Random Forest

LEO BREIMAN

2021. 1. 6

윤준영

Problem

- Precedent research on random forest
  - Bagging (decision tree)
  - Random split selection
  - Random noise into the outputs

These < Adaboost

Random Forest

- Random Forest: algorithm

1. For $b = 1$ to B:
   (a) Draw a $\boxed{\text{bootstrap sample}}$ $\mathbf{Z}^*$ 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 $\boxed{m \text{ variables at random}}$ from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter 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: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\mathrm{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

To improve accuracy

-> using randomness

minimize the correlation
maintaining the strength

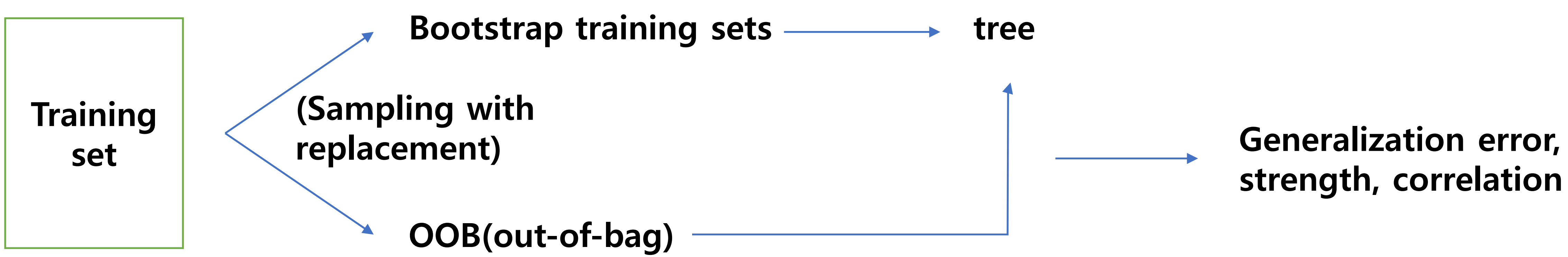
Random Forest

- A specialized bagging for decision tree algorithms
- Method used
  - Bagging
  - Randomly chosen input variables

Definition 1.1. A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k),\ k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$.

Random Forest

- A specialized bagging for decision tree algorithms
- Method used
  - Bagging
  - Randomly chosen input variables



Random Forest

- A specialized bagging for decision tree algorithms
- Method used
    - Bagging
    - Randomly chosen input variables

