MS1b Statistical Data Mining
Part 4: Supervised Learning
Ensemble Methods

Yee Whye Teh
Department of Statistics
Oxford

http://www.stats.ox.ac.uk/~teh/datamining.html

Outline

Supervised Learning: Ensemble Methods

- Bagging
- Random Forests
- Boosting

Outline

Supervised Learning: Ensemble Methods

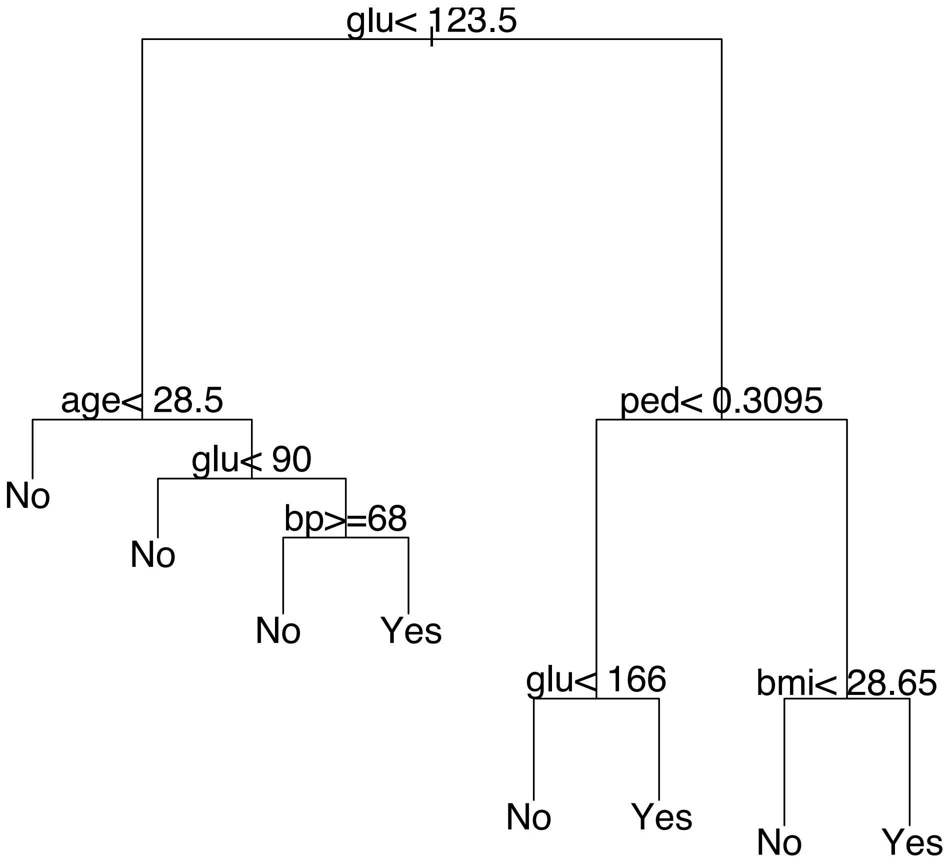
- Bagging
- Random Forests
- Boosting

Bagging

An appeal of trees is their interpretability. Recall classification tree for Pima Indians example.

```
library(rpart)
library(MASS)
data(Pima.tr)                                              ## load data
Diabetes <- Pima.tr[,8]                                    ## response
X <- Pima.tr[,-8]                                          ## predictor
tree <- rpart(Diabetes ~ ., data=X,
                   control=rpart.control(xval=10))  ## 10-fold CV
```

```
> plot(tree); text(tree)
```



Tree is very interpretable, selecting a subset of all predictor variables. Is the tree also ‘stable’ under small perturbations of the data or if we have slightly different training data? Can we do formal ‘significance testing’ as in linear models? How do we know we are not including irrelevant variables?

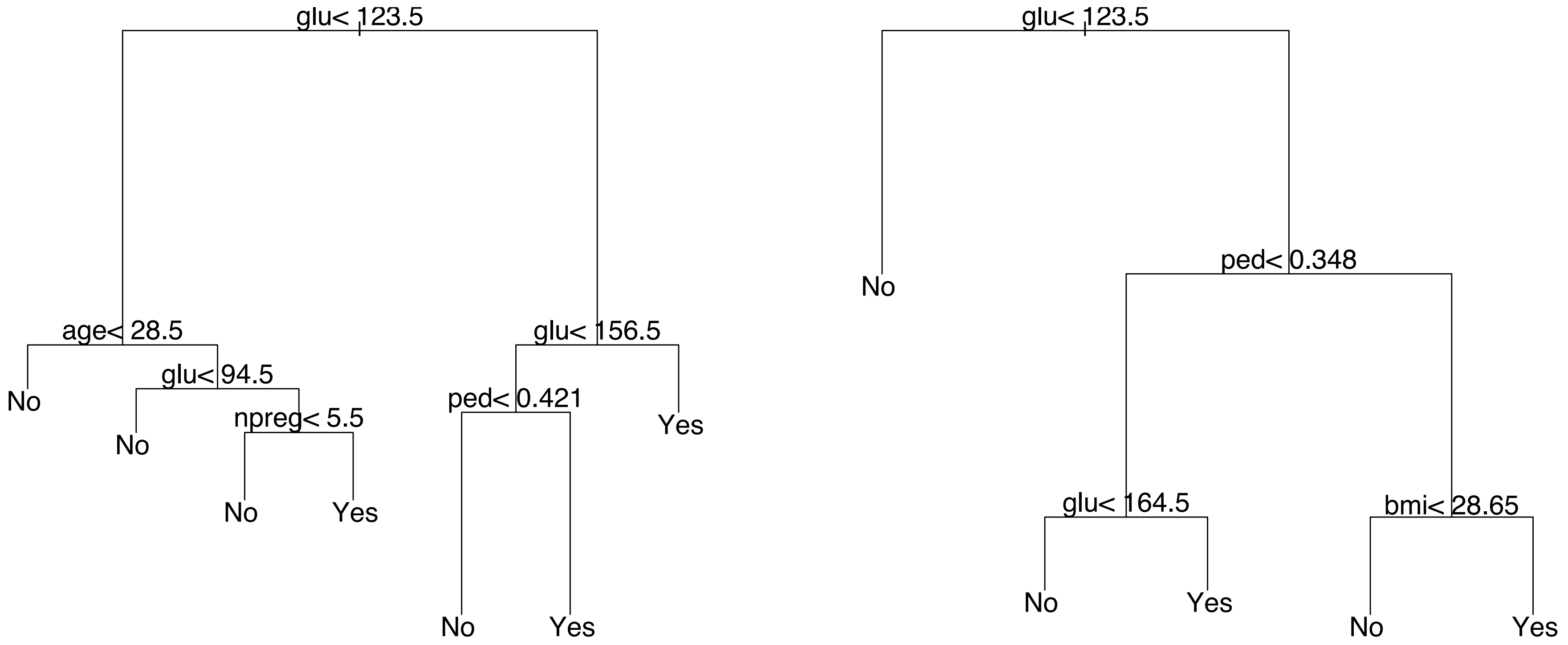
To fit the classification tree, we used all observations $i = 1, \ldots, n$ with $n = 200$. What would the tree have looked like for a slightly different set of observations?
The Bootstrap (Efron, 79) is a natural way to assess the variance of estimators, fitting the tree repeatedly on so-called **bootstrap samples**. These are random sets of size n, where each element is drawn **with replacement** from the original n observations $\{1, \ldots, n\}$.

```
> n <- nrow(X)
> subsample <- sample(1:n, n , replace=TRUE)

> sort(subsample)
[1]  2 4 4 5 6 7 9 10 11 12 12 12 12 13 13 15 15 20 ...
```

Some of the original observations do not appear in the bootstrap sample (e.g. $i = 1$ or $i = 3$); some appear once (e.g. $i = 2$ or $i = 5$) and some twice or more often (e.g. $i = 4$).
Fit the tree on these resampled observations.

```
> tree_boot <- rpart(Diabetes ~ ., data=X,  subset=subsample,
                      control=rpart.control(xval=10)) ## 10-fold CV
```

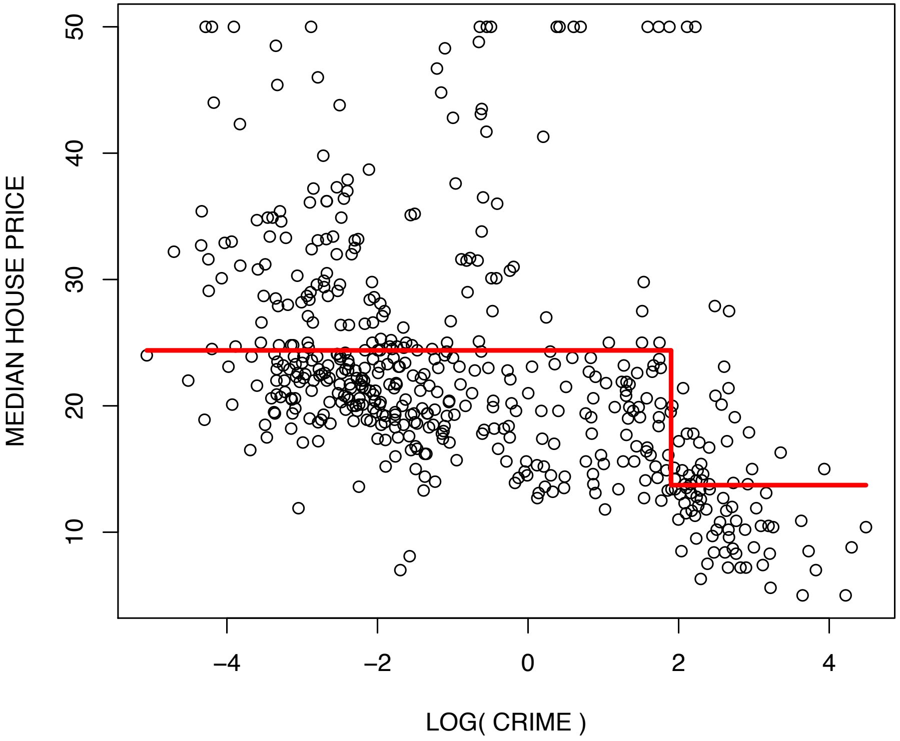
Doing this twice, we get the two following trees, each fitted on a different (random) subset of the data.

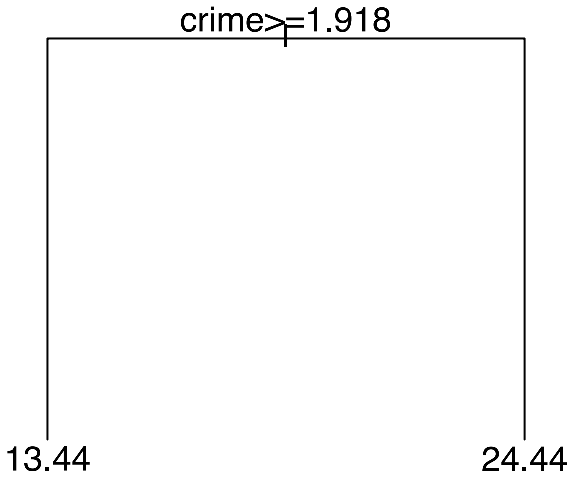


Classification trees are typically not very stable under subsampling of the data. This affects both interpretability and also prediction.

