Bagging Trees and Random Forests

Wei Li

Syracuse University

Spring 2021

Bagging Trees

Random Forests

Interpretation of Random Forests

Random Forests and Overfitting

Bagging Trees

Additive Trees

Trees can be simple, but often lack of prediction accuracy.

One major problem with trees is their high variance. Often a small change in the data can result in a very different series of splits, making interpretation somewhat precarious.

- Bagging (Breiman 1996): Fit many large trees to bootstrap-resampled versions of the training data, and classify by a majority vote.
- Boosting (Freund & Shapire 1996): Fit many small trees sequentially to re-weighted versions of the training data. Classify by a weighted majority vote.

Bagging

Recall first the bagging procedure:

- Given data $Z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,
- Generate $Z^{*b} = \{(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)\} \sim \hat{P}_n, b = 1, \ldots, B$ where the empirical distribution $\hat{P}_n$: putting equal probability $1/n$ on each (x_i, y_i) (discrete)
- Obtain $\hat{f}^{*b}(x)$ using the bootstrap sample Z^{*b}, for $b = 1, \ldots, B$
- The Monte Carlo estimate of the bagging estimate

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

  - For classification, use hard classification (where $\hat{f}$ yields the class prediction) or soft classification (where $\hat{f}$ yields the probabilities estimates).
- Each individual trees: low bias, high variance.
- bagging trees is used to to reduce the variance (hopefully)?

OOB estimate

For bagged trees,

- we may use CV to estimate the number of trees B used (i.e., B bootstrap samples) selecting among different choices of B), or
- estimate the OOB error (as the test error estimate).

Using a ver large B will not lead to overfit

- may choose some B sufficiently large
- use the OOB error as the estimate of the prediction error.

OOB estimate

- Obtain B bootstrap samples, and fitted $\hat{f}^{*b}, b = 1, \ldots, B$.
- For i-th observation, make prediction using each of the trees in which this observation was OOB (yielding around $B/3$ predictions for the i-th observation)
- Take average of all the predicted values for this i-th observation. This leads to a single **OOB prediction** for i-th observation. for each observation $z_i = (x_i, y_i)$, its **out-of-bag estimate** is (if regression)

$$\hat{f}_{\text{oob}}\ (x_i) = \sum_{b \in B_i^{out}} f^{*b}\,(x_i)\,/|B_i^{out}|$$

where B_i^{out} is the index of the bootstrap samples in which z_i did not appear. If classification, take the majority vote from these B_i^{out} many predictions.

- Obtain an OOB predictions for each of n original data points.
- Compute the overall OOB error rate: MSE (if regression), or classification error rate (if classification).

With B sufficiently large, overall OOB error is virtually equivalent to leave-one-out cross-validation error.

Variable importance measure

One can obtain an overall summary of the importance of each predictor using the RSS (for bagging regression trees) or the Gini index (for bagging classification trees).

- bagging regression trees, the total amount that the RSS is decreased due to splits over a given predictor, averaged over all B trees.
- bagging classification trees, the total amount that the Gini index is decreased by splits over a given predictor, averaged over all B trees.

say for each X_j:

- for each tree, find all splits on variable X_j
- add up total amount of RSS (or Gini index) decreases from these splits
- average over all B tress
- measure importance of X_j (larger the better)

For a single regression tree T:

The **squared relative importance** of variable X_ℓ is the sum of such squared improvements over all internal nodes for which it was chosen as the splitting variable. Let

$$\mathcal{I}_\ell^2(T) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 I(v(t) = \ell)$$

be a measure of relevance for each predictor variable X_ℓ. The sum is over the $J - 1$ internal nodes of the tree.

- At each node t, $X_{v(t)}$ is used to partition the region associated with that node into two subregions; within each a separate constant is fit to the response values.
- The particular variable chosen is the one that gives maximal estimated improvement $\hat{\imath}_t^2$ in squared error risk over that for a constant fit over the entire region.

For additive tree expansions $f_B(x) = \frac{1}{B}\sum_{b=1}^{b} T\left(x;\Theta_b\right)$ (or e.g., $f_M(x) = \sum_{m=1}^{M} T\left(x;\Theta_m\right)$ based on boosting); it is simply averaged over the trees

$$\mathcal{I}_\ell^2 = \frac{1}{B}\sum_{b=1}^{B}\mathcal{I}_\ell^2\left(T_b\right)$$

- Above is called **squared relevance** or **actual relevance**;
- customary to assign the largest a value of 100 and then scale the others accordingly.

The simple structure in the model can be lost due to bagging

- A bagged tree is no longer a tree.
- The bagged estimate is not easy to interpret.
- Under $0 - 1$ loss for classification, bagging may not help due to the nonadditivity of bias and variance. But in general, bagging *independent* and *good* weak learners is a good idea.
- The bagged trees however are not independent of each other.

