S&DS 365 / 565

Data Mining and Machine Learning

Trees and Ensemble Methods

Yale

Bias vs Variance

Deep trees (like k-NN with small k) have low bias, but suffer from high variance. One remedy is to prune the tree.

In today's lecture, we consider a different approach, through the use of ensembles.

Ensemble: Intuition

A good analogy for thinking about ensemble methods is to consider the term “**Wisdom of the Crowds**”.

Ensemble: Intuition

A good analogy for thinking about ensemble methods is to consider the term "**Wisdom of the Crowds**".

- Tree — One Person
- Forest — Crowd

Smoothing the tree, improve variance
not affect bias too much

Ensemble: Intuition (Groupthink)

You don't want everybody think the same

It is not enough to just have a crowd (forest) – if everyone thinks the same, then you get groupthink and the value of a crowd $\equiv$ one person. $\equiv$ same bias, same variance

Ensemble: Intuition (Groupthink)

It is not enough to just have a crowd (forest) – if everyone thinks the same, then you get groupthink and the value of a crowd $\equiv$ one person. The people's opinions in the crowd should be uncorrelated. We shall make this more explicit later on.

independent

Trees vs. other methods

Decision trees are similar in spirit to k-nearest neighbors.

Trees vs. other methods

Decision trees are similar in spirit to k-nearest neighbors.

- Both produce simple predictions (averages/maximally occurring) based on “neighborhoods” in the predictor space.

Trees vs. other methods

Decision trees are similar in spirit to k-nearest neighbors.

regression ↑ classification ↑

- Both produce simple predictions (averages/maximally occurring) based on "neighborhoods" in the predictor space.
- However, decision trees use adaptive neighborhoods.

Trees vs. other methods

Recall that linear regression fits models of the form an additive model

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Trees vs. other methods

Recall that linear regression fits models of the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Regression trees are like fitting linear regression models with a bunch of indicators!