We might for example be suspicious of a particular classification (e.g. “No”) if a large fraction of resampled trees is classifying otherwise (classifying as “Yes”).

Can also look at regression trees.
Take the previous example of the Boston Housing data, trying to predict median house prices, based on the (univariate) predictor variable crime rate.





Fit a stump (the simplest tree – just a root node) to the data. This yields the fitted function $\hat{Y}(x)$, shown as a red solid line.

We fitted (tree) predictor $\hat{Y}(x)$ on the observations

$$(X_1, Y_1), \dots, (X_n, Y_n), \qquad i = 1, \dots, n.$$

Assess the variance of the fitted function $\hat{Y}(x)$ by taking $B = 20$ random subsamples of the original data. Fit bootstrap estimators (trees)

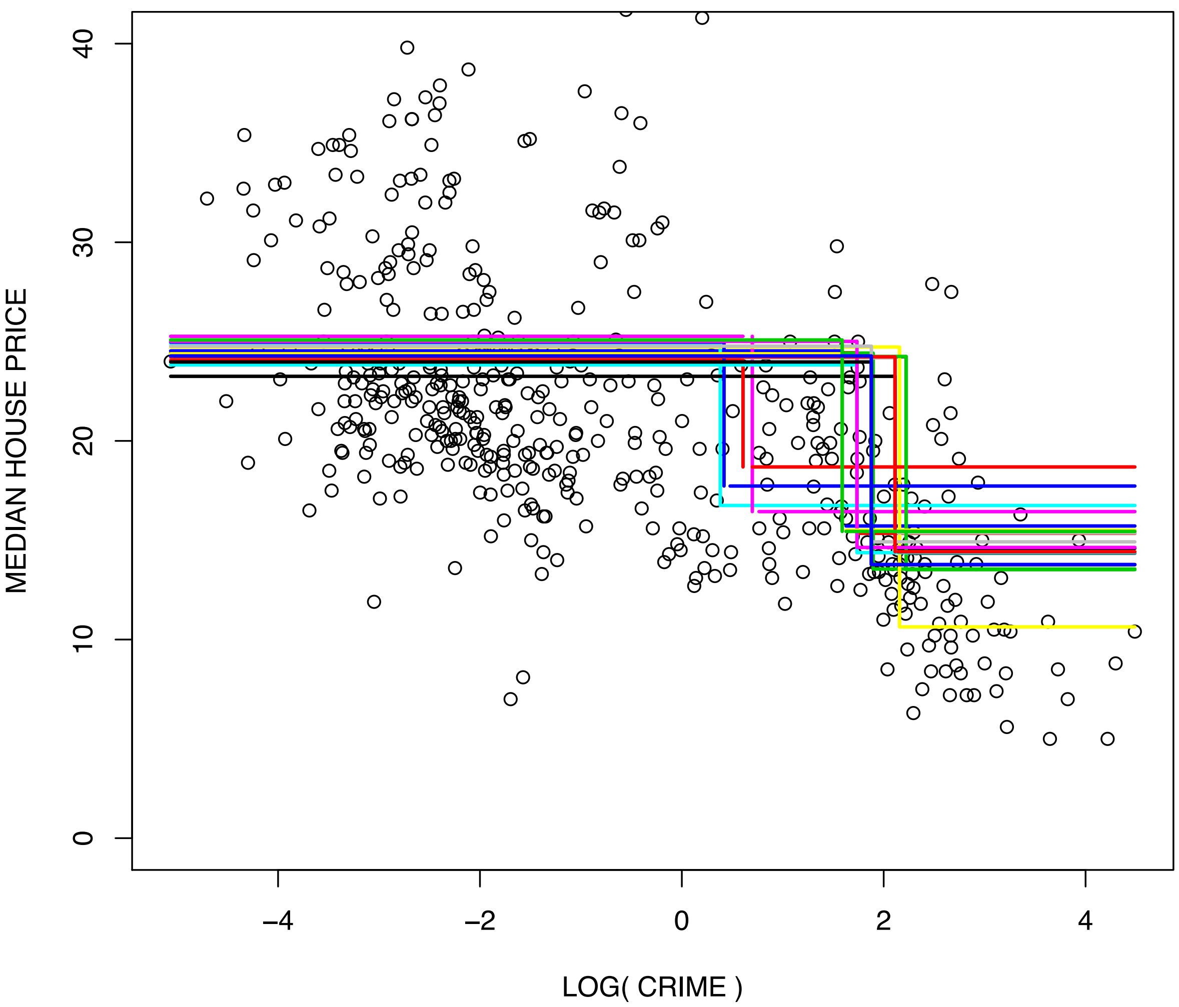
$$\hat{Y}^{*,b}(x), \qquad b = 1, \dots, B$$

where each tree $\hat{Y}^{*,b}$ is fitted on the resampled data

$$(X_{j_1}, Y_{j_1}), \dots, (X_{j_n}, Y_{j_n}), \qquad i = 1, \dots, n,$$

each index j_k, $k = 1, \dots, n$, being chosen at random from the set $\{1, \dots, n\}$ with replacement.

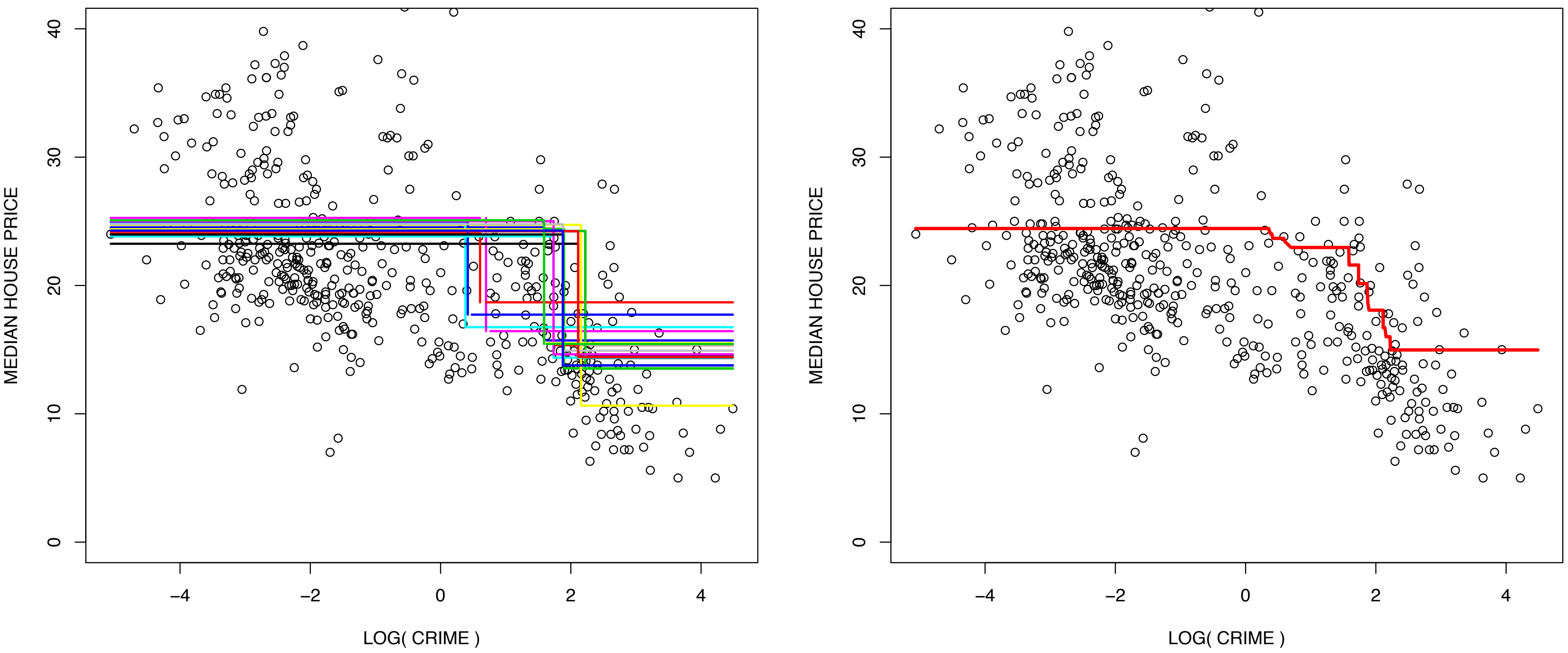
Trees $\hat{Y}^{*,1}, \ldots, \hat{Y}^{*,20}$ each fitted on a different (random) bootstrap sample of the original $n = 500$ observations.



The variance of the fit $\hat{Y}^*$ is high in the region where the splitpoint is placed.

Idea of Bagging (**B**ootstrap **Agg**regation): average across all B trees fitted on different bootstrap samples,

$$\hat{Y}_{Bag} = \frac{1}{B} \sum_{i=1}^{B} \hat{Y}^{*,i}.$$



Empirically, Bagging seems to reduce the variance of $\hat{Y}$, e.g.

$$E\Big((\hat{Y} - E(\hat{Y}))^2\Big) \;\; \geq \;\; E\Big((\hat{Y}_{Bag} - E(\hat{Y}_{Bag}))^2\Big).$$

Bagged trees are an example of an **ensemble of trees**, as prediction is based on many individual predictors.
In summary, bagging trees has the following algorithm. Let $\hat{Y}$ be a tree (or other predictor), based on samples $(X_1, Y_1), \dots, (X_n, Y_n)$.

1. Draw indices $(j_1, \dots, j_n)$ from the set $\{1, \dots, n\}$ with replacement. Fit the tree $\hat{Y}^*$ based on samples
$$(X_{j_1}, Y_{j_1}), \dots, (X_{j_n}, Y_{j_n}).$$
2. Repeat first step B times to obtain
$$\hat{Y}^{*,1}, \dots, \hat{Y}^{*,B}.$$
3. Bagged estimator is
$$\hat{Y}_{Bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{Y}^{*,b}.$$

Variance reduction

Suppose, in an ideal world, we could instead base trees $\tilde{Y}^{*,b}$, $b = 1, \ldots, B$ on n samples drawn from the (unknown) joint distribution of (X, Y), instead of resampling from the original n observations.
The bagged estimator is then

$$\tilde{Y}_{Bag} \;=\; \frac{1}{B} \sum_{b=1}^{B} \tilde{Y}^{*,b}.$$

For $B \to \infty$ (many bootstrap samples),

$$\tilde{Y}_{Bag} \;\to\; E(\hat{Y}),$$

where the expectation is with respect to the random sample of n observations and $\hat{Y}$ is the standard estimator (tree) fitted on these n observations.

