Bagged Trees and Random Forests

Wei Li

Syracuse University

Spring 2024

OVERVIEW

Bagged Trees

Random Forests

Interpretation of Random Forests

Random Forests and Bias-Variance tradeoff

Bagged Trees

Additive Trees

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

One major problem with trees is their high variance. A small change in the data can result in a very different series of splits.

- Bagging or random forests (Breiman 1996): Fit many trees to bootstrap-resampled versions of the training data and average them.
- Boosting (Freund & Shapire 1996): Fit many small trees sequentially to re-weighted versions of the training data.

Bagging

Recall first the bagging procedure:

- Given data $\mathbf{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$,
- Generate $\mathbf{Z}^{*(b)} = \{(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)\} \sim \hat{\mathbb{P}}_n, b = 1, \ldots, B$ where the empirical distribution $\hat{\mathbb{P}}_n$
- Obtain $\hat{f}^{*(b)}(x)$ using the bootstrap sample $\mathbf{Z}^{*(b)}$, 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.
- bagged trees is used to to reduce the variance
  - better only when weakly correlated (random forests)

Choice of B

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

Value of B: one may use CV to estimate the number of trees B used (i.e., B bootstrap samples) selecting among different choices of B.

- the number of trees B is not a critical parameter
- a too small B produces inaccurate estimate
- pick sufficiently large B
  - using a ver large B will not lead to overfit.

Estimate prediction error

One can compute the so-called OOB error as the estimate of the prediction error:

- 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. For each observation $z_i = (x_i, y_i)$, its **OOB (out-of-bag estimate)** is (if regression)

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

where C^{-i} is the index of the bootstrap samples in which z_i did not appear.

  - If classification, take the majority vote from these C^{-i} many predictions.

OOB (out-of-bag estimate) for i-th observation:

$$\hat{f}_{\text{oob}}\ (x_i) = \sum_{b \in C^{-i}} \hat{f}^{*(b)}\ (x_i)\ /|C^{-i}|$$

- 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).

$$\frac{1}{n} \sum_{i=1}^{n} L\Big(y_i, \hat{f}_{\text{oob}}\ (x_i)\Big)$$

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

The OOB approach is convenient when performing bagging on large data sets for which cross-validation would be computationally expensive.

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_j is the sum of such squared improvements over all internal nodes for which it was chosen as the splitting variable. Let

$$\mathcal{I}_j^2(T) = \sum_{t=1}^{|T|-1} \hat{\imath}_t^2 I(v(t) = j)$$

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

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

squared relevance

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

$$\mathcal{I}_j^2 = \frac{1}{B}\sum_{b=1}^{B} \mathcal{I}_j^2(T_b)$$

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

Note: For bagged tree estimate, there is a different way to construct variable importance measure using the OOB error estimate.

Limitations of bagged trees

- A bagged tree is no longer a tree.
- The bagged estimate is not easy to interpret.
- For regression, bagged trees of high variance usually help improve MSE.
- For classification ($0-1$ loss), bagging *independent* and *good but weak* learners usually help improve accuracy.
  - The bagged trees however are not independent of each other.

This motivates the de-correlating bagged trees (random forests).

Random Forests

Random Forests

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

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

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 predictors.

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.

Algorithm (random forests)

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 x :

- Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$
- Classification:
  - 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, so $\hat{h}_{\text{rf}}^B(x) =$ majority vote $\{\hat{h}_b(x)\}_1^B$

Interpretation of Random Forests

Variable importance measure: bagged tree

Besides **squared relevance**, bagged tree (random forests) can also use the OOB samples to construct a variable importance measure.

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

- for the b-th tree, the OOB observations are passed down the tree, compute OOB prediction accuracy
  - *randomly permute* the values for the jth variable in the OOB observations of the tree, compute OOB prediction accuracy
  - find the decrease in accuracy due to the permutation for the tree
- average all the above decreases *over all trees*

Local importance measure

For each observation and predictor, the (i, j) **local importance measure** can be computed:

- for observation i, take all of the trees of which i is OOB
  - aggregate the predictions for this observation based on these trees
  - calculate the OOB accuracy based on these trees
- for the variable j. Permute the values of j for the OOB observations of every tree not containing i (this should alter the value for the j variable for observation i)
  - do the same as in the above step to obtain the OOB accuracy for the observation i, that is aggregate the predictions for this (j-variate permuted) observation based on these trees, then aggregate the prediction, and obtain OOB accuracy.
- subtracting the permuted-j OOB accuracy from the non-permuted OOB accuracy gives the (i, j) local importance measure.
- averaging (i, j) local importance measure over all i gives the importance measure for variable j.