$$f(X) = \sum_{j=1}^{J} \beta_j \mathbb{1}\{X \in R_j\}$$

number of regions

adaptive feature

build J new features

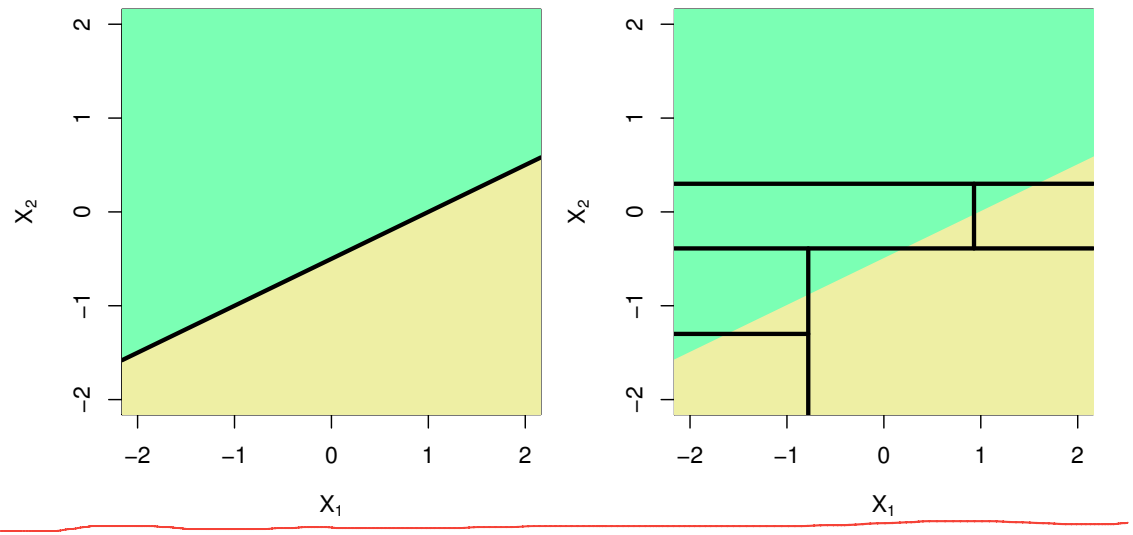
Learn params β_j

Trees vs. other methods

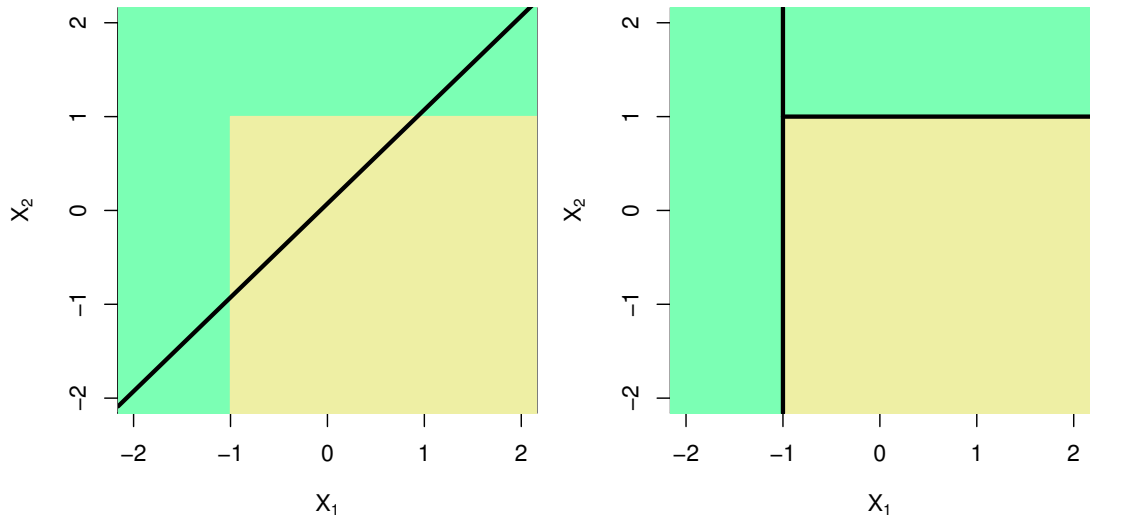
Are trees always better than linear methods?

No

tree bad



tree better



Summary

- trees are intuitive
- prediction rules easy to explain, interpret
- trees are sensitive to underlying data
- trees produce non-smooth prediction surfaces (esp. problematic for regression)
- prediction accuracy can be so-so

Visualizations

http://www.r2d3.us/
visual-intro-to-machine-learning-part-1/

Issues with trees

- Instability. Full trees generally have high variance. As data change, tree topology can change dramatically, making interpretation difficult.
- Lack of smoothness. The splits lead to a “jagged” decision boundary. More of a problem for regression than classification.
- Difficulty capturing additive structure. If actual model is additive, this may not be captured by the tree with limited data.

Ensemble methods

Ensemble methods pool together multiple models to arrive at more reliable predictions.

Ensemble methods

Ensemble methods pool together multiple models to arrive at more reliable predictions.

basically using average

- bagging
- random forests
- boosting

Bagging

Bagging or (bootstrap aggregation) exploits the idea that averaging reduces variability.

Bagging

Bagging or (bootstrap aggregation) exploits the idea that averaging reduces variability.

e.g. given $Y_1, Y_2, \ldots,$ *iid* with mean μ and variance σ^2:

- suppose we want to estimate μ
- consider estimators Y_1 and $\bar{Y}$ (both unbiased)
- $Var(\bar{Y}) = \sigma^2/n < Var(Y_1)$ $= \sigma^2$ bad

good

both $\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_1] = \mu$ unbiased estimate

Bootstrap

Suppose instead $Y_1, Y_2, \ldots, Y_B$ are only *id*, not independent, but with correlation ρ. Then,

$$Var(\bar{Y}) = \frac{\sigma^2}{B}(1-\rho) + \rho\sigma^2$$

in extreme case, $\rho = 1$ $Var(\bar{Y}) = \sigma^2 = Var(Y_i)$ listening to person 1 is same as person i, ..., every one