Now compare the squared error loss of $\tilde{Y}_{Bag}$ with the loss of the original tree estimator $\hat{Y}$,

$$E\big((Y-\hat{Y})^2\big),$$

where both $\hat{Y}=\hat{Y}(x)$ and $\hat{Y}_{Bag}=\hat{Y}_{Bag}(x)$ are evaluated at some $x\in\mathbb{R}^p$ and the expectation is with respect to a random new observation Y and a new training sample on which $\hat{Y}$ is fitted.
Using $\tilde{Y}_{Bag}\to E(\hat{Y})$ for $B\to\infty$,

$$\begin{aligned} E\big((Y-\hat{Y})^2\big) &= E\big((Y-\tilde{Y}_{Bag}+\tilde{Y}_{Bag}-\hat{Y})^2\big) \\ &= E\big((Y-\tilde{Y}_{Bag})^2\big)+E\big((\tilde{Y}_{Bag}-\hat{Y})^2\big) \\ &\geq E\big((Y-\tilde{Y}_{Bag})^2\big). \end{aligned}$$

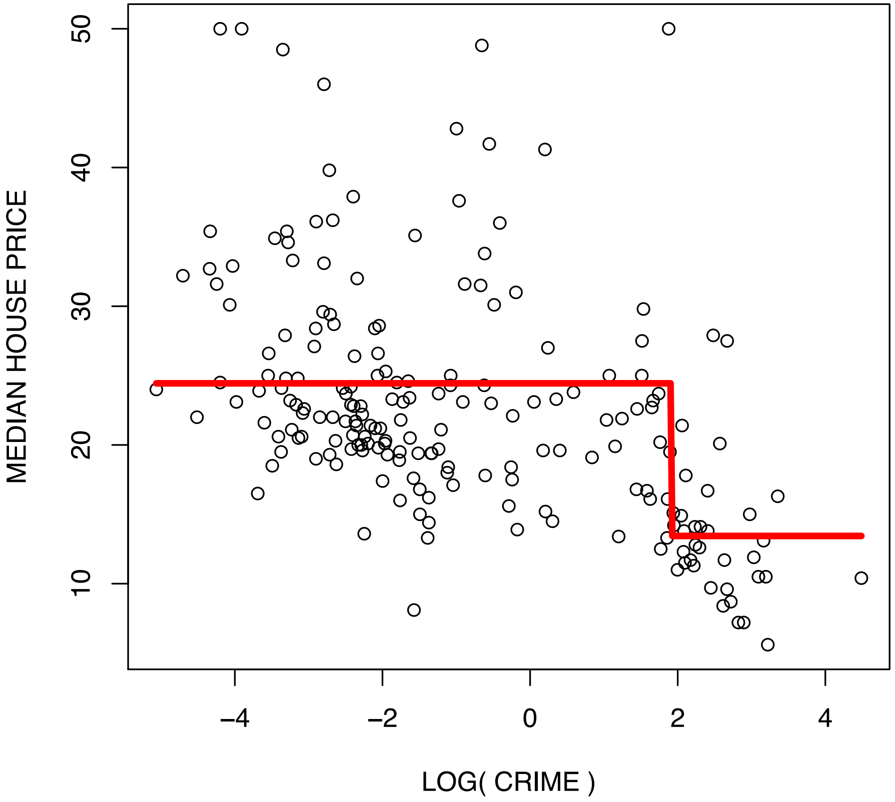
The (population) bagging estimator $\tilde{Y}_{Bag}$ thus reduced the squared error loss by eliminating the variance of $\hat{Y}$ around its mean $E(\hat{Y})$.

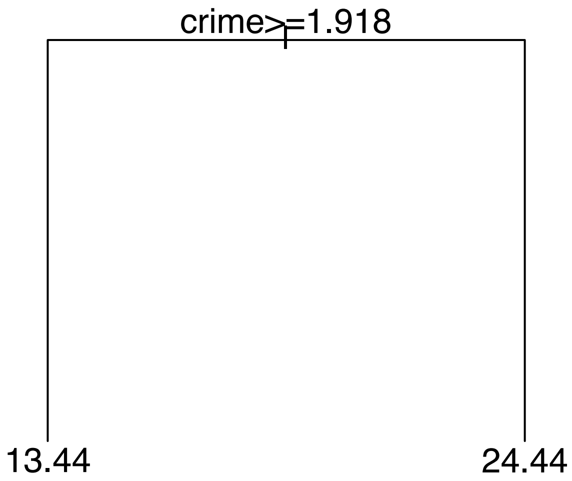
The variance reduction still applies if the idealized (population) estimate $\tilde{Y}_{Bag}$ is replaced by the actual bagging estimator $\hat{Y}$. This variance reduction is traded for a (small) increase in the bias in the procedure.
Bagging helps thus most for ‘flexible’ estimators $\hat{Y}$ which have a high variance

$$E\big((\hat{Y} - E(\hat{Y}))^2\big).$$

For trees, this means that bagging has a very beneficial effect on trees with a large size (number of leaf nodes), whereas the benefit of bagging on trees with small size is much less pronounced.

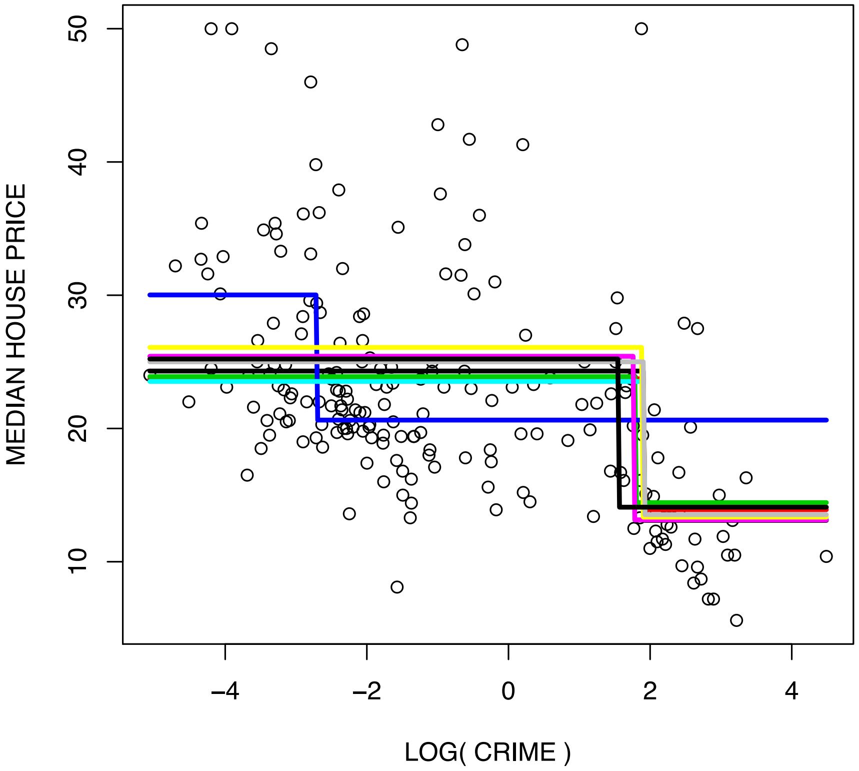
Look again at previous example of predicting house prices, using crime rate as the univariate predictor.



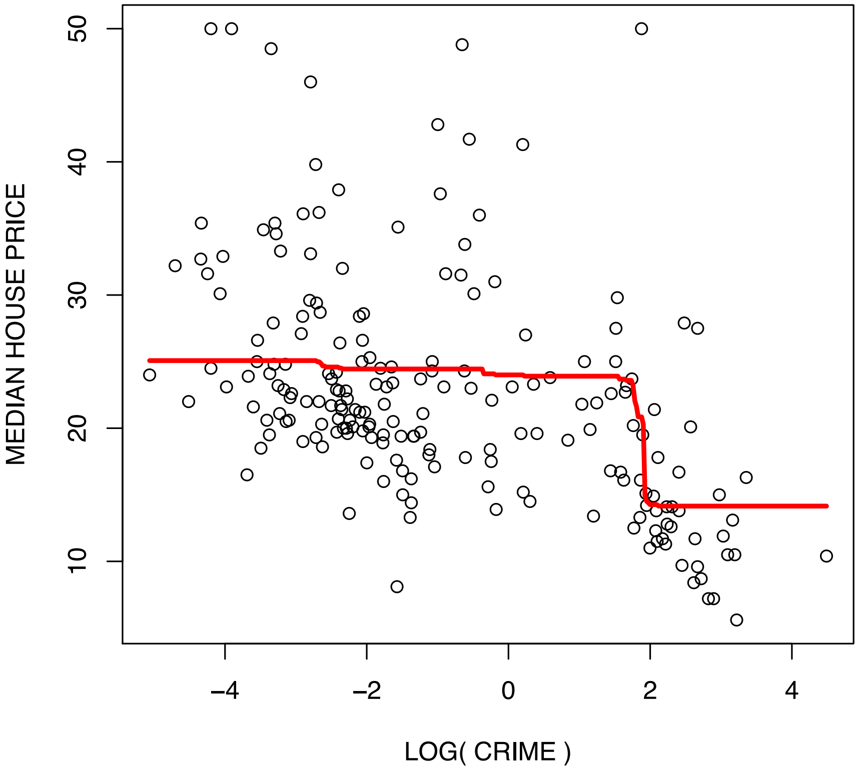


Fitting a single tree with depth $d = 1$ (a stump).

Bagged stumps $\hat{Y}^{*,b}$, $b = 1, \ldots, 10$.