→ 1. Bagging seems to enhance accuracy when random feature is used

2. Can used to give ongoing estimates of the generalization error of the combined ensemble trees

Random Forest

- Generalization Error

$$Generalization\ Error \ \leq \ \frac{\bar{\rho}(1-s^2)}{s^2}$$

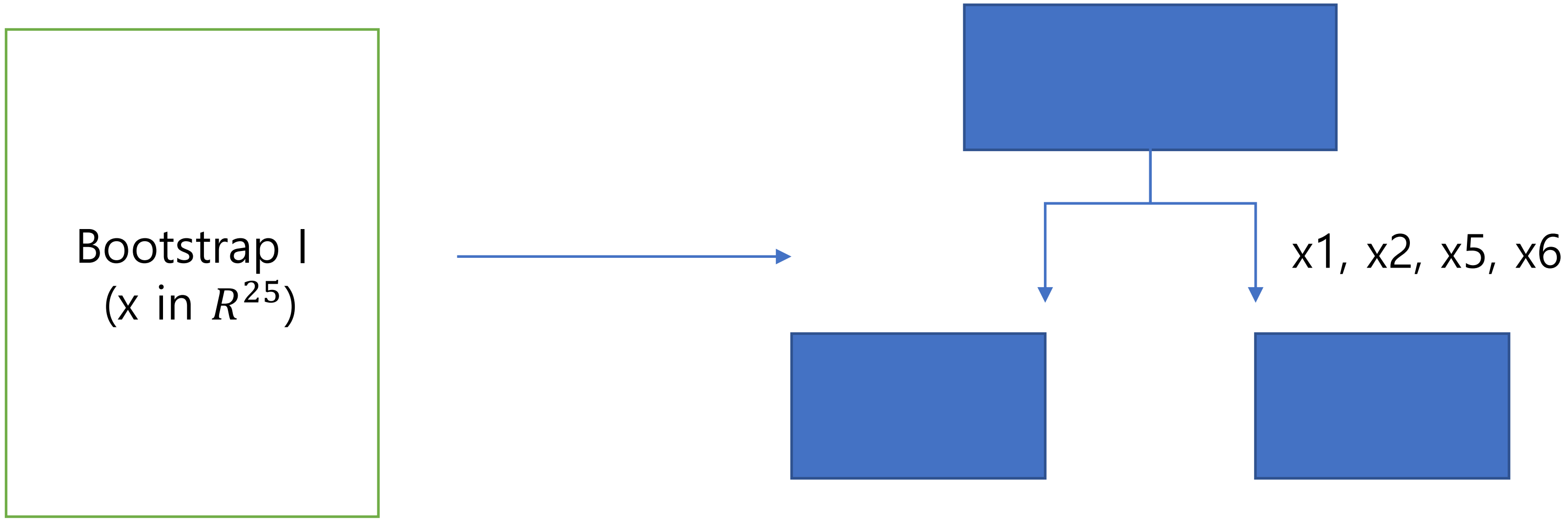
$\bar{\rho}$ is the mean value of the correlation between individual trees

s^2 is the average difference proportions between correct and incorrect trees

Random Forest

- A specialized bagging for decision tree algorithms
- Method used
  - Bagging
  - Randomly chosen input variables

- randomly selected variables to split in each node



In each Bootstrap forms tree
-> predict results by voting

Random Forest

- A specialized bagging for decision tree algorithms
- Method used
  - Bagging
  - Randomly chosen input variables

- randomly selected variables to split in each node

  using random input selection (Forest-RI)

  - selecting at random, at each node, a small group of input variables to split on

Random Forest

Test set errors (%).

Data set	Adaboost	Selection	Forest-RI single input	One tree
Glass	22.0	20.6	21.2	36.9
Breast cancer	3.2	2.9	2.7	6.3
Diabetes	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vowel	4.1	3.4	3.3	30.4
Ionosphere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credit	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Votes	4.8	4.1	4.6	7.4
Liver	30.7	25.1	24.7	40.6
Letters	3.4	3.5	4.7	19.8
Sat-images	8.8	8.6	10.5	17.2
Zip-code	6.2	6.3	7.8	20.6
Waveform	17.8	17.2	17.3	34.0
Twonorm	4.9	3.9	3.9	24.7
Threenorm	18.8	17.5	17.5	38.4
Ringnorm	6.9	4.9	4.9	25.7

Random input selection can be faster than Adaboost or Bagging

Error rate is similar to Adaboost

Random Forest

- A specialized bagging for decision tree algorithms
- Method used
    - Bagging
    - Randomly chosen input variables

- randomly selected variables to split in each node

  using random input selection (Forest-RI)

  - selecting at random, at each node, a small group of input variables to split on

  using linear combination of inputs (Forest-RC)

  - take random linear combinations of number of the input variables

Random Forest

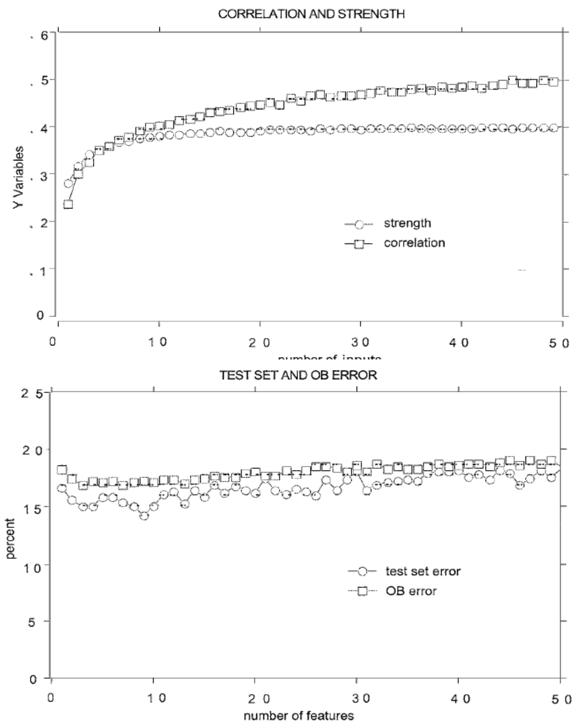
Test set errors (%).