Bagging

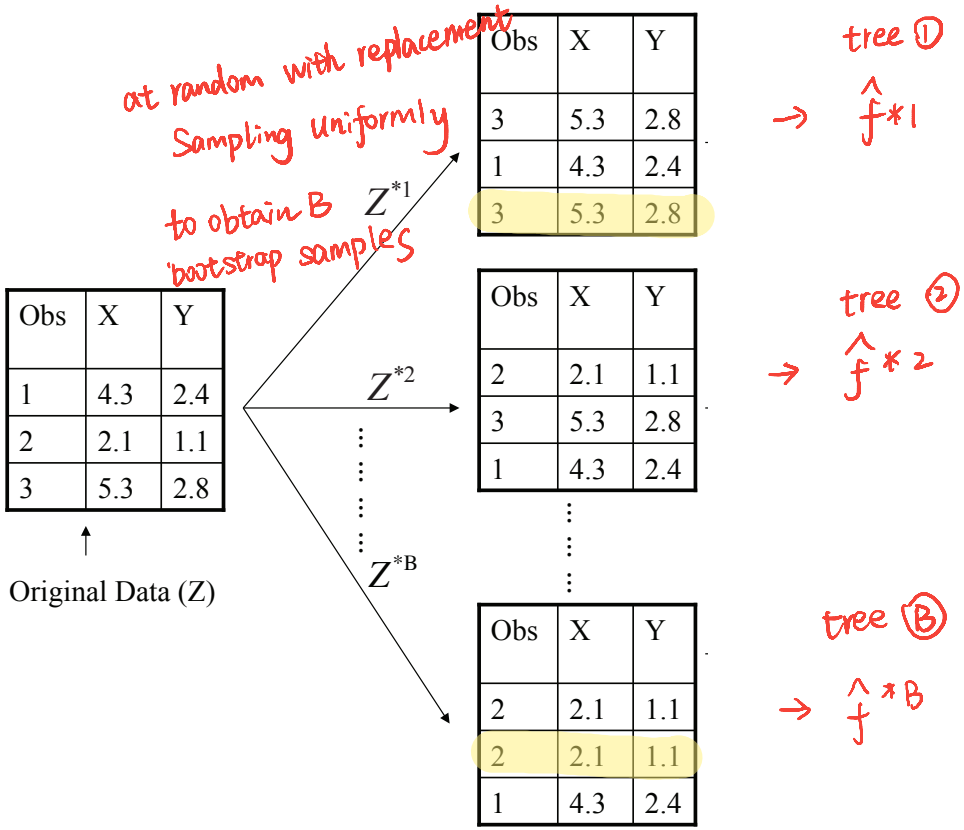
Regression trees: For example, if we had multiple training sets, growing multiple trees, then take an average.

Bagging

Regression trees: For example, if we had multiple training sets, growing multiple trees, then take an average.

independent traing sets

In practice, we only has 1 traning set



bootstrap 自手起家，自力更生

pull oneself up by one's own bootstrap

用自己的靴子带子把自己拉起来

Bagging

Create B bootstrap samples, grow tree (without pruning) using each $\widehat{f}^{*1}, \widehat{f}^{*2}, \ldots \widehat{f}^{*B}$. For prediction at x, we take an average:

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

Bagging

Regression trees

Create B bootstrap samples, grow tree (without pruning) using each $\widehat{f}^{*1}, \widehat{f}^{*2}, \ldots \widehat{f}^{*B}$. For prediction at x, we take an average:

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

Classification trees: For prediction at x, $\widehat{f}_{bag}(x)$ is decided by *majority vote*.

Out-of-Bag error estimation

We can show that each bagged tree uses about 2/3 of all observations (with repeats).

sampling

Out-of-Bag error estimation

We can show that each bagged tree uses about 2/3 of all observations (with repeats).

$$1 - \left(1 - \frac{1}{n}\right)^{n} \to 1 - \frac{1}{e} \approx 0.6321 \approx \frac{2}{3}$$

n different trials

$\uparrow$ probability of picking a data points in n data points

The probability of pick a data point in n samples (bags) is $\frac{2}{3}$

$\Rightarrow$ proportion of total data in a bag is $\frac{2}{3}$

$\Rightarrow$ $\frac{1}{3}$ of total data remained unuse

Out-of-Bag error estimation

We can show that each bagged tree uses about 2/3 of all observations (with repeats).

$$1 - \left(1 - \frac{1}{n}\right)^n \to 1 - \frac{1}{e} \approx 0.6321$$

The remaining data (*out-of-bag* (OOB) observations) can be put to good use.

