2.3:  
Bagging and Random Forests

Reto Wüest

Department of Political Science and International Relations  
University of Geneva



# Bagging

---

# Bagging

- Decision trees suffer from **high variance**: small changes in the training data can lead to quite different results.
- We would like to have a method with **low variance**: the results are similar if the method is applied repeatedly to distinct data sets (generated by the same DGP).
- **Bootstrap aggregation**, or **bagging**, is a general-purpose procedure for **reducing the variance** of a learning method.
- Bagging is frequently used in the context of decision trees.

## Averaging Reduces Variance

Suppose there is a set of  $n$  independent variables  $X_1, \dots, X_n$  with variance  $V(X_i) = \sigma^2$ ,  $\forall i = 1, \dots, n$ . The variance of the mean  $\bar{X}$  of these variables then is  $V(\bar{X}) = \sigma^2/n$ .

# Bagging

- Averaging a set of variables **reduces variance**.
- Hence, we could **reduce the variance** (increase the prediction accuracy!) of a machine learning method as follows:
  - Take  $B$  training sets from the population;
  - Train the method on each training set to get predictions  $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ ;
  - Average the resulting predictions

$$\hat{f}^{\text{avg}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x). \quad (2.3.1)$$

# Bagging

- We generally **do not have access** to multiple training sets.
- Instead, we can **bootstrap**:
  - Generate  $B$  bootstrapped training sets by taking repeated samples from the (single) training set;
  - Train the method on the  $b$ th bootstrapped training set to get prediction  $\hat{f}^{*b}(x)$ ;
  - Average all predictions to obtain

$$\hat{f}^{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x). \quad (2.3.2)$$

- This approach is called **bagging**!

## Bagging – Applied to Regression Trees

- Construct  $B$  regression trees using  $B$  bootstrapped training sets, and average the resulting predictions.
- Each tree is grown deep and is not pruned. Hence, each tree has low bias, but high variance.
- Averaging these  $B$  trees reduces the variance.
- Bagging has been shown to give impressive improvements in accuracy by combining hundreds or thousands of trees.

## Bagging – Applied to Classification Trees

- How can bagging be extended to a classification problem?
- Construct  $B$  **classification trees** using  $B$  bootstrapped training sets.
- For a given test observation, we record the class predicted by each of the  $B$  trees, and take a “**majority vote**.”
- Hence, the overall prediction is the **most commonly occurring class** among the  $B$  predictions.



# Bagging

- Increasing the number of trees  $B$  will not lead to overfitting. In practice, we want to use a value of  $B$  that is sufficiently large for the test error to have settled down.
- How do we estimate the test error of a bagged model?

# Bagging

---

## Out-of-Bag Error Estimation

# Out-of-Bag Error Estimation

- With bagging, we can estimate the test error **without the need to perform CV**.
- Recall that the trees are repeatedly fit to **bootstrapped subsets** of the training set.
- It turns out that, on average, each tree is fit to around  $2/3$  of the training observations. The remaining  $1/3$  of the training observations not used to fit a given tree are called the **out-of-bag (OOB)** observations.

## Out-of-Bag Error Estimation

- We can predict the response for observation  $i$  using all trees in which that observation was OOB. This will yield about  $B/3$  predictions.
- To obtain a single prediction for observation  $i$ , we can **average** these predicted responses (regression) or take a **majority vote** (classification).
- After doing this for all  $n$  observations, we can compute the **overall OOB MSE** (regression) or **classification error** (classification).
- The resulting OOB error is a **valid estimate** of the test error for the bagged model.

# Bagging

---

## Measuring Variable Importance

# Measuring Variable Importance

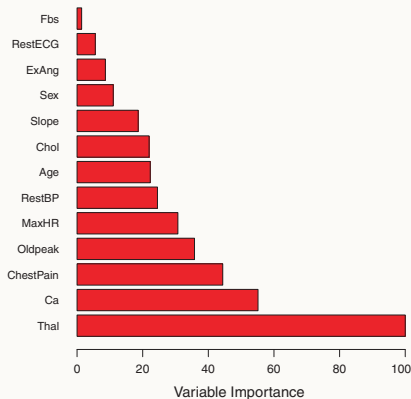
- Bagging typically has a **better prediction accuracy** than a single tree.
- This comes at the **expense of interpretability**: it is no longer possible to represent the model as a single tree and it is no longer clear which variables are most important.
- Therefore, it can be useful to compute an overall **summary of the importance** of each predictor using the RSS (regression) or the Gini index (classification).

# Measuring Variable Importance

- For **regression trees**: we can record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all  $B$  trees.
- For **classification trees**: we can record the total amount that the Gini index is decreased due to splits over a given predictor, averaged over all  $B$  trees.
- In both cases, a **large value** indicates an **important predictor**.

# Measuring Variable Importance – Example

A Variable Importance Plot for the Heart Disease Data



(Source: James et al. 2013, 320)

The plot shows the mean decrease in the Gini index for each variable, relative to the largest.



# Random Forests

---

# Random Forests

- Random forests provide an improvement over bagged trees.
- They involve a small tweak that decorrelates the trees:
  - As in bagging, we build a number of decision trees on bootstrapped training samples.
  - At each split in the tree-building process, we consider a random sample of  $m$  predictors,  $m < p$ , as candidates for the split.
  - A new sample of  $m$  predictors is taken at each split, typically of size  $m \approx \sqrt{p}$ .
- Therefore, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors.

# Random Forests

- Does this sound crazy?
- Suppose that there is **one very strong predictor** in the training data, along with a number of moderately strong predictors.
- In bagging, most or all of the individual trees will use this strong predictor in the **top split**.
- Consequently, all bagged trees will look **quite similar** to each other, so the predictions from these trees will be **highly correlated**.

# Random Forests

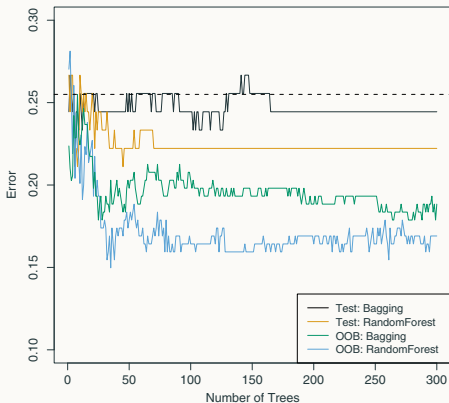
- Averaging highly correlated quantities leads to a **smaller reduction in variance** than averaging uncorrelated quantities.
- Therefore, bagging will not lead to a substantial reduction in variance over a single tree.
- In random forests, on average  $(p - m)/p$  of the splits will **not even consider** the strong predictor.
- Random forests **decorrelate** the trees, making the average of the trees **less variable** and hence **more reliable**.

# Random Forests

- The difference between bagging and random forests depends on the choice of predictor subset size  $m$ .
- If  $m = p$ , then the random forest is **equivalent** to bagging.
- As with bagging, random forests will not overfit if we increase  $B$ , so in practice we use a **sufficiently large** value of  $B$  ( $B$  is sufficiently large when the error rate has settled down).

# Bagging and Random Forests – Example

## Bagging and Random Forest Results for the Heart Disease Data



(Source: James et al. 2013, 318)

The dashed line indicates the test error resulting from a single classification tree. Random forests were applied with  $m = \sqrt{p}$ .