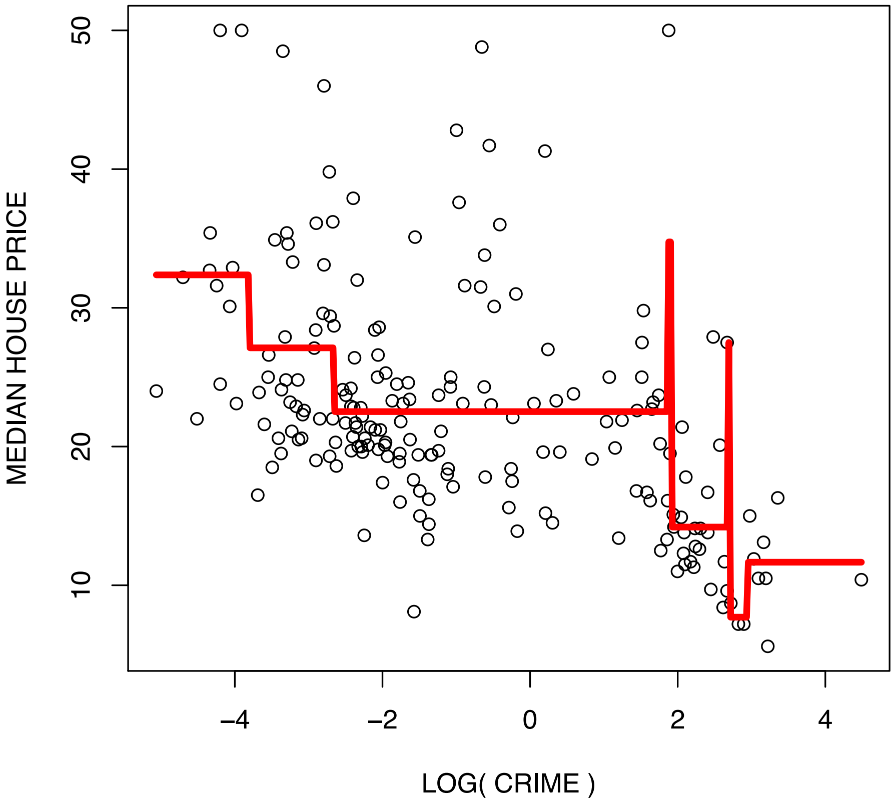


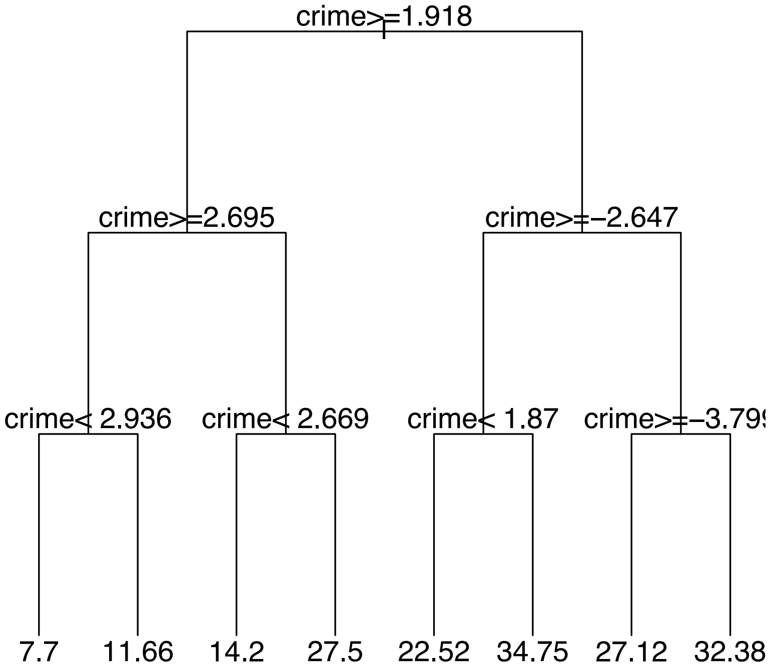
Averaged bagged estimator $\hat{Y}_{Bag}$.



A stump $\hat{Y}$ has (relative to larger trees) a low variance (and a high bias). Bagging leads to a small but not a dramatic improvement.

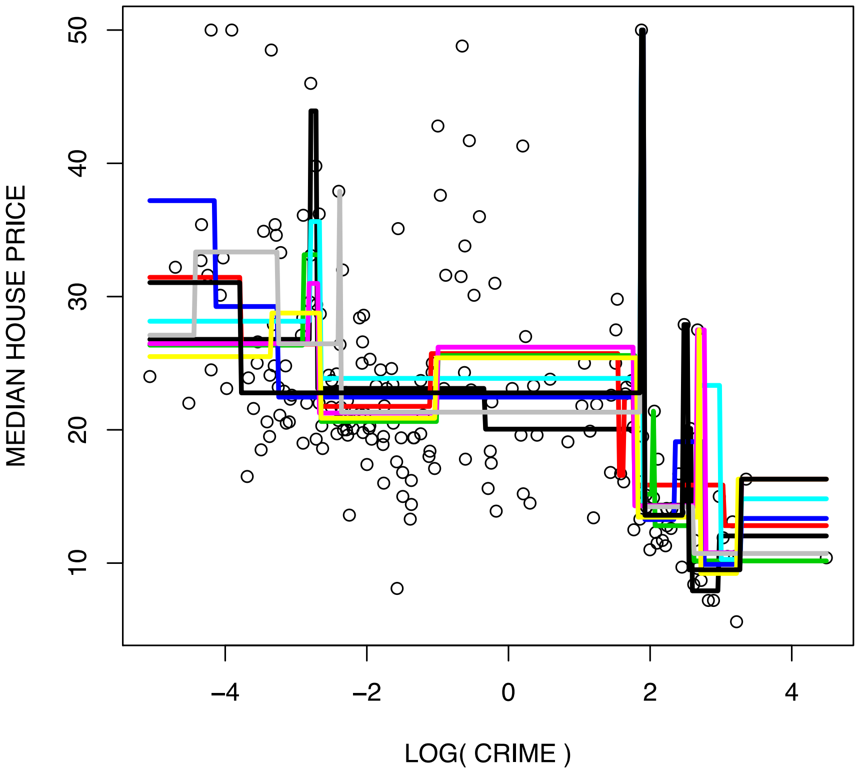
Now fit a tree with depth $d = 3$.



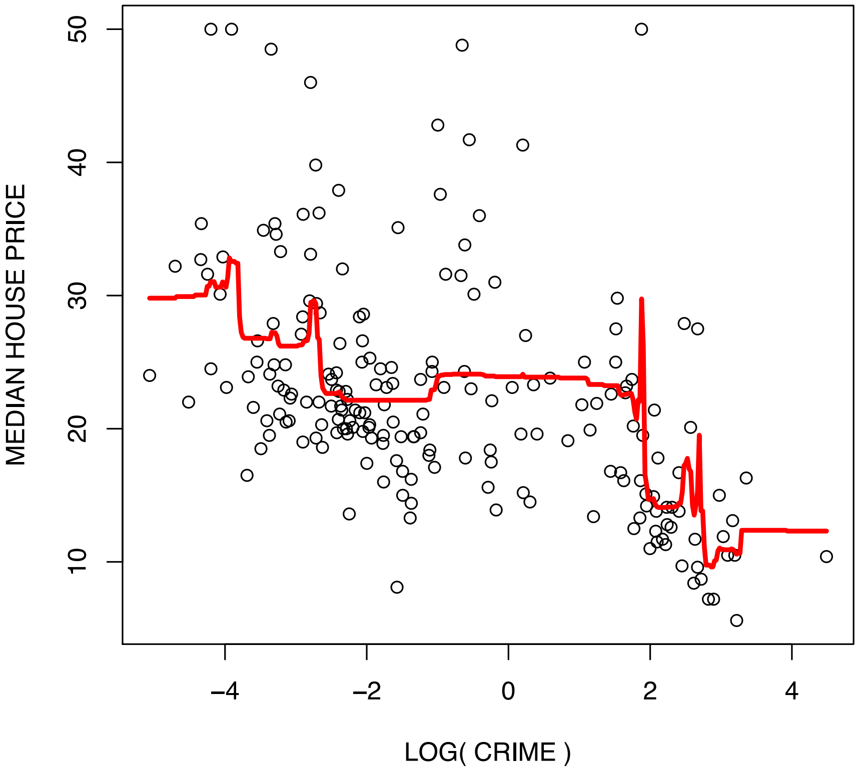


The fit of a single tree has a high variance and will have poor performance (when trying to predict new observations).

Bagged trees of depth $d = 3$, $\hat{Y}^{*,b}$, $b = 1, \ldots, 10$.



Averaged bagged estimator $\hat{Y}_{Bag}$.



As $\hat{Y}$ has a high variance (and a low bias), bagging leads to a large improvement.

Even though **improvement** through bagging is largest in general for trees with large depths, the optimal tree depth (yielding smallest prediction error when bagging) is not obvious a priori.

Out-of-bag test error estimation

To answer this question, we need again a good approximation to the test error (here for the squared error loss function L),

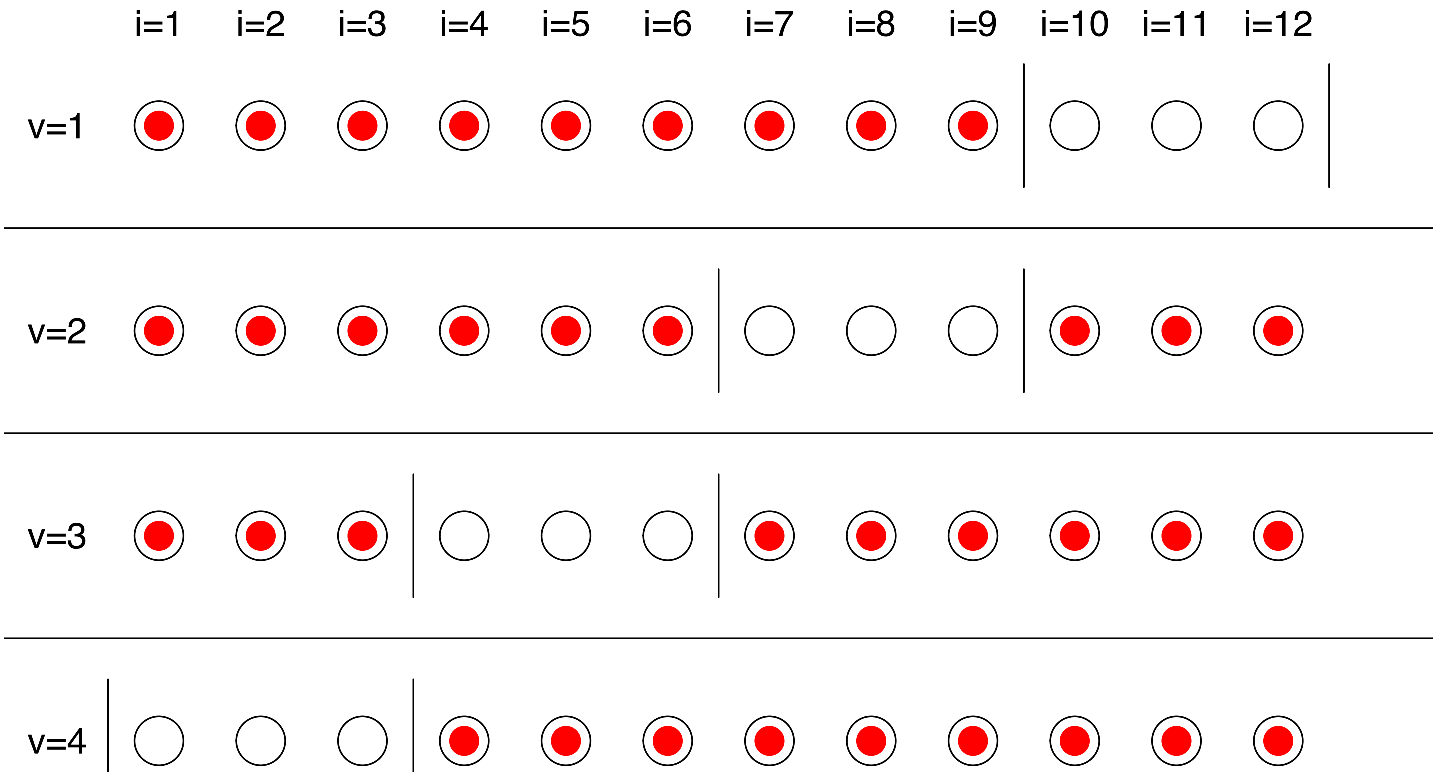
$$R_{test} := E(L(Y, \hat{Y}_{Bag})),$$

where the expectation is with respect to new random pairs (X, Y) and $\hat{Y}_{Bag} = \hat{Y}_{Bag}(X)$, to

- tune the parameters of the algorithm (e.g. select depth of the tree)
- or assess the true performance (and compare with other approaches).