Out-of-Bag error estimation

	Bagging iteration					OOB Est.
Obs	1	2	3	...	B	
1	OOB	train	train	...	train	$\widehat{y}_1$
2	train	OOB	train	...	train	$\widehat{y}_2$
3	train	train	OOB	...	train	$\widehat{y}_3$
4	OOB	train	train	...	OOB	$\widehat{y}_4$
...	...	...	...	...	...	...
n	train	train	OOB	...	train	$\widehat{y}_n$

Out-of-Bag error estimation

	Bagging iteration					OOB Est.
Obs	1	2	3	...	B	
1	OOB	train	train	...	train	$\widehat{y}_1$
2	train	OOB	train	...	train	$\widehat{y}_2$
3	train	train	OOB	...	train	$\widehat{y}_3$
4	OOB	train	train	...	OOB	$\widehat{y}_4$
...	...	...	...	...	...	...
n	train	train	OOB	...	train	$\widehat{y}_n$

- For each bagged tree, we can make predictions for the OOB observations.
- At the end, we can aggregate over (take average) all predictions for the i-th observation to arrive at a OOB prediction $\widehat{y}_i$.
- We can compute prediction error based on these OOB predictions $\widehat{y}_1, \ldots, \widehat{y}_n$. (OOB observ. are like test data)

Variable importance

While bagging improves upon the predictive ability of trees, it kills off the interpretability of the model.

reduce variance

coz each bag tree become different (but you want the differences coz you want them independent)

Variable importance

While bagging improves upon the predictive ability of trees, it kills off the interpretability of the model.

A good tool for interpreting a bagged tree model (and other tree-based ensemble methods like forests) is the **variable importance measure**.

Variable importance

While bagging improves upon the predictive ability of trees, it kills off the interpretability of the model.

A good tool for interpreting a bagged tree model (and other tree-based ensemble methods like forests) is the **variable importance measure**.

Variable importance can be measured by the amount that the RSS (or Gini index) is reduced due to splits over a given predictor, averaged over all B trees.

for regression tree

for classification tree

X_k

Random Forests

Similar to bagging, but takes the averaging idea even further – averaging uncorrelated things decreases the error!

Random Forests

Similar to bagging, but takes the averaging idea even further – averaging uncorrelated things decreases the error!

Still use bootstrap samples, but only use m out of p predictors at each split, *chosen randomly*

m often very small eg. 3, 5

thus decrease dependence of each tree

Bagging and Random Forests

Ideas.

good: maintain low bias, but decrease variance

- Grow many trees and average their predictions
- Trees are grown deep, to have low bias, but high variance

key:

- To "decorrelate" the predictors, each tree is

For randomforest

  - grown on a bootstrap sample of the data
  - grown with random subsets of the predictors at each split

- Tree growing can be done in parallel

Split decision rule can be various: $X > 3$ or linear function, etc

Simple decision rule can help tree smooth

Random Forests Algorithm

1. For $b = 1$ to B: grow B trees
   - **(a)** Draw a bootstrap sample Z^* of size n from the training data with replacement
   - **(b)** Grow a random-forest tree T_b to the bootstrapped data, recursively repeating following steps, until minimum node size reached:

difference between RF & bootstrapping: this step help decorrelate predictors

     - **i.** Select m variables at random from the p variables
     - **ii.** Pick the best variable/split-point among the m
     - **iii.** Split the node into two children nodes
2. Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.