Proximity plot

In growing a random forest, an $n \times n$ **proximity matrix** is constructed 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.

The **dissimilarity matrix** D then is derived as the opposite of the *proximity matrix*.

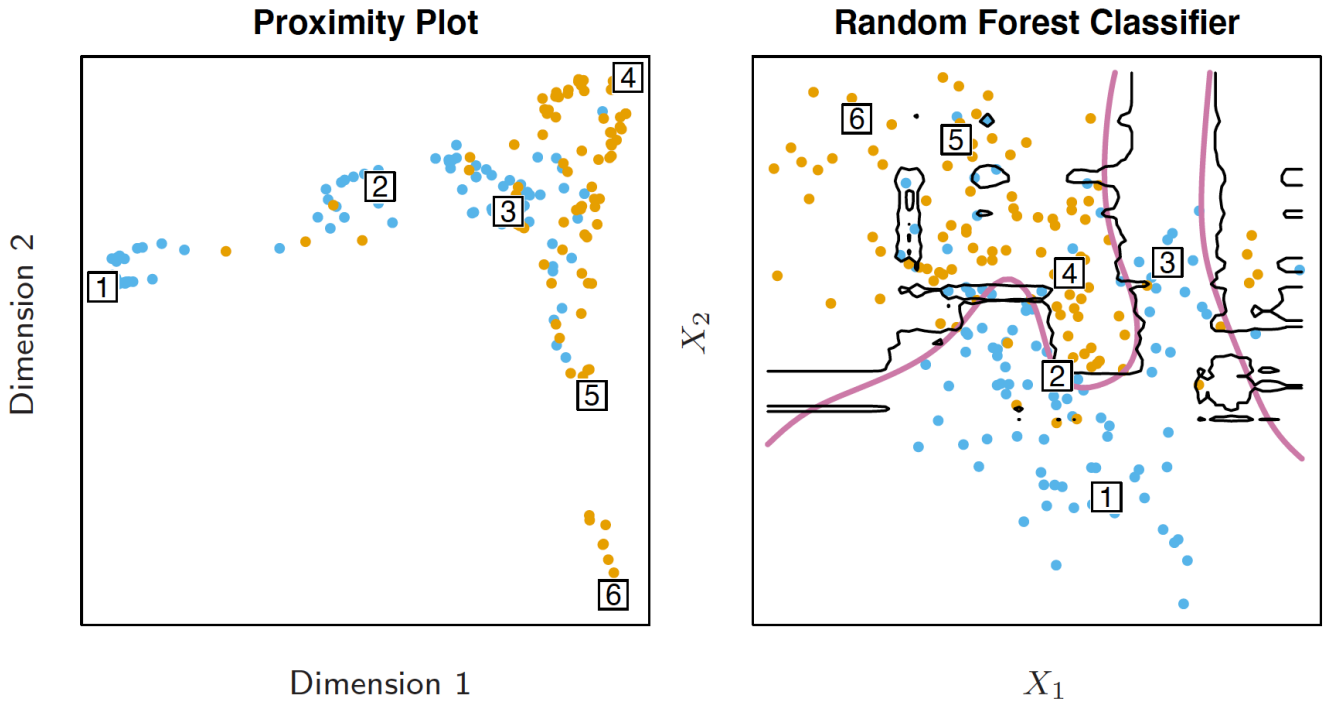
Apply **Multidimensional scaling** to seek values $z_1, z_2, \ldots, z_n \in \mathbb{R}^k$ (say $k = 2, 3$) minimizing the the stress function

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

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

Observations that are more similar (i.e., share more common features or patterns in how they are classified or regressed by the trees, i.e., shaping the same prediction at the terminal node) will have lower dissimilarity values.

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



- Points in pure regions class-wise map to the extremities (terminals) of the star;
- Points nearer the decision boundaries map nearer the center.

ESL: Fig 15.6

Average Dependence Plots

Consider the subvector $X_{\mathcal{S}}$ of $\ell < p$ of the predictor variables $X = (X_1, X_2, \ldots, X_p)$, indexed by $\mathcal{S} \subset \{1, 2, \ldots, p\}$. Let $\mathcal{C}$ be the complement set, with $\mathcal{S} \cup \mathcal{C} = \{1, 2, \ldots, p\}$.