Data set	Adaboost	Selection	Forest-RI single input	One tree
Glass	22.0	20.6	21.2	36.9
Breast cancer	3.2	2.9	2.7	6.3
Diabetes	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vowel	4.1	3.4	3.3	30.4
Ionosphere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credit	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Votes	4.8	4.1	4.6	7.4
Liver	30.7	25.1	24.7	40.6
Letters	3.4	3.5	4.7	19.8
Sat-images	8.8	8.6	10.5	17.2
Zip-code	6.2	6.3	7.8	20.6
Waveform	17.8	17.2	17.3	34.0
Twonorm	4.9	3.9	3.9	24.7
Threenorm	18.8	17.5	17.5	38.4
Ringnorm	6.9	4.9	4.9	25.7

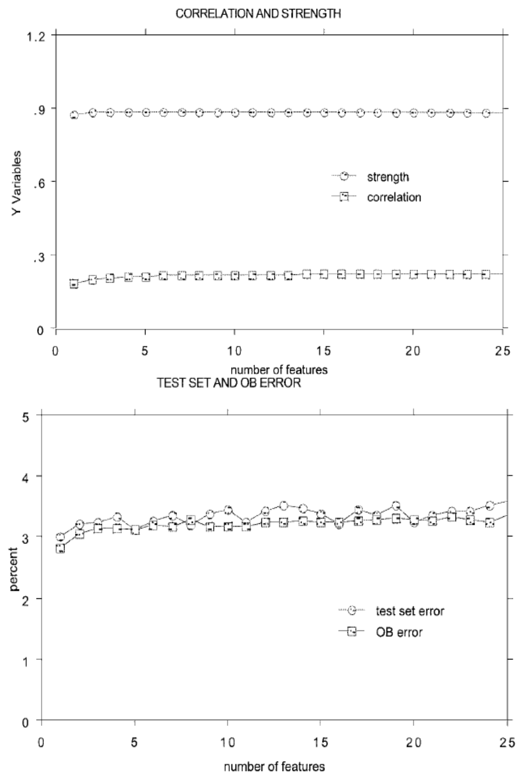
Forest-RC exceptionally does well on the synthetic data-set