To make a prediction at a new point x:

Regression: Average $\widehat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Classification: Majority vote of the individual trees

Random forests–recommended parameters

$\lfloor x \rfloor$: floor of x. the greatest integer $\leq x$

$\lfloor 3.4 \rfloor = 3$

p: # of predictors

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

MSRP: Manufacturer's Suggested Randomization Parameters

Out of bag (OOB) prediction

As before, can use out-of-bag (OOB) samples:

- For each observation $z_i = (x_i, y_i)$, construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which z_i did not appear
- Thus, cross-validation can be performed "along the way"

The chance a sample x_i does not appear in a bootstrap sample is

$$\left(1 - \frac{1}{n}\right)^n \longrightarrow \frac{1}{e} \approx 0.368$$

Out of bag (OOB) prediction

Plug in << Elements of statistical Learning >>

OOB Error matches test error pretty well

means OOB Error is a good surrogate for test Error

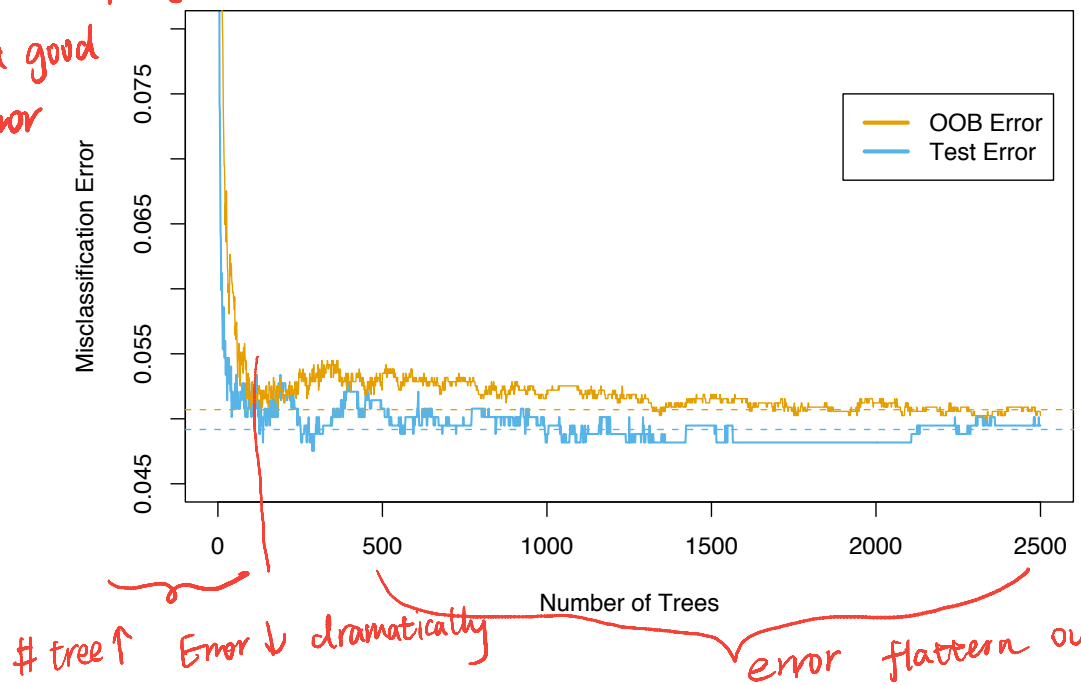
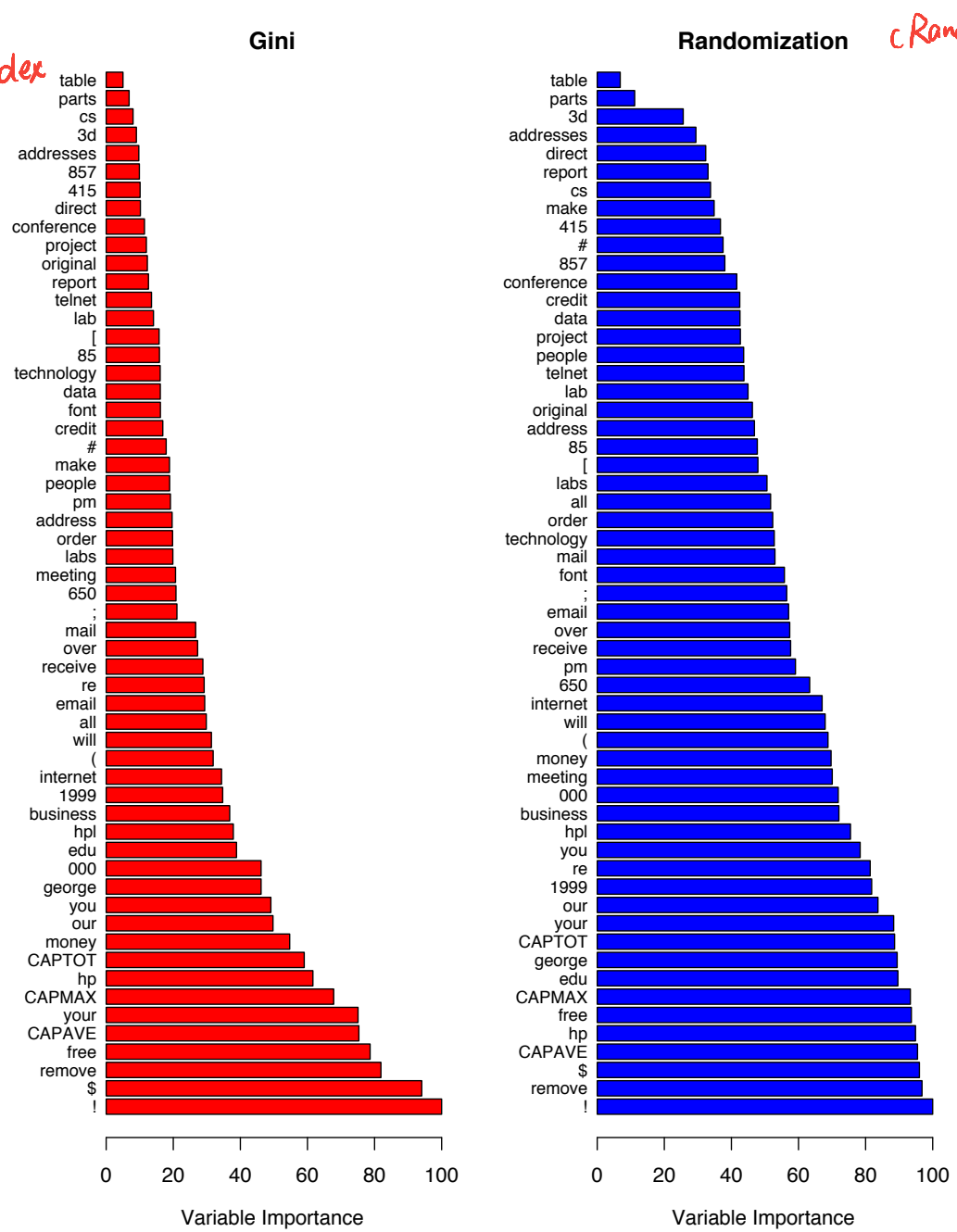