Random Forests

Random Forests

Random forests is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them.

For bagged trees, the expectation of an average of B trees is the same as the expectation of any of them.

The only improvement is through variance reduction.

An average of B i.i.d. random variables, each with variance σ^2, has variance $\frac{1}{B}\sigma^2$.

If the variables are simply i.d. (identically distributed) with positive pairwise correlation ρ, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

The random forests improve the variance reduction of bagging by

- reducing the correlation between the trees, without increasing the variance too much.
  - achieved in the tree-growing process through random selection of the input variables.

Algorithm

```
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
        recursively 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 ($m \leq p$).
         ii. Pick the best variable/split-point among the $m$ candidates.
        iii. Split the node into two child nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
```

To make a prediction at a new point x :

- Regression: $\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
- Classification: As regression, average over the predicted probabilities over all B trees. Or, let $\hat{h}_b(x)$ be the class prediction of the b th random-forest tree. Then $\hat{h}_{\mathrm{rf}}^{B}(x) =$ majority vote $\left\{\hat{h}_b(x)\right\}_1^B$

Interpretation of Random Forests

Variable importance measure: random forests

To measure the prediction strength of each variable, random forests also use the OOB samples to construct a different variable importance measure.

The idea is to use randomization to voids the effect of the j-th variable:

When the b-th tree is grown, the OOB observations are passed down the tree, and the OOB prediction accuracy is recorded. Then the values for the j th variable are *randomly permuted* in the OOB observations of every tree, and the accuracy is again computed. Then the decrease in accuracy as a result of this permuting is *averaged over all trees*. These are typically expressed as a percent of the maximal decrease in accuracy.

Proximity plot

In growing a random forest, an $n \times n$ **proximity matrix** is accumulated for the training data.

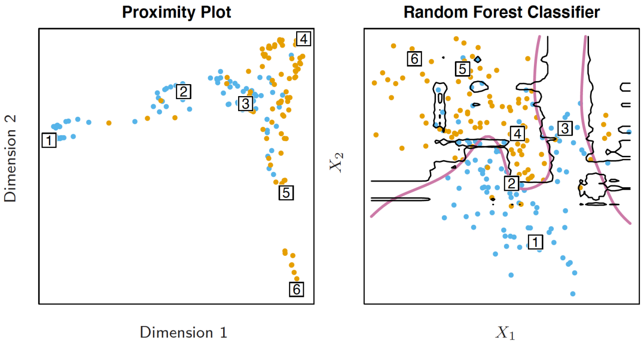
- For every tree, any pair of OOB observations sharing a terminal node has their proximity increased by one.
- The matrix is then represented using multidimensional scaling.

Multidimensional scaling seeks values $z_1, z_2, \ldots, z_n \in \mathbb{R}^k$ (say $k = 2, 3$) to minimize the so-called stress function

$$S_M\left(z_1, z_2, \ldots, z_n\right) = \sum_{i \neq i'} \left(d_{ii'} - \|z_i - z_{i'}\|\right)^2$$

where $d_{ii'}$ is some dissimilarity measure.

Example: when X is two-dimensional, using $k = 2$



ESL: Fig 15.6

Random Forests and Overfitting

Random Forests and Overfitting

Use a sufficiently large B for the error rate to settle down.

How about m? Focus regression problems:

- After B such trees $\{T(x;\Theta_b)\}_1^B$ are grown
    - Θ_b characterizes the b-th random forest tree in terms of split variables, cutpoints at each node, and terminal-node values

random forest (regression) predictor is

$$\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$$

So

$$\lim_{B\to\infty} \hat{f}_{\mathrm{rf}}^B(x) = \hat{f}_{\mathrm{rf}}(x) := \mathrm{E}_{\Theta|Z} T(X;\Theta(Z)),$$

here $\Theta|Z$ is conditional on the training data Z.

Effect of $m < p$: bias

The bias of a random forest is the same as the bias of any of the individual sampled trees $T(x; \Theta(\mathbf{Z}))$,

$$\begin{aligned}\text{Bias}(x) &= f(x) - \mathrm{E}_{\mathbf{Z}} \hat{f}_{\text{rf}}(x)\\ &= f(x) - \mathrm{E}_{\mathbf{Z}} \mathrm{E}_{\Theta|\mathbf{Z}} T(x; \Theta(\mathbf{Z}))\end{aligned}$$

- General trend is as m decreases, the bias increases.

