Lecture 19 Tree-based Methods III: Bagging and Random Forest

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

March 23, 2021

Outline

Bagging

- Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method;
- we introduce it here because it is particularly useful and frequently used in the context of decision trees.
- In other words, averaging the predictions, each trained on a different training set, can reduce the prediction variance.
- Of course, this is not practical because we generally do not have access to multiple training sets.

Bagging

- Instead, we can bootstrap, by taking repeated samples from the (single) training data set.
- In this approach we generate B different bootstrapped training data sets.
- We then train our method on the b-th bootstrapped training set in order to get $\hat{f}^{*b}(x)$, the prediction at a point x.
- We then average all the predictions to obtain

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

- This is called bagging.

Bagging classification trees

- The above prescription applied to regression trees.
- For classification trees: for each test observation, we record the class predicted by each of the B trees, and take a majority vote: the overall prediction is the most commonly occurring class among the B predictions.

Out-of-Bag Error Estimation

P[an observation is not sampled] = (1-1/n)^n = 0.37 (if n is large) approximately 1/3

- It turns out that there is a very straightforward way to estimate the test error of a bagged model.
- Recall that the key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. One can show that on average, each bagged tree makes use of around two-thirds of the observations. if n samples, E[|one bootstrapped dataset|] = 2n/3
- The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations.
- We can predict the response for the ith observation using each of the trees in which that observation was OOB. This will yield around $B/3$ predictions for the i-th observation, which we average. for bagging, if B is large.

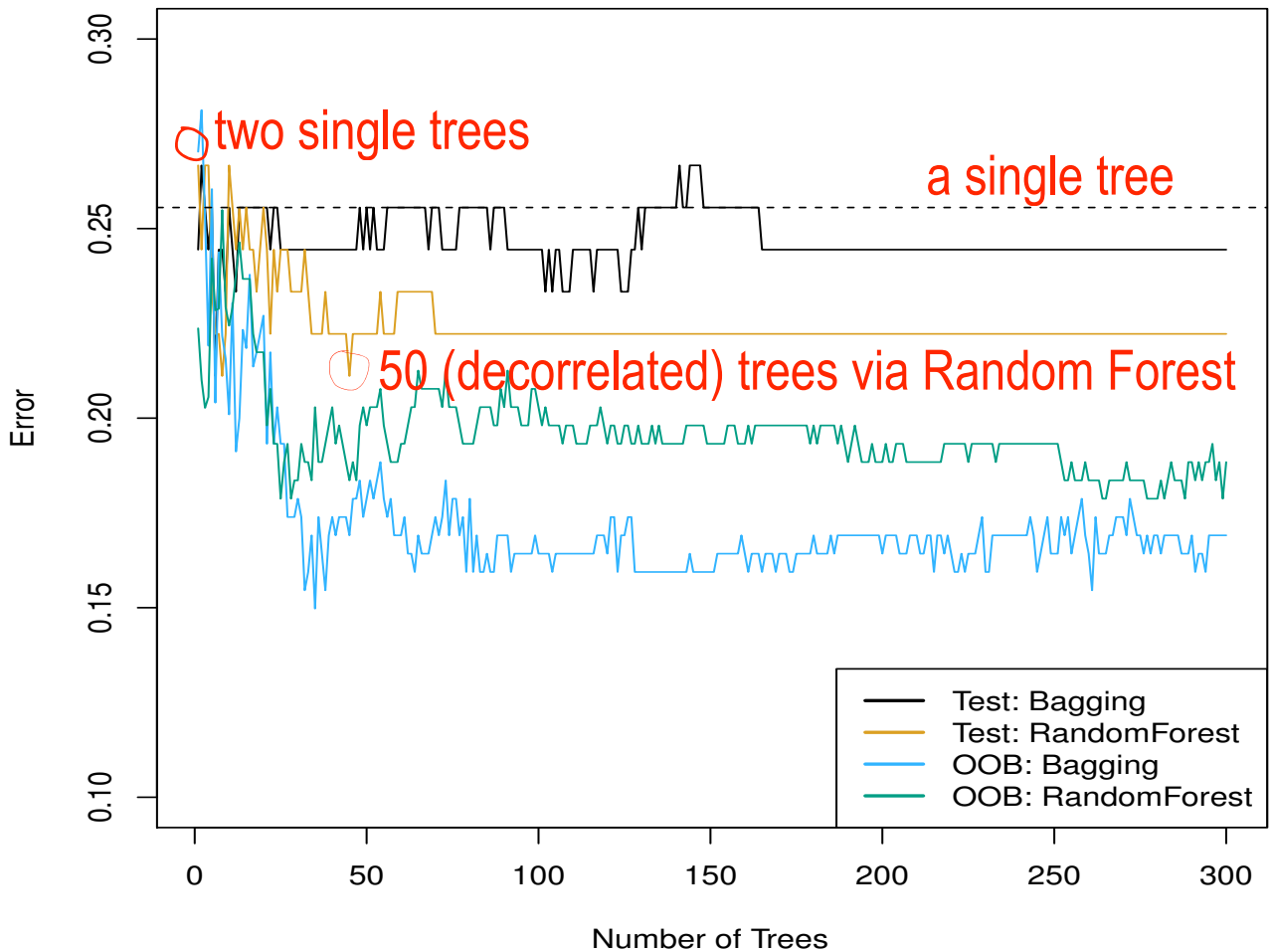
overall OOB error = average of OOBerr_i

recursive binary splitting, each step:
1. pick a terminal node
2. pick a predictor
3. pick a cutpoint

Random Forests

- Random forests provide an improvement over bagged trees by way of a small tweak that decorrelates the trees. This reduces the variance when we average the trees.
- As in bagging, we build a number of decision trees on bootstrapped training samples.
- But when building these decision trees, each time a split in a tree is considered, a random selection of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$ — that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (4 out of the 13 for the Heart data).