FIGURE 15.4. OOB *error computed on the* `spam` *training data, compared to the test error computed on the test set.*

Important features: Spam

dataset

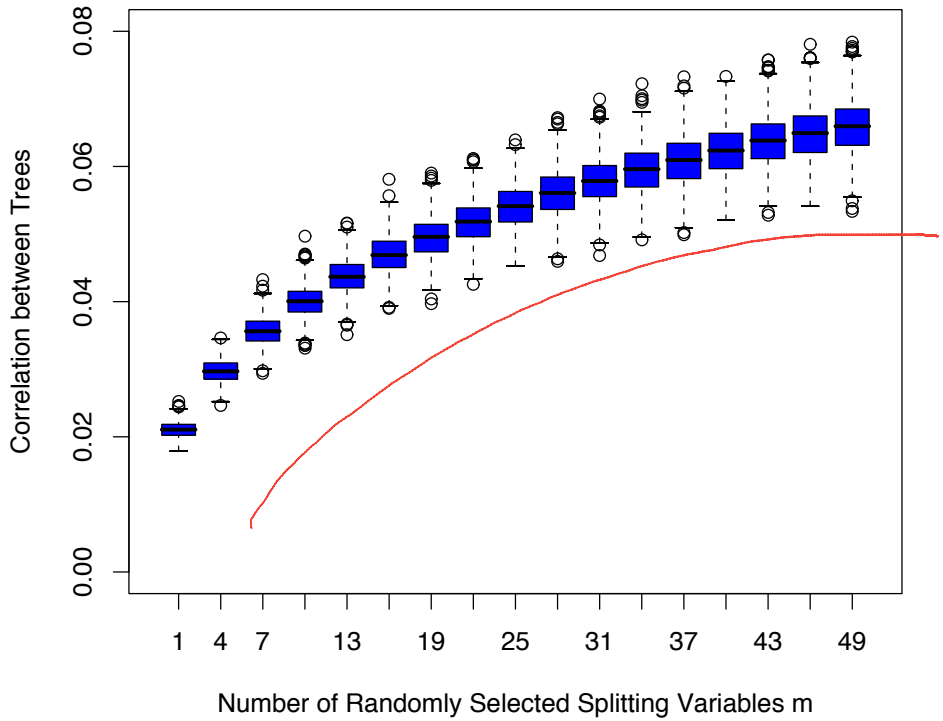
Sort variables by Gini index



(Random Forest)

Random forest correlation

We want trees are decorrelated, so when we average them, reduce variance

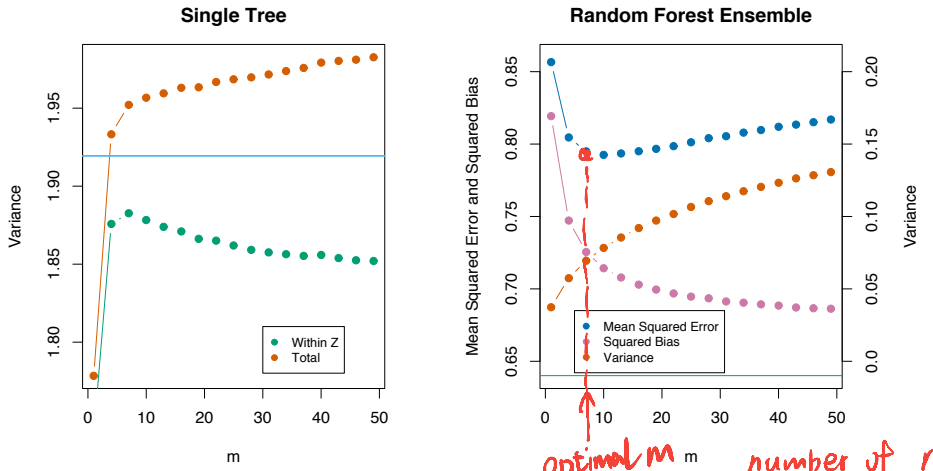


m↑ correlation of trees ↑

FIGURE 15.9. *Correlations between pairs of trees drawn by a random-forest regression algorithm, as a function of* m*. The boxplots represent the correlations at* 600 *randomly chosen prediction points* x*.*

Random forest MSE (mean squared error)