Could compute generalization error $\widehat{R}_{test}$ by cross-validation (CV), as discussed previously.

Here schematic illustration of $V = 4$-fold CV for $n = 12$ samples.

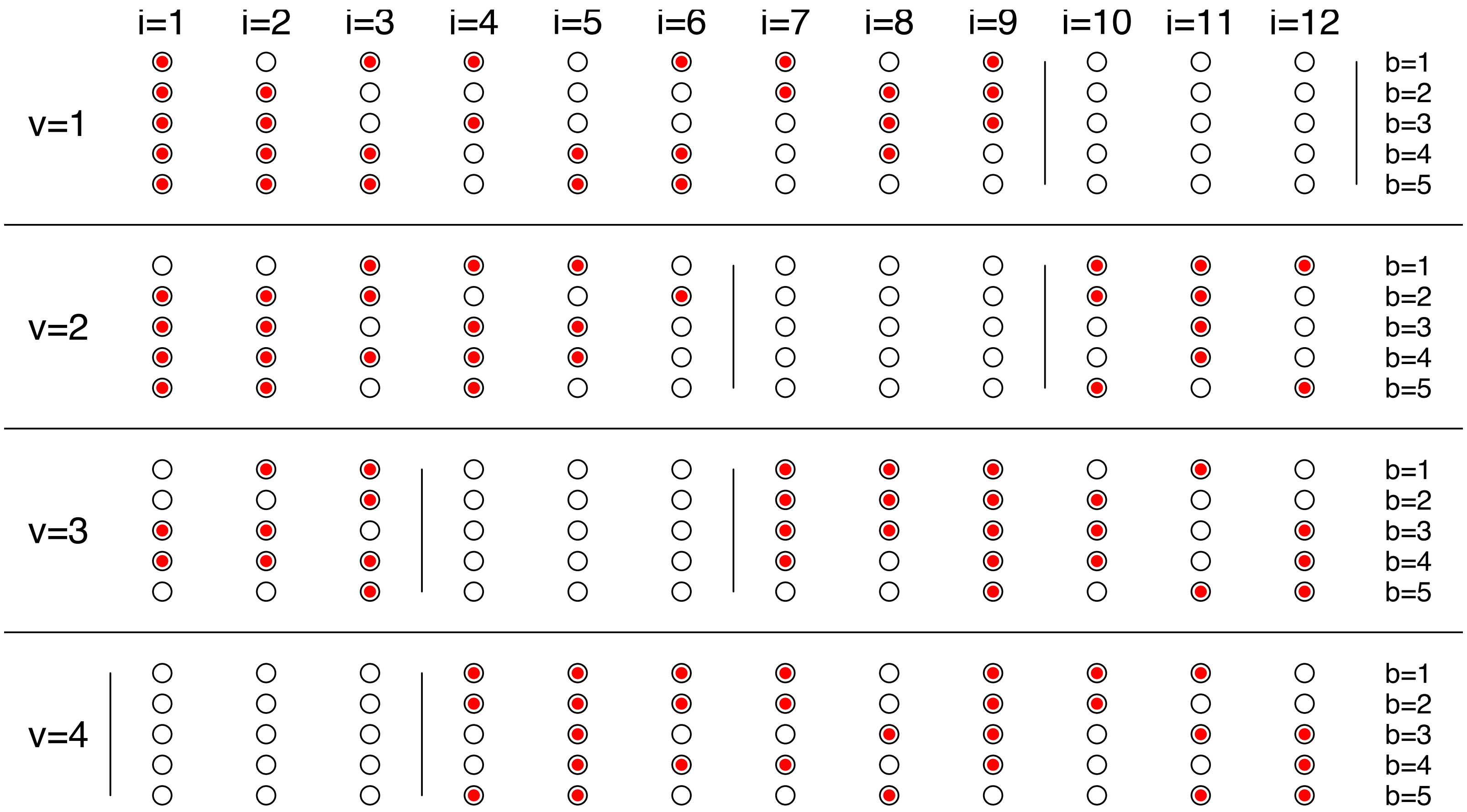


For each $v = 1, \ldots, V$,

- fit $\hat{Y}_{Bag}$ on the training samples, shown as red and filled dots.
- predict with this tree the left-out test observations, shown as white unfilled circles.

Compute the CV test error by averaging the loss across all test observations.

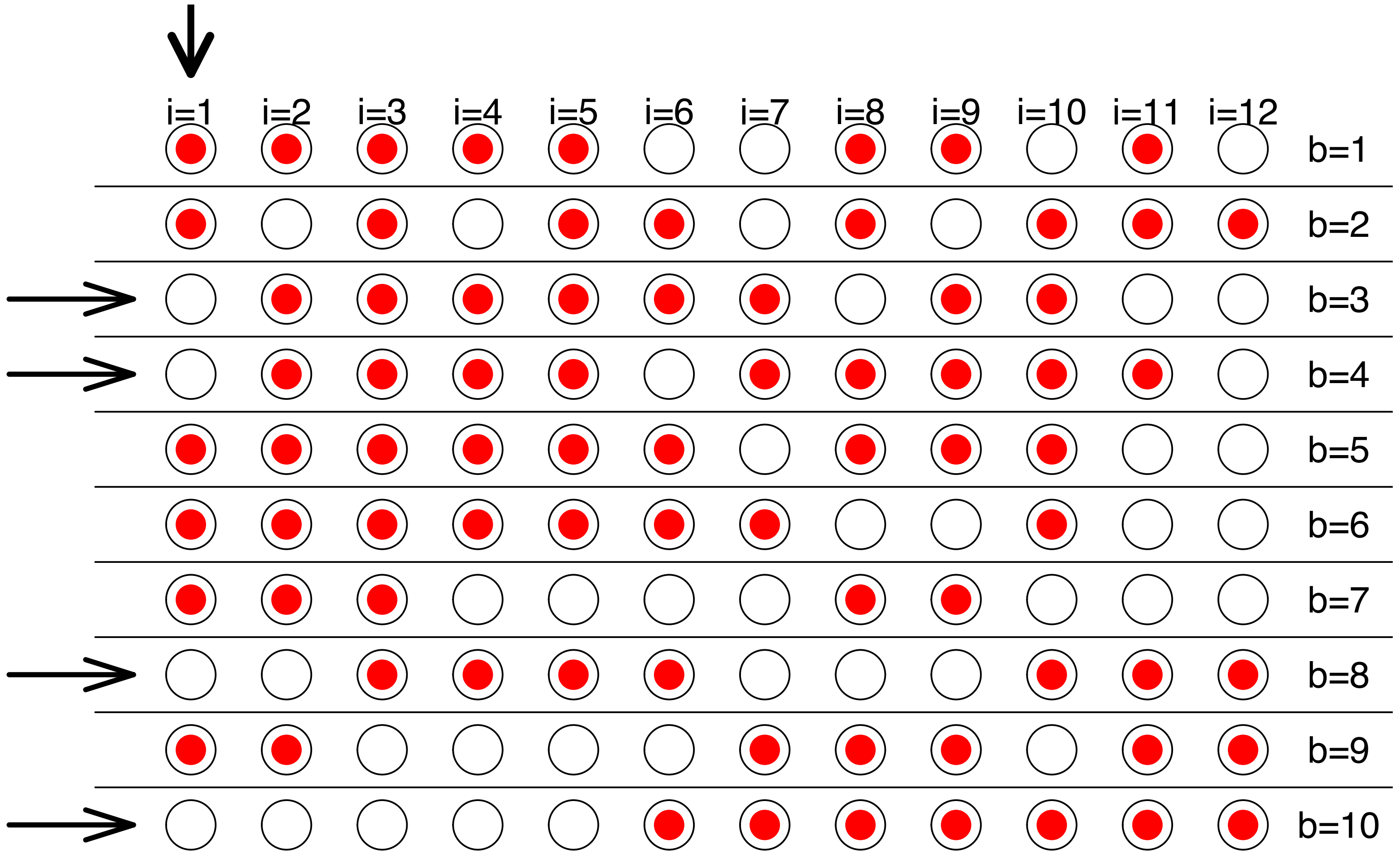
But to fit $\hat{Y}_{Bag}$ on the training samples for each $v = 1, \ldots, V$, need another set of B bootstrap samples on which the original tree is fitted (and whose average gives the $\hat{Y}_{Bag}$ for these training observations).



For each $v = 1, \ldots, V$, the tree needs to be fitted B times. In total, $V \times B$ fits are necessary. This can be very expensive computationally.

$\Rightarrow$ Out-of-bag estimation !

Idea: test on the “unused” data points in each bootstrap iteration to estimate the test error.



i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10	i=11	i=12	
●	●	●	●	●	○	○	●	●	○	●	○	b=1
●	○	●	○	●	●	○	●	○	●	●	●	b=2
○	●	●	●	●	●	●	○	●	●	○	○	b=3
○	●	●	●	●	○	●	●	●	●	●	○	b=4
●	●	●	●	●	●	○	●	●	●	○	○	b=5
●	●	●	●	●	●	●	○	○	●	○	○	b=6
●	●	●	○	○	○	○	●	●	○	○	○	b=7
○	○	●	●	●	●	○	○	○	●	●	●	b=8
●	●	○	○	○	○	●	●	●	○	●	●	b=9
○	○	○	○	○	●	●	●	●	●	●	●	b=10

If fitting B bootstrap estimates $\hat{Y}^{*,b}$, to assess the prediction for $i = 1$, average only over such b, where observation $i = 1$ has not been used in fitting $\hat{Y}^{*,b}$.