The **average or partial dependence** of $f(X) = f(X_{\mathcal{S}}, X_{\mathcal{C}})$ on $X_{\mathcal{S}}$ is

$$f_{\mathcal{S}}(X_{\mathcal{S}}) = \mathrm{E}_{X_{\mathcal{C}}} f(X_{\mathcal{S}}, X_{\mathcal{C}})$$

They can be estimated by

$$\bar{f}_{\mathcal{S}}(X_{\mathcal{S}}) = \frac{1}{n} \sum_{i=1}^{n} f(X_{\mathcal{S}}, x_{i\mathcal{C}})$$

where $\{x_{1\mathcal{C}}, x_{2\mathcal{C}}, \ldots, x_{n\mathcal{C}}\}$ are the values of $X_{\mathcal{C}}$ occurring in the training data.

Note: this is not the same as $\tilde{f}_{\mathcal{S}}(X_{\mathcal{S}}) = \mathrm{E}(f(X_{\mathcal{S}}, X_{\mathcal{C}}) \mid X_{\mathcal{S}})$.

For K-class classification, we may plot partial dependence plot for the function $\log(p_{k_1}(x)/p_{k_2}(x))$.

Random Forests and Bias-Variance tradeoff

Random Forests and Bias-Variance tradeoff

Use a sufficiently large B.

How about m? Focus on regression problems.

- Before sampling, there is sampling variability of $\mathbf{Z}$ itself.
- Given $\mathbf{Z}$, B such trees $\{T(x;\Theta_b)\}_1^B$ are grown
  - Θ_b characterizes the b-th random forest tree, involving
    - bootstrap sampling
    - choosing split variables, cutpoints at each node, and terminal-node values

Random forest (regression) predictor is

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

So

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

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

Effect of $m < p$: bias

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

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

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

Effect of $m < p$: variance

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

- $\rho(x)$ is

$$\rho(x) = \text{corr}\left(T\left(x;\Theta_1(\mathbf{Z})\right), T\left(x;\Theta_2(\mathbf{Z})\right)\right)$$

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

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

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

- variability conditional on $\mathbf{Z}$
- sampling variability of $\mathbf{Z}$

On $\sigma^2(x)$:

$$\sigma^2(x) = \mathrm{Var}_{\Theta,\mathbf{Z}}\, T(x;\Theta(\mathbf{Z})) = \mathrm{Var}_{\mathbf{Z}}\, \mathrm{E}_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z})) + \mathrm{E}_{\mathbf{Z}}\, \mathrm{Var}_{\Theta|\mathbf{Z}}\, T(x;\Theta(\mathbf{Z}))$$

Total Variance = sampling variance of $\hat{f}_{\mathrm{rf}}(x)$ + within-Z variance of a tree

- The first term: smaller m 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, $\sigma^2(x)$ is insensitive to m.

On $\rho(x)$:

$$\begin{aligned}\rho(x) &= \operatorname{corr}\left(T\left(x;\Theta_1(\mathbf{Z})\right), T\left(x;\Theta_2(\mathbf{Z})\right)\right) \\ &= \frac{0 + \operatorname{cov}_{\mathbf{Z}}\left(E_{\Theta_1|\mathbf{Z}}\left[T\left(x;\Theta_1(\mathbf{Z})\right], E_{\Theta_2|\mathbf{Z}}\left[T\left(x;\Theta_2(\mathbf{Z})\right]\right)\right.\right.}{\operatorname{Var}_{\mathbf{Z}} E_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z})) + E_{\mathbf{Z}} \operatorname{Var}_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z}))}. \\ &= \frac{\operatorname{Var}_{\mathbf{Z}} E_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z}))}{\operatorname{Var}_{\mathbf{Z}} E_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z})) + E_{\mathbf{Z}} \operatorname{Var}_{\Theta|\mathbf{Z}} T(x;\Theta(\mathbf{Z}))}.\end{aligned}$$

- numerator: sampling variability of the ideal random forest, tends to increase in m
- denominator: within-sample variability of a tree, not sensitive to m.

The correlation $\rho(x)$ between pairs of trees decreases as m decreases: pairs of tree predictions at x for *different* $\mathbf{Z}$ are less similar if m is small.