Effect of $m < p$: variance

$$var(\hat{f}_{\mathrm{rf}}(x)) \approx \rho(x)\sigma^2(x) + \lim_{B\to\infty} \frac{1-\rho(x)}{B}\rho^2(x) \approx \rho(x)\sigma^2(x)$$

- $\rho(x)$ is

  $$\rho(x) = \operatorname{corr}\left[T\left(x;\Theta_1(Z)\right), T\left(x;\Theta_2(Z)\right)\right]$$

  where $\Theta_1(Z)$ and $\Theta_2(Z)$ are a randomly drawn pair of random forest trees grown to the randomly sampled Z.
- $\sigma^2(x)$ is the sampling variance of any single randomly drawn tree,

  $$\sigma^2(x) = \operatorname{Var} T(x;\Theta(Z))$$

The variability averaged-over in $\rho(x)$ and $\sigma^2(x)$ is due both to

- variability conditional on Z: due to the bootstrap sampling and feature sampling at each split
- sampling variability of Z itself.

On $\rho(x)$:

$$\rho(x) = \operatorname{corr}\left[T\left(x;\Theta_1(Z)\right), T\left(x;\Theta_2(Z)\right)\right]$$

The correlation $\rho(x)$ between pairs of trees decreases as m decreases: pairs of tree predictions at x for different training sets Z are likely to be less similar if they do not use the same splitting variables.

On $\sigma^2(x)$:

$$\sigma^2(x) = \mathrm{Var}_{\Theta,\mathbf{Z}}\, T(x;\Theta(\mathbf{Z})) = \mathrm{Var}_{\mathbf{Z}}\, \mathrm{E}_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z})) + \mathrm{E}_{\mathbf{Z}}\, \mathrm{Var}_{\Theta|\mathbf{Z}}\, T(x;\Theta(\mathbf{Z}))$$
$$\text{Total Variance} = \mathrm{Var}_{\mathbf{Z}}\, \hat{f}_{\mathrm{rf}}(x) \quad + \quad \text{within-Z Variance}$$

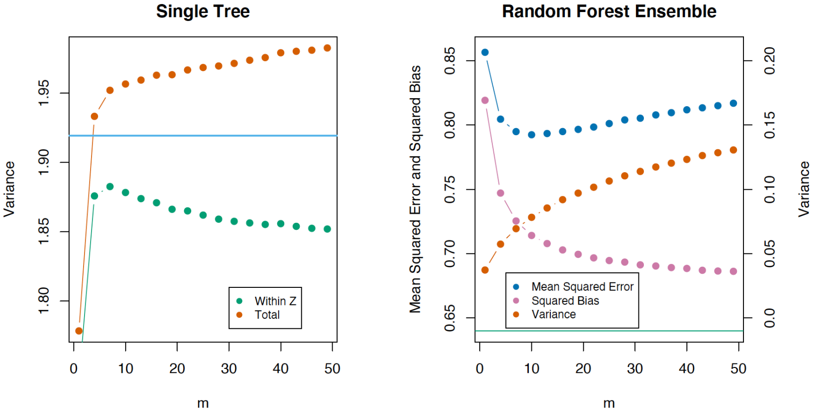
- The first term: smaller m tends to increase the bias of $\hat{f}_{rf}$ tends to decreases variance of $\hat{f}_{rf}$.
- The second term: smaller m tends to induce more variability in $\mathrm{Var}_{\Theta|\mathbf{Z}}\, T(x;\Theta(\mathbf{Z}))$.
- overall, $\sigma^2(x)$ does not change much as m varying.

The overall effect of $\sigma^2(x)\rho(x)$ decreases as m decreases.

The improvements made by random forests are solely a result of variance reduction through the decorrrelation.

Let $x \in R^p$: using m relatively smaller than p

- For classification, the default value for m is $\lfloor \sqrt{p} \rfloor$ and the minimum node size is one.
- For regression, the default value for m is $\lfloor p/3 \rfloor$ and the minimum node size is five.



Single Tree
Variance
1.80
1.85
1.90
1.95
0
10
20
30
40
50
m
● Within Z
● Total
Random Forest Ensemble
Mean Squared Error and Squared Bias
0.65
0.70
0.75
0.80
0.85
Variance
0.0
0.05
0.10
0.15
0.20
● Mean Squared Error
● Squared Bias
● Variance

ESL: Fig 15.10

Sub-sampling

Sub-sampling (without replacement) is an effective alternative to bagging.

Each tree is constructed on a subset of a_n examples picked within the original sample (without replacement) and forms estimate. Then taking average of these estimates is approximately equivalent to bagging, while using smaller fractions of the original sample reduces the variance even further (through decorrelation).

- the median forest is consistent if $a_n = o(n)$ (Scornet, 2015)
- asymptotic normality holds if $a_n = o(\sqrt{n})$ (Mentch and Hooker, 2015).