Heart Data


two single trees
a single tree
50 (decorrelated) trees via Random Forest
Error
0.10
0.15
0.20
0.25
0.30
0
50
100
150
200
250
300
Number of Trees
Test: Bagging
Test: RandomForest
OOB: Bagging
OOB: RandomForest

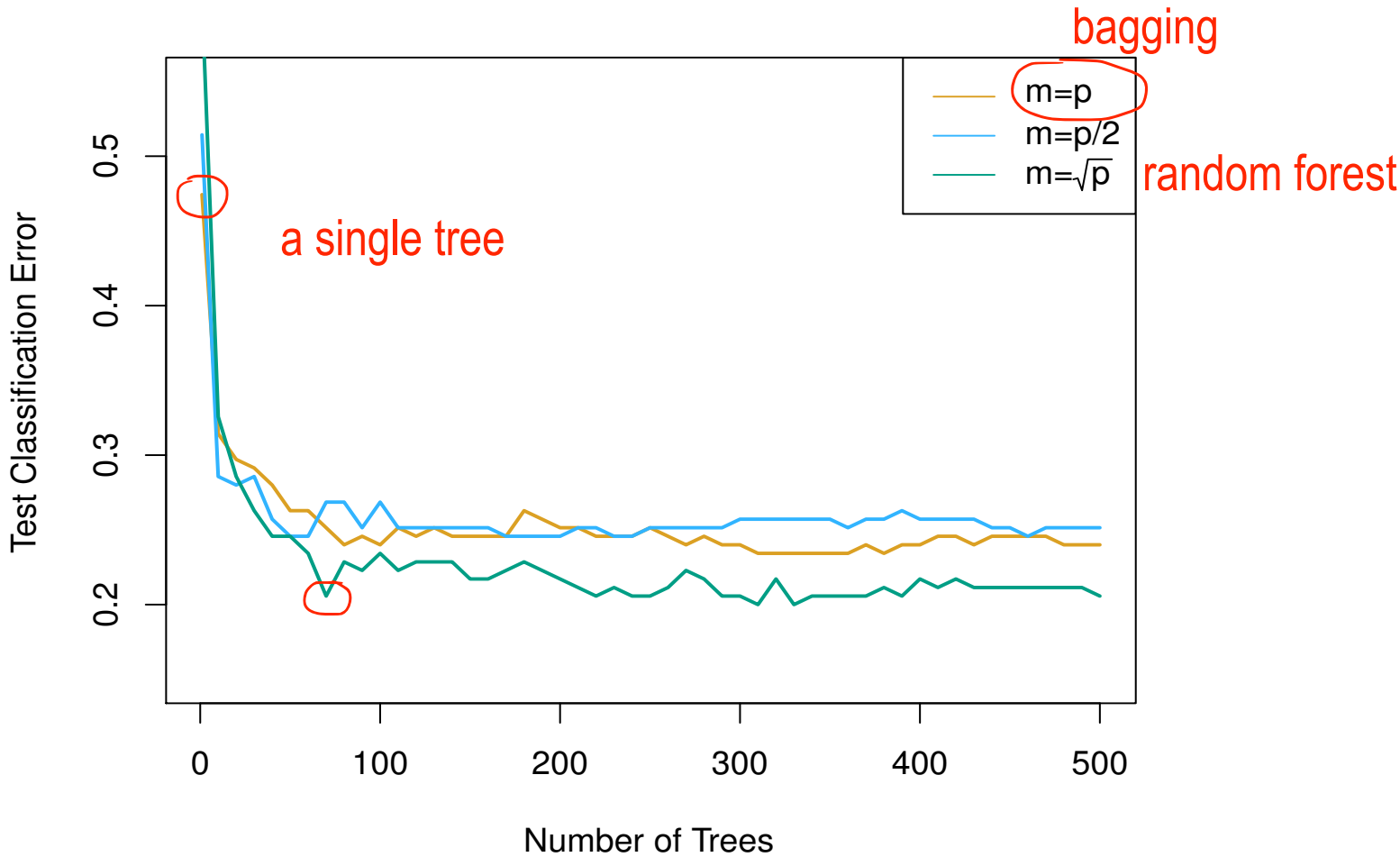
Heart Data

- Bagging and random forest results for the Heart data.
- The test error (black and orange) is shown as a function of B, the number of bootstrapped training sets used.
- Random forests were applied with $m = \sqrt{p}$.
- The dashed line indicates the test error resulting from a single classification tree.
- The green and blue traces show the OOB error, which in this case is considerably lower.

Gene expression data

- We applied random forests to a high-dimensional biological data set consisting of expression measurements of $4,718$ genes measured on tissue samples from 349 patients.
- There are around $20,000$ genes in humans, and individual genes have different levels of activity, or expression, in particular cells, tissues, and biological conditions.
- Each of the patient samples has a qualitative label with 15 different levels: either normal or one of 14 different types of cancer.
- We use random forests to predict cancer type based on the 500 genes that have the largest variance in the training set.
- We randomly divided the observations into a training and a test set, and applied random forests to the training set for three different values of the number of splitting variables m.

Gene expression data



Gene expression data

- Results from random forests for the fifteen-class gene expression data set with $p = 500$ predictors.
- The test error is displayed as a function of the number of trees. Each colored line corresponds to a different value of m, the number of predictors available for splitting at each interior tree node.
- Random forests ($m < p$) lead to an obvious improvement over bagging ($m = p$). A single classification tree has an error rate of 45.7%.

Summary and Remark

- Bagging
- Random Forests
- Read textbook Chapter 8, Chapter 15 and `R` code
- Do `R` lab