Forest-RC compares more favorably to Adaboost than Forest-RI

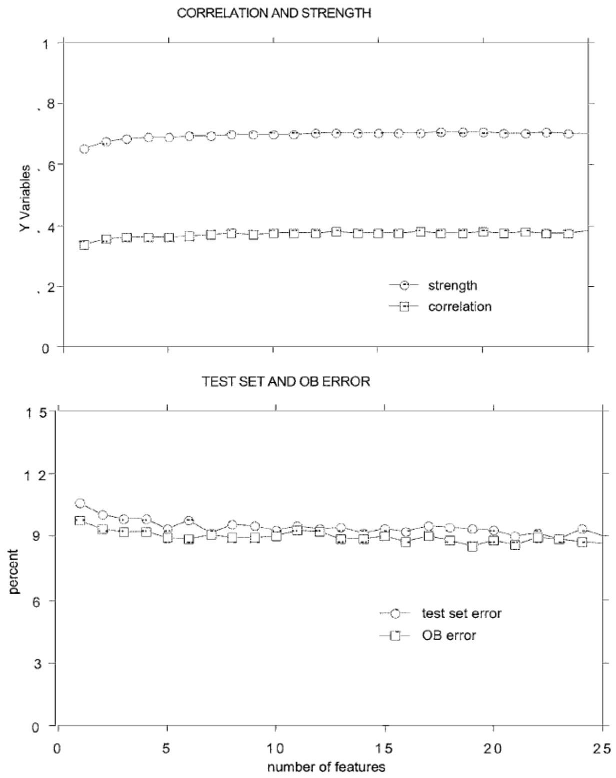
Effect of strength & correlation



Effects of number of inputs

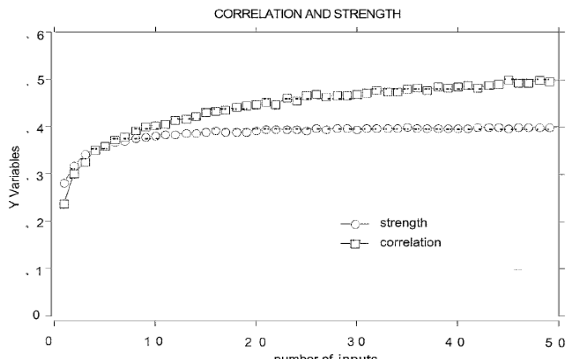


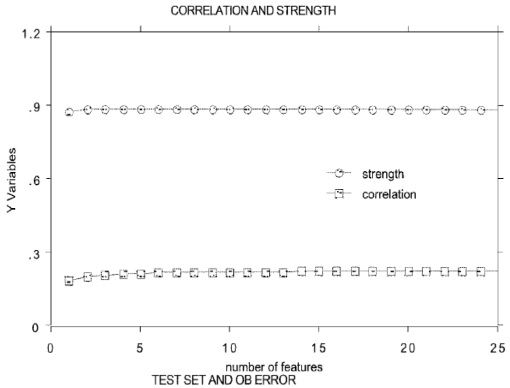
Effects of number of feature (small data)

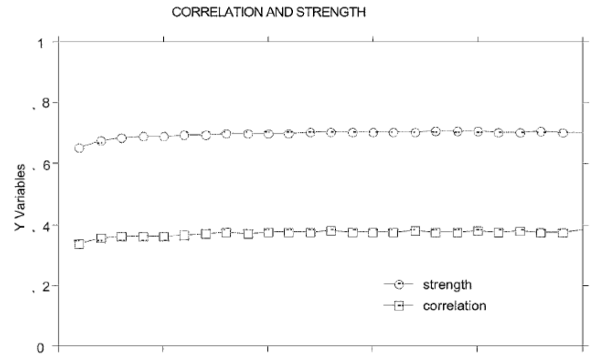


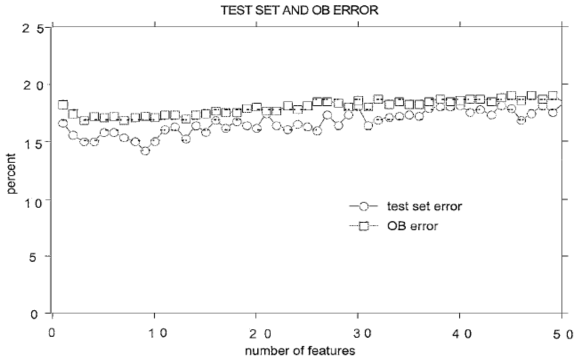
Effects of number of feature (large data)