Recall that, for B bootstrap samples $\hat{Y}^{*,b}$, the bagged estimator at observation i is given by $\hat{Y}_i := \hat{Y}_{Bag}(X_i)$,

$$\hat{Y}_i = \frac{1}{B} \sum_{b \in \{1,\dots,B\}} \hat{Y}^{*,b}(X_i)$$

Instead, let now

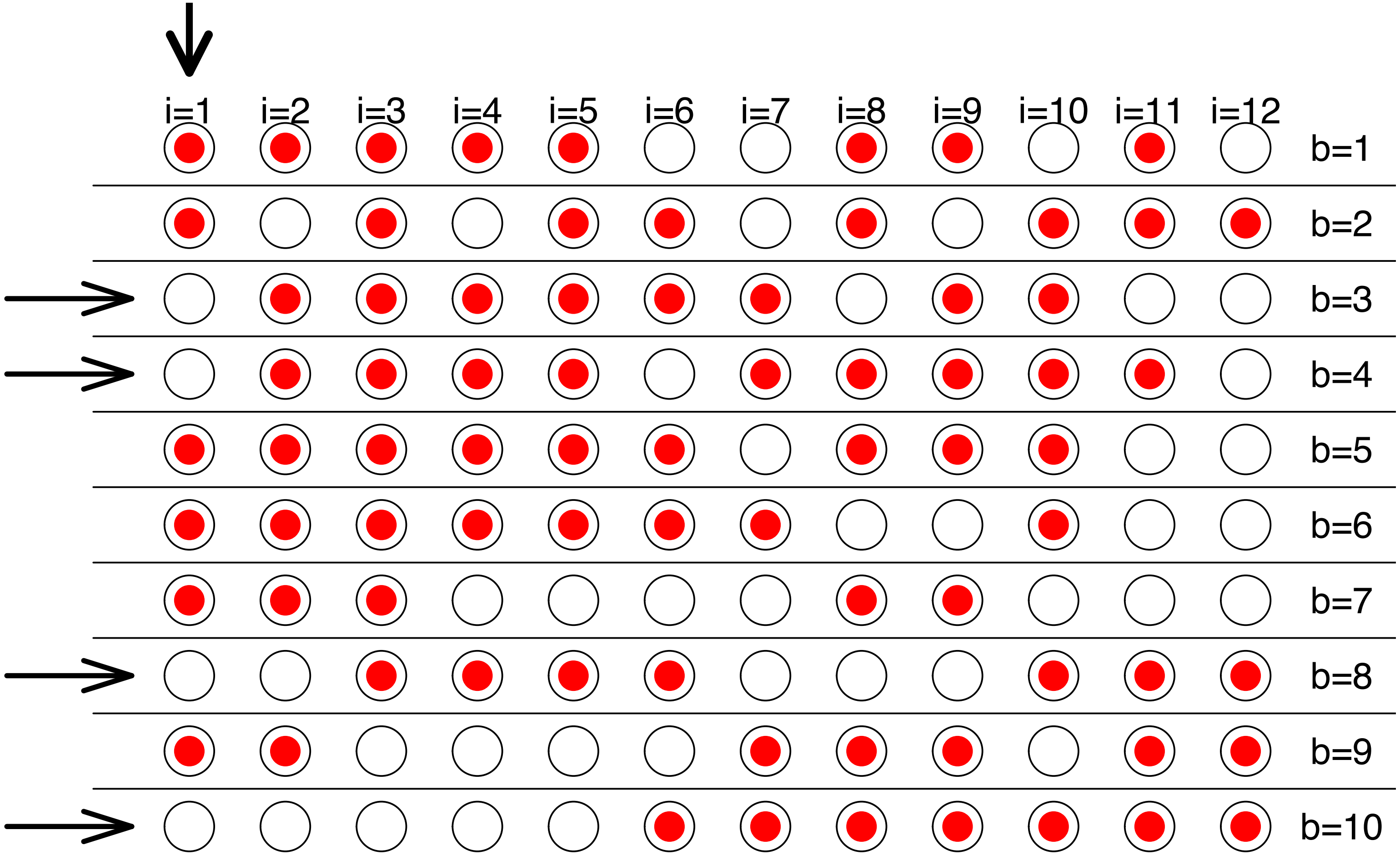
$$\hat{Y}_i^{oob} = \frac{1}{|\tilde{B}_i|} \sum_{b \in \tilde{B}_i} \hat{Y}^{*,b}(X_i),$$

where the sum is only taken over the set

$$\tilde{B}_i = \{b : X_i \text{ is not in training set}\} \subseteq \{1, \dots, B\}.$$

The estimate of the test error is then computed, as usual, by

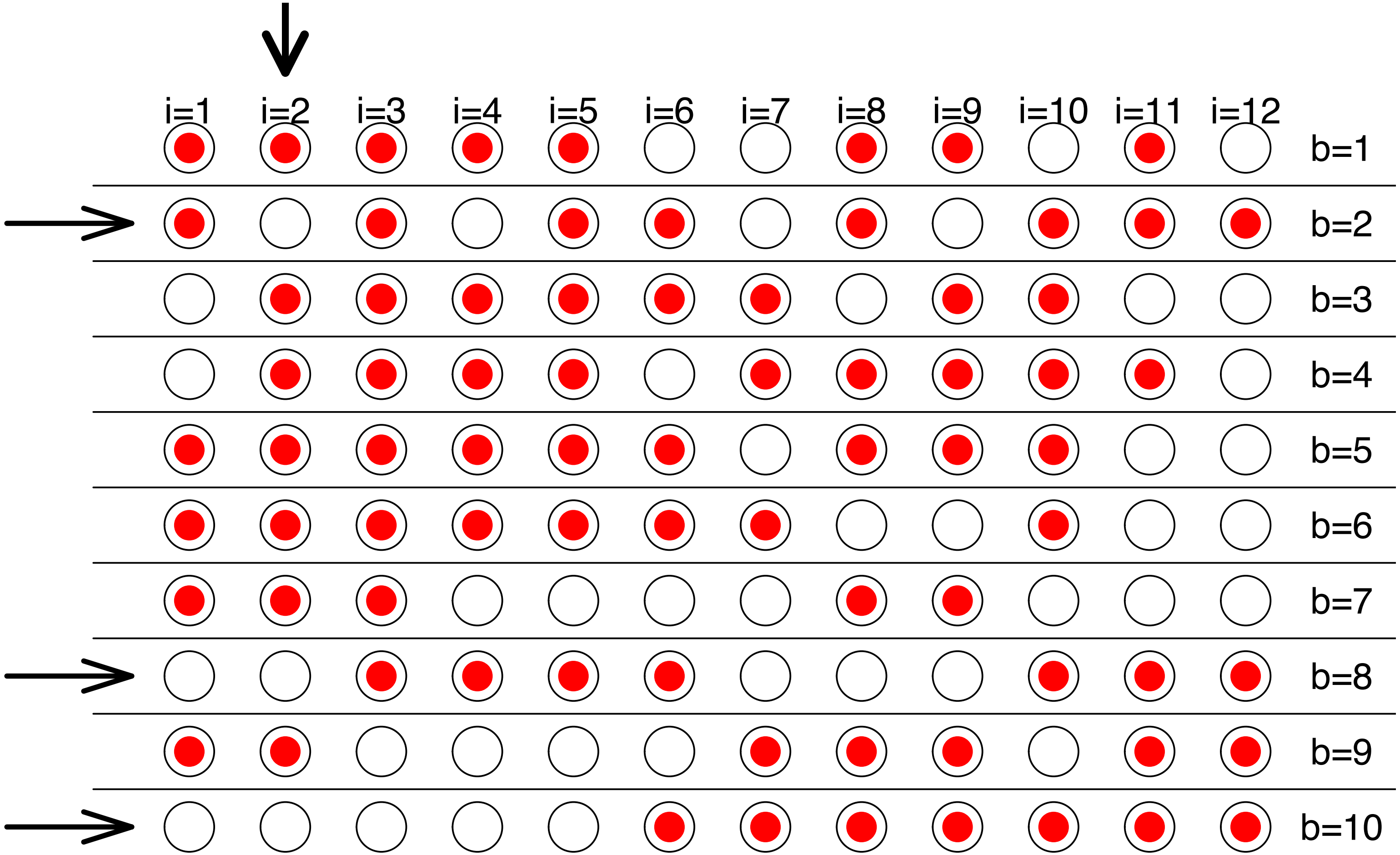
$$\widehat{R}_{test} = n^{-1} \sum_{i=1}^{n} L(Y_i, \hat{Y}_i^{oob}).$$



	↓ i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10	i=11	i=12	
	◉	◉	◉	◉	◉	○	○	◉	◉	○	◉	○	b=1
	◉	○	◉	○	◉	◉	○	◉	○	◉	◉	◉	b=2
→	○	◉	◉	◉	◉	◉	◉	○	◉	◉	○	○	b=3
→	○	◉	◉	◉	◉	○	◉	◉	◉	◉	◉	○	b=4
	◉	◉	◉	◉	◉	◉	○	◉	◉	◉	○	○	b=5
	◉	◉	◉	◉	◉	◉	◉	○	○	◉	○	○	b=6
	◉	◉	◉	○	○	○	○	◉	◉	○	○	○	b=7
→	○	○	◉	◉	◉	◉	○	○	○	◉	◉	◉	b=8
	◉	◉	○	○	○	○	◉	◉	◉	○	◉	◉	b=9
→	○	○	○	○	○	◉	◉	◉	◉	◉	◉	◉	b=10

In this example with $B = 10$ and $n = 12$, to get prediction for $i = 1$, average only over trees $\hat{Y}^{*,b}(X_1)$ with $b \in \{3, 4, 8, 10\}$, e.g.

$$\hat{Y}_1^{oob} = \frac{1}{4} \sum_{b \in \{3,4,8,10\}} \hat{Y}^{*,b}(X_1).$$



	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10	i=11	i=12	
	●	●	●	●	●	○	○	●	●	○	●	○	b=1
→	●	○	●	○	●	●	○	●	○	●	●	●	b=2
	○	●	●	●	●	●	●	○	●	●	○	○	b=3
	○	●	●	●	●	○	●	●	●	●	●	○	b=4
	●	●	●	●	●	●	○	●	●	●	○	○	b=5
	●	●	●	●	●	●	●	○	○	●	○	○	b=6
	●	●	●	○	○	○	○	●	●	○	○	○	b=7
→	○	○	●	●	●	●	○	○	○	●	●	●	b=8
	●	●	○	○	○	○	●	●	●	○	●	●	b=9
→	○	○	○	○	○	●	●	●	●	●	●	●	b=10

For predicting observation $i = 2$, average only over trees $\hat{Y}^{*,b}(X_2)$ with $b \in \{2, 8, 10\}$.

$$\hat{Y}_2^{oob} = \frac{1}{3} \sum_{b \in \{2,8,10\}} \hat{Y}^{*,b}(X_2).$$

We clearly need to average over many bootstrap samples in practice to get accurate results, e.g. $|\tilde{B}_i|$ needs to be reasonably large for all $i = 1, \ldots, n$.