bias-variance tradeoff

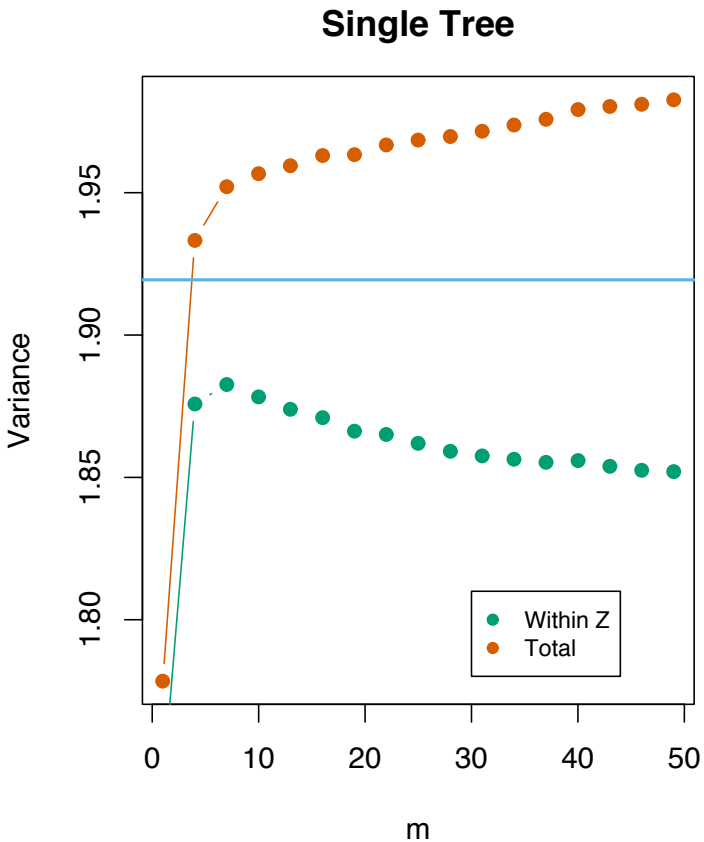


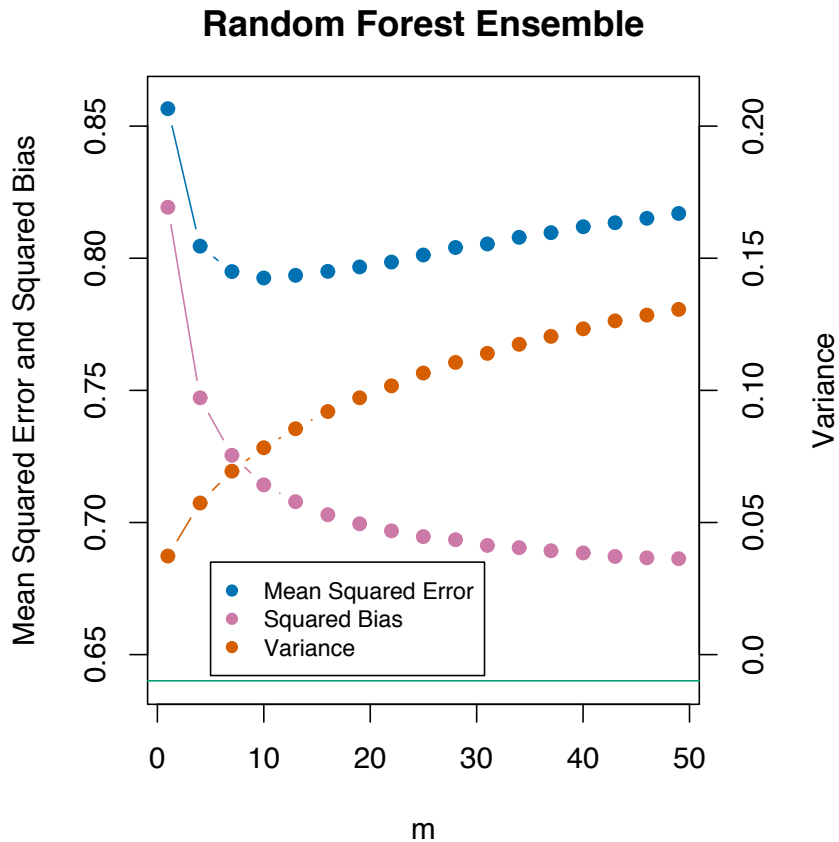
P600

number of randomly selected splitting variables

FIGURE 15.10. *Simulation results. The left panel shows the average variance of a single random forest tree, as a function of m. "Within* $\mathbf{Z}$*" refers to the average within-sample contribution to the variance, resulting from the bootstrap sampling and split-variable sampling (17.9). "Total" includes the sampling variability of* $\mathbf{Z}$*. The horizontal line is the average variance of a single fully grown tree (without bootstrap sampling). The right panel shows the average mean-squared error, squared bias and variance of the ensemble, as a function of m. Note that the variance axis is on the right (same scale, different level). The horizontal line is the average squared-bias of a fully grown tree.*

Random forest MSE





Conclusion: ensemble methods

- Bagging tries to reduce variance by averaging many trees
- Random forests decorrelate by random sampling of predictors
- Manage bias-variance tradeoff with randomization