Note: the 0 is the term $E_{\mathbf{Z}}\left[\operatorname{cov}_{\Theta|\mathbf{Z}}\left(T\left(x;\Theta_1(\mathbf{Z})\right), T\left(x;\Theta_2(\mathbf{Z})\right)\right)\right]$

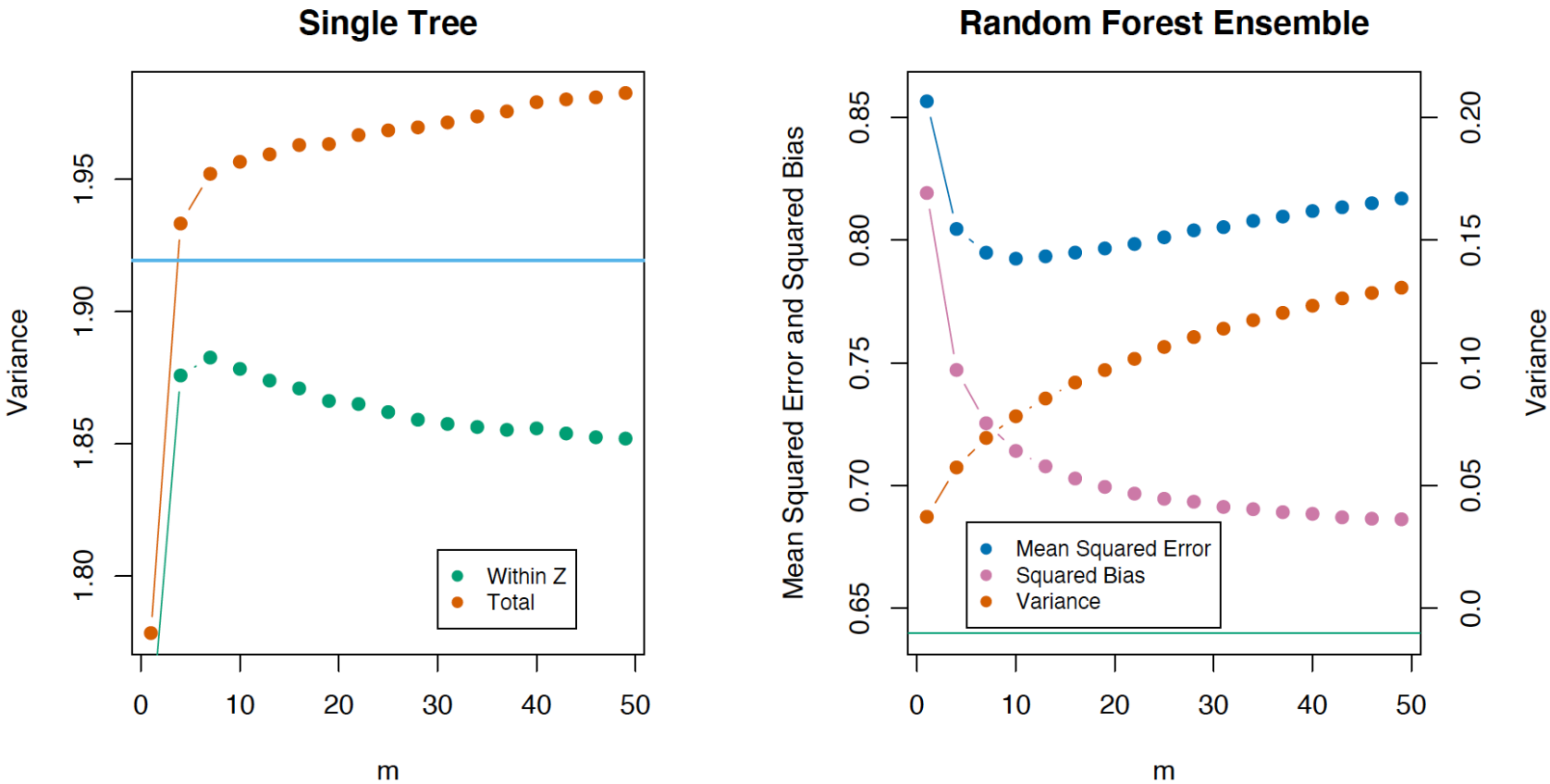
The overall effect of $\sigma^2(x)\rho(x)$ increases in m.

- $m \uparrow$, bias $\downarrow$
- $m \uparrow$, variance $\rho(x)\sigma^2(x) \uparrow$
  - $\sigma^2(x)$: insensitive to m
    - $m \uparrow$, sampling variance $\uparrow$
    - $m \uparrow$, within-sample variance $\downarrow$
  - $\rho(x)$: $m \uparrow$, $\rho(x) \uparrow$

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

Practical choice: for $x \in \mathbb{R}^p$: 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.



ESL: Fig 15.10

The left panel shows the average variance of a *single* random forest tree, as a function of m. The horizontal line is the average variance of a single *fully* grown tree (with-out bootstrap sampling).

The right panel: MSE, squared bias and variance of the ensemble, as a function of m. The horizontal line is the average squared-bias of a fully grown tree.

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).