What is the relation between $|\tilde{B}_i|$ and B?
The probability π^{oob} of an observation NOT being included in a bootstrap sample $(j_1, \ldots, j_n)$ (and hence being 'out-of-bag') is, as all j_k for $k = 1, \ldots, n$ are drawn with replacement from $\{1, \ldots, n\}$,

$$\pi^{oob} = \prod_{i=1}^{n} (1 - \frac{1}{n}) \quad \overset{n \to \infty}{\to} \quad \exp(-1) \approx 0.367.$$

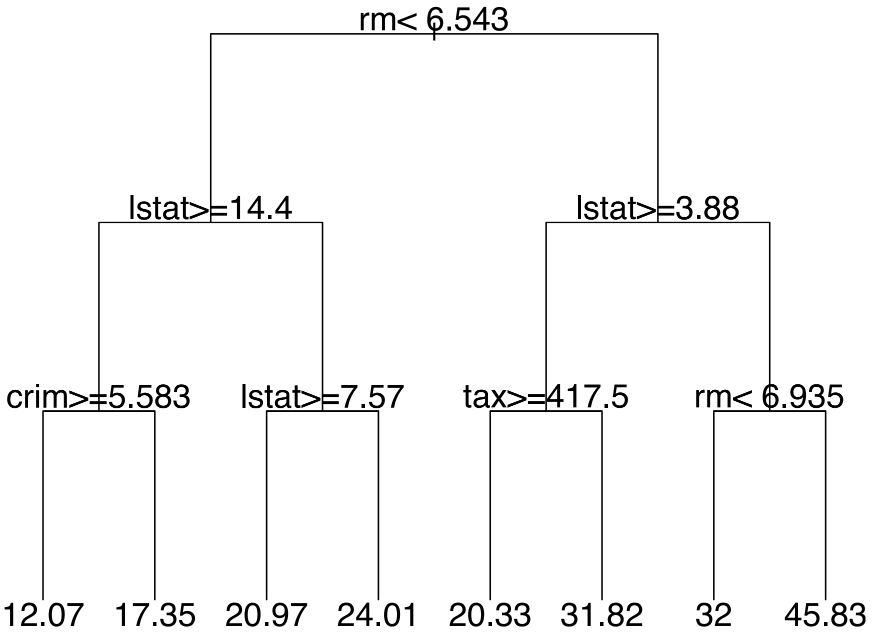
Hence $E(|\tilde{B}_i|) = \exp(-1) \cdot B \approx 0.367 \cdot B$ for all $i = 1, \ldots, n$.
In practice, number of bootstrap samples B is typically between 200 and 1000, meaning that the number $|\tilde{B}_i|$ of out-of-bag samples will be approximately in the range $70 - 350$. The obtained test error estimate is asymptotically unbiased for large number B of bootstrap samples and large sample size n.

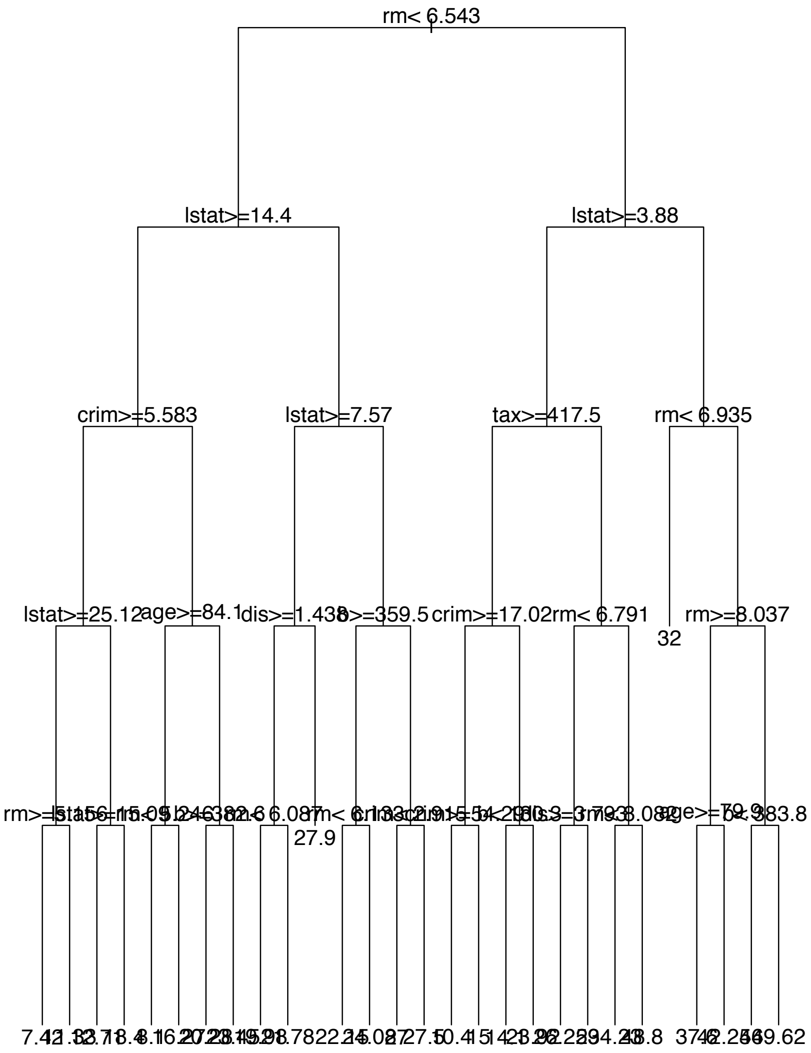
Apply out of bag estimation to select optimal tree depth and assess performance of bagged trees for Boston Housing data.
Use the entire dataset with $p = 13$ predictor variables. Fit first an ordinary tree of depth $d \in \{1, 2, 3, \ldots, 30\}$.

```
n <- nrow(BostonHousing)    ## n samples

X <- BostonHousing[,-14]
Y <- BostonHousing[,14]

maxdepth <- 3              ## fit here trees of depth 3
                           ## use function ‘rpart’ to fit tree
tree <- rpart(Y ~ ., data=X ,
       control=rpart.control(maxdepth=maxdepth,minsplit=2))
```

Plot trees of depth $d = 3$ and $d = 5$.

```
plot(tree,  margin=.1,  uniform=TRUE)
text(tree,  cex=1.3)
```





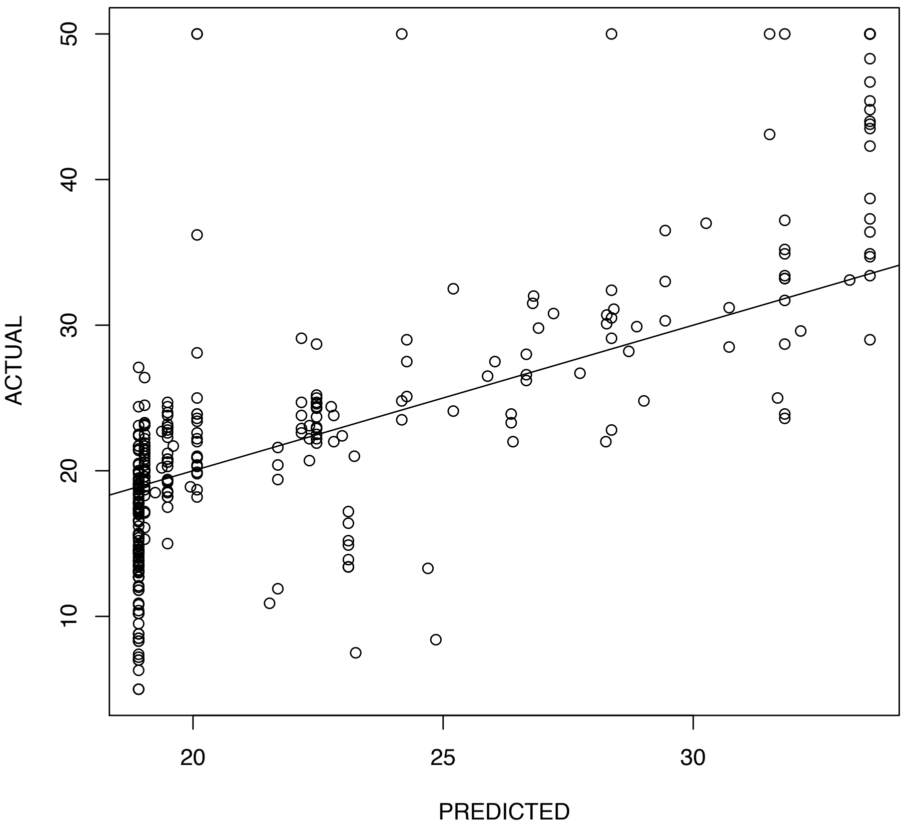
Bagging with $B = 100$ bootstrap samples, computing the out-of-bag (OOB) estimate of prediction error.

```
B <- 100
prediction_oob <- rep(0,length(Y))     ## vector with oob predictions
numbertrees_oob <- rep(0,length(Y))    ## how many oob trees
                                       ## for each sample ?

for (b in 1:B){                        ## loop over bootstrap samples
  subsample <- sample(1:n,n,replace=TRUE)      ## "in-bag" samples
  outofbag <- (1:n)[-subsample]                ## "out-of-bag" samples

                                       ## fit tree on "in-bag" samples
  treeboot <- rpart(Y ~ ., data=X, subset=subsample,
        control=rpart.control(maxdepth=maxdepth,minsplit=2))

                                       ## predict on oob-samples
  prediction_oob[outofbag] <- prediction_oob[outofbag] +
                     predict(treeboot, newdata=X[outofbag,])
  numbertrees_oob[outofbag] <- numbertrees_oob[outofbag] +  1
}
## final oob-prediction is average across all "out-of-bag" trees
prediction_oob <- prediction_oob / numbertrees_oob
```

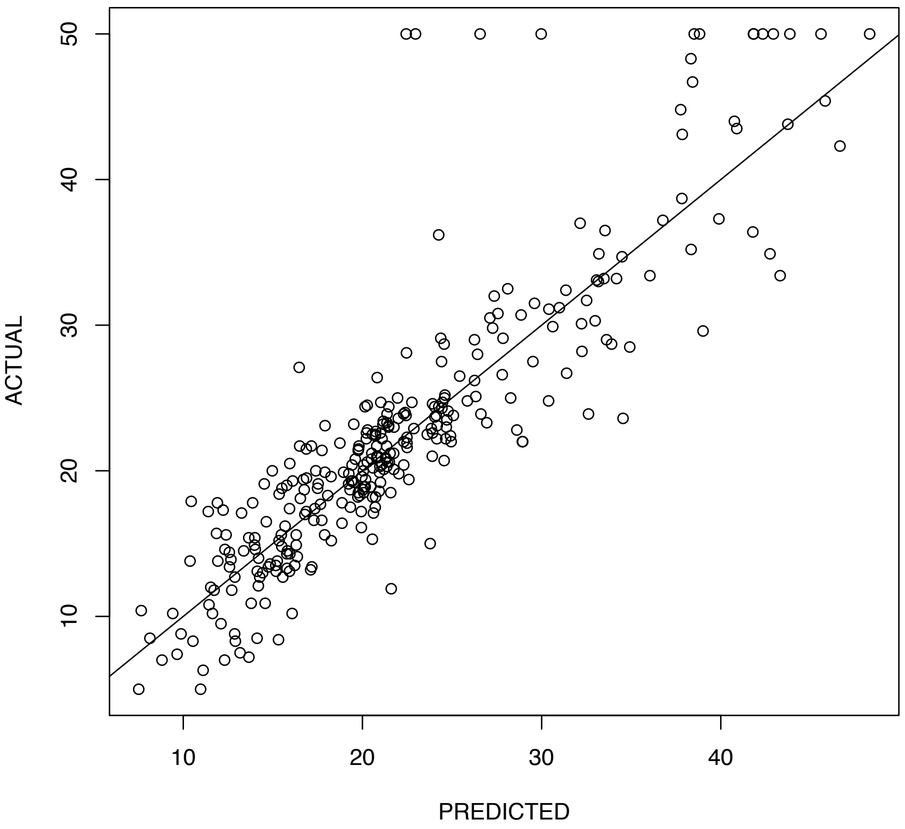
Plot out-of-bag predictions.

```
plot(prediction_oob, Y,  xlab="PREDICTED",  ylab="ACTUAL")
```

For depth $d = 1$.