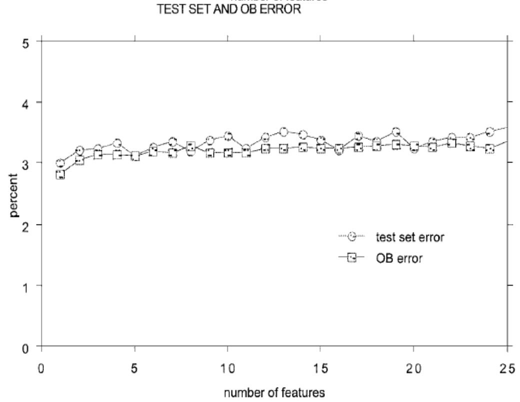
Effect of strength & correlation

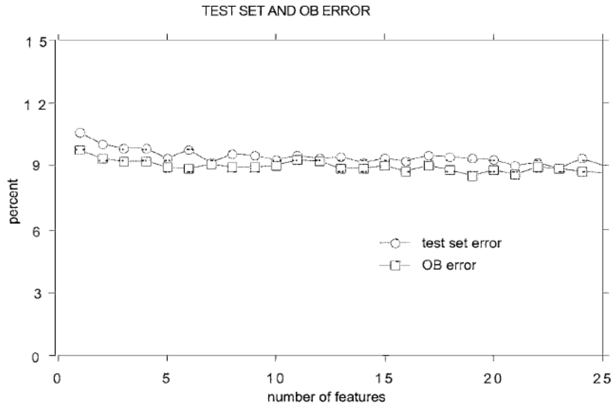












=> Better random forest have lower correlation between trees and higher strength

The effect of noise

Data set	Adaboost	Forest-RI	Forest-RC
Glass	1.6	.4	−.4
Breast cancer	43.2	1.8	11.1
Diabetes	6.8	1.7	2.8
Sonar	15.1	−6.6	4.2
Ionosphere	27.7	3.8	5.7
Soybean	26.9	3.2	8.5
Ecoli	7.5	7.9	7.8
Votes	48.9	6.3	4.6
Liver	10.3	−.2	4.8

Increase of error rates due to noise

Adaboost deteriorates markedly.

Random forest shows small changes

Random Forest: variable importance

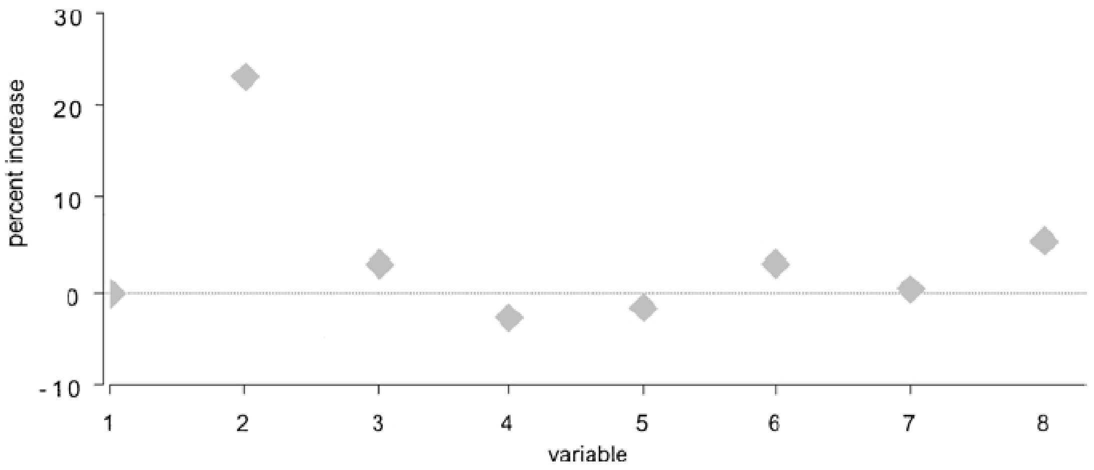
- Variable Importance
  1. OOB error for the original dataset (e_i)
  2. OOB error for the dataset in which variable x_i with noise (OOB is randomly permuted) (p_i)
  3. Compute variable importance based on mean and standard deviation of($p_i - e_i$) over all trees