For depth $d = 10$.



Out-of-bag estimates of test error

$$E\big((\hat{Y} - Y)^2\big)$$

as a function of tree depth d. Table shows CV-mean squared error loss (with out-of-bag prediction for the bagged estimator).

tree depth d	1	2	3	4	5	10	30
single tree $\hat{Y}$	60.7	44.8	32.8	31.2	27.7	26.5	27.3
bagged trees $\hat{Y}_{Bag}$	43.4	27.0	22.8	21.5	20.7	20.1	20.1

Without bagging, the optimal tree depth seems to be $d = 10$. With bagging, we could also take the depth up to $d = 30$.

Bagging strongly improves performance.

On the other hand, bagged trees cannot be displayed as nicely as single trees and some of the interpretability of trees is lost.

For classification, it is easily possible to construct (artifical) examples where bagging leads to a deterioration of performance.

Consider a two-class problem $Y \in \{0, 1\}$. Suppose all labels are truly $Y = 1$ and there is a random predictor $\hat{Y}$ which predicts

$$\hat{Y} = \begin{cases} 1 & \text{with probability } 0.3 \\ 0 & \text{with probability } 0.7 \end{cases} .$$

This classifier would have a misclassification error of 0.7.

Now bag this classifier by taking a mean $\hat{Y}_{Bag} = \sum_{b=1}^{B} \hat{Y}^{*,b}$ and classify by majority decision among all bagged trees, i.e. classify as $Y = 1$ if and only if $\hat{Y}_{Bag} > 0.5$.

The misclassification error of the bagged trees is now 1 and bagging made a bad predictor even worse.

Bagging trees typically improves prediction for real-life datasets. Consider the following datasets.

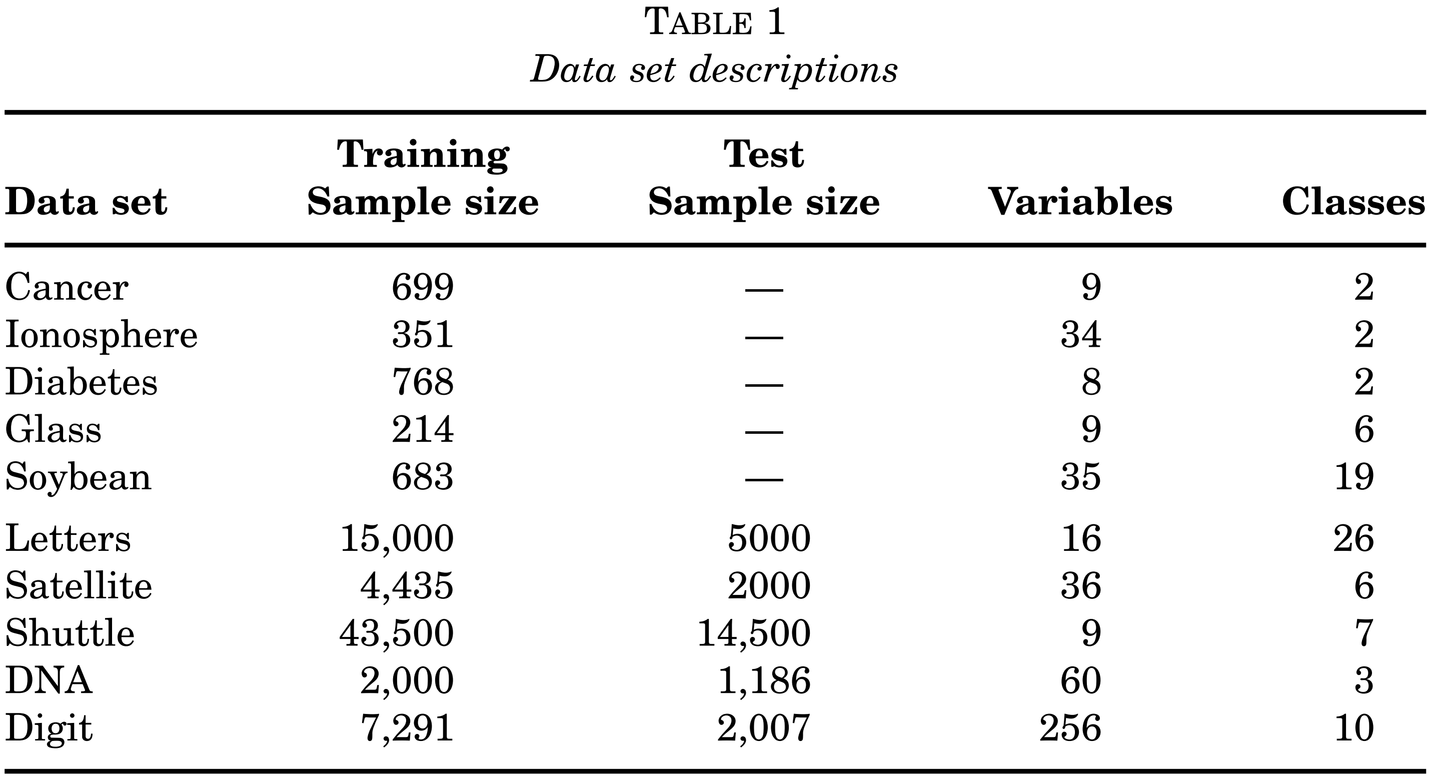


TABLE 1
Data set descriptions

Data set	Training Sample size	Test Sample size	Variables	Classes
Cancer	699	—	9	2
Ionosphere	351	—	34	2
Diabetes	768	—	8	2
Glass	214	—	9	6
Soybean	683	—	35	19
Letters	15,000	5000	16	26
Satellite	4,435	2000	36	6
Shuttle	43,500	14,500	9	7
DNA	2,000	1,186	60	3
Digit	7,291	2,007	256	10

Both trees and bagged trees (Forests) are fitted on these data.

The misclassification errors on the test sets for single trees and bagged trees (‘Forests’).

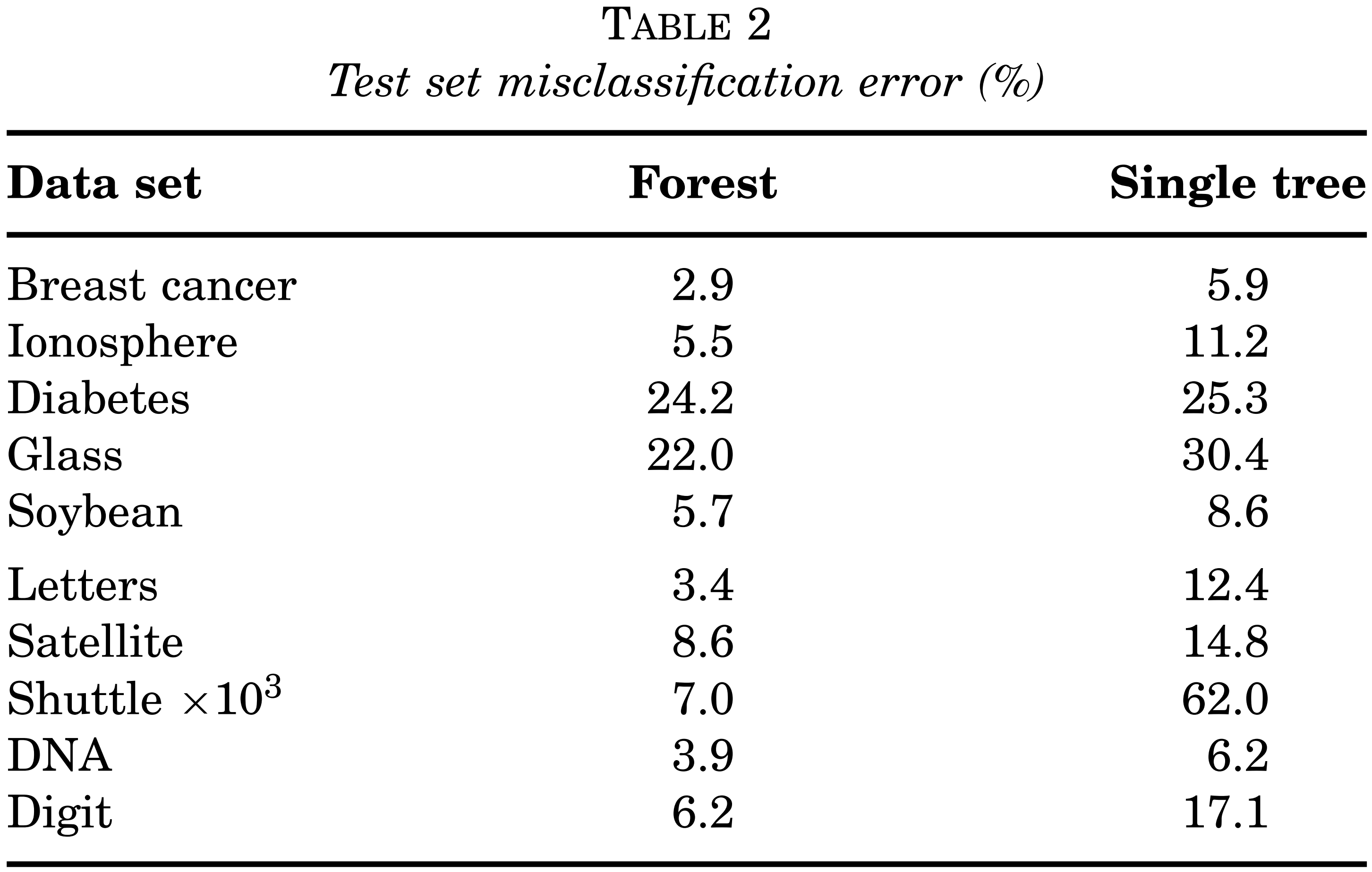


TABLE 2

Test set misclassification error (%)

Data set	**Forest**	**Single tree**
Breast cancer	2.9	5.9
Ionosphere	5.5	11.2
Diabetes	24.2	25.3
Glass	22.0	30.4
Soybean	5.7	8.6
Letters	3.4	12.4
Satellite	8.6	14.8
Shuttle $\times 10^3$	7.0	62.0
DNA	3.9	6.2
Digit	6.2	17.1

from Breiman: ‘Statistical Modelling: the two cultures’.

Note that ‘Forests’ are not standard bagged trees, but so-called Random Forests, which employ additional randomization (more later).