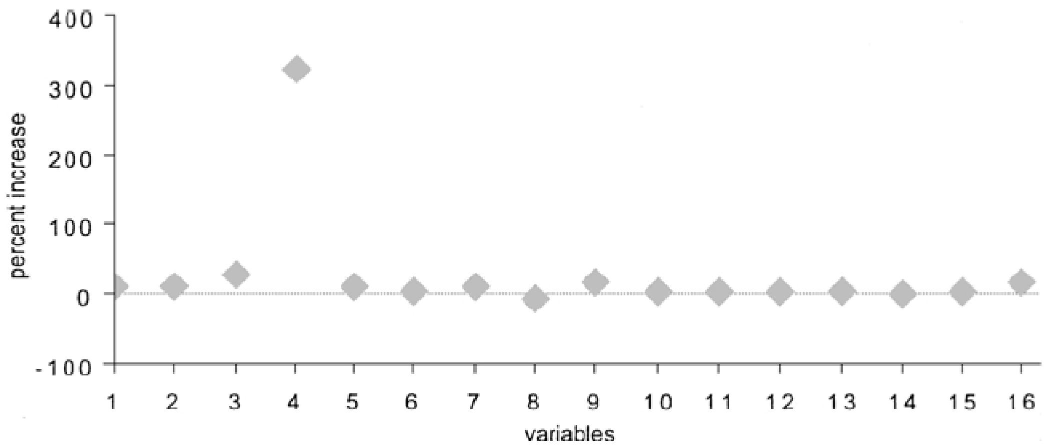
If the variable is important

Gap between random permutation is big

deviation between individual trees are small

Random Forest: variable importance





Regression

Theorem 11.1. *As the number of trees in the forest goes to infinity, almost surrely,*

$$E_{\mathbf{X},Y}(Y - av_k h(\mathbf{X}, \Theta_k))^2 \rightarrow E_{\mathbf{X},Y}(Y - E_{\Theta} h(\mathbf{X}, \Theta))^2. \qquad (12)$$

Theorem 11.2. *Assume that for all* Θ, $EY = E_{\mathbf{X}} h(\mathbf{X}, \Theta)$. *Then*

$$PE^*(forest) \leq \bar{\rho} PE^*(tree)$$

where $\bar{\rho}$ *is the weighted correlation between the residuals* $Y - h(\mathbf{X}, \Theta)$ *and* $Y - h(X, \Theta')$ *where* Θ, Θ' *are independent.*

Regression

- Experiment

Data set	Bagging	Adapt. bag	Forest
Boston Housing	11.4	9.7	10.2
Ozone	17.8	17.8	16.3
Servo $\times 10-2$	24.5	25.1	24.6
Abalone	4.9	4.9	4.6
Robot Arm $\times 10-2$	4.7	2.8	4.2
Friedman #1	6.3	4.1	5.7
Friedman #2 $\times 10+3$	21.5	21.5	19.6
Friedman #3 $\times 10-3$	24.8	24.8	21.6

Data Set	Test error	OB error	PE*(tree)	Cor.
Boston Housing	10.2	11.6	26.3	.45
Ozone	16.3	17.6	32.5	.55
Servo $\times 10-2$	24.6	27.9	56.4	.56
Abalone	4.6	4.6	8.3	.56
Robot Arm $\times 10-2$	4.2	3.7	9.1	.41
Friedman #1	5.7	6.3	15.3	.41
Friedman #2 $\times 10+3$	19.6	20.4	40.7	.51
Friedman #3 $\times 10-3$	21.6	22.9	48.3	.49

Data set	With bagging	With Noise
Boston Housing	10.2	9.1
Ozone	17.8	16.3
Servo $\times 10-2$	24.6	23.2
Abalone	4.6	4.7
Robot Arm $\times 10-2$	4.2	3.9
Friedman #1	5.7	5.1
Friedman #2 $\times 10+3$	19.6	20.4
Friedman #3 $\times 10-3$	21.6	19.8

Using feature is too small,
the tree gets small.

using feature is too large,
Error gets big

Correlation increase slower than classification

OB error are consistently high

Conclusion

- Random forest are an effective prediction tool
- Injecting right kind of randomness makes accurate classifiers and regressors
- Random inputs and Random features produce good in classification
- Different injection of randomness can produce better results