ARTICLE PRESENTATION ←

or just in xls

send info on paper for me to
add to wikipage!

22.02 ~ 22.03

30 min: 20 min
+
10 min ← discussant : group email if preferences
me for.
article

!

The basic ← not too elaborate maths

take home message → trend on lightning talks

Bagging

Bagging can be used with different regression and classification methods, but we will focus on trees.

High variance of trees

Decision trees often suffer from high variance.

- By this we mean that the trees are sensitive to small changes in the predictors:
- If we change the observation set, we may get a very different tree.
- This is due to the fact that small changes in the data can result in a large effect on which splits is done.
- A small effect on the top level is propagated down in the tree.

For the Boston data, we saw that changing the train/test split gave very different trees

To reduce the variance of decision trees we can apply *bootstrap aggregating* (*bagging*), invented by Leo Breiman in 1996 (after he retired in 1993).

1) Motivation: $X_1, .., X_B$ i.i.d $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

$$Var(\underset{\overset{\uparrow}{\frac{1}{B}\sum_{b=1}^{B} X_b}}{\bar{X}}) = \frac{\sigma^2}{B}$$

2) Enter tree: $\underline{X}$ is replaced by $\overset{\swarrow \text{ tree}}{\hat{f}}(\underline{X})$

3) Do not have B indep. data sets to be used to make B trees

$\Rightarrow$ use bootstrap samples:

$$\begin{tikzcd} Z=(x,y) \arrow[d, "\text{bootstrap}"'] \\ Z^{*1}, Z^{*2}, \dots, Z^{*B} \end{tikzcd}$$

F P for Z

$\hat{F} = \frac{1}{N}$ for each pair Z

new data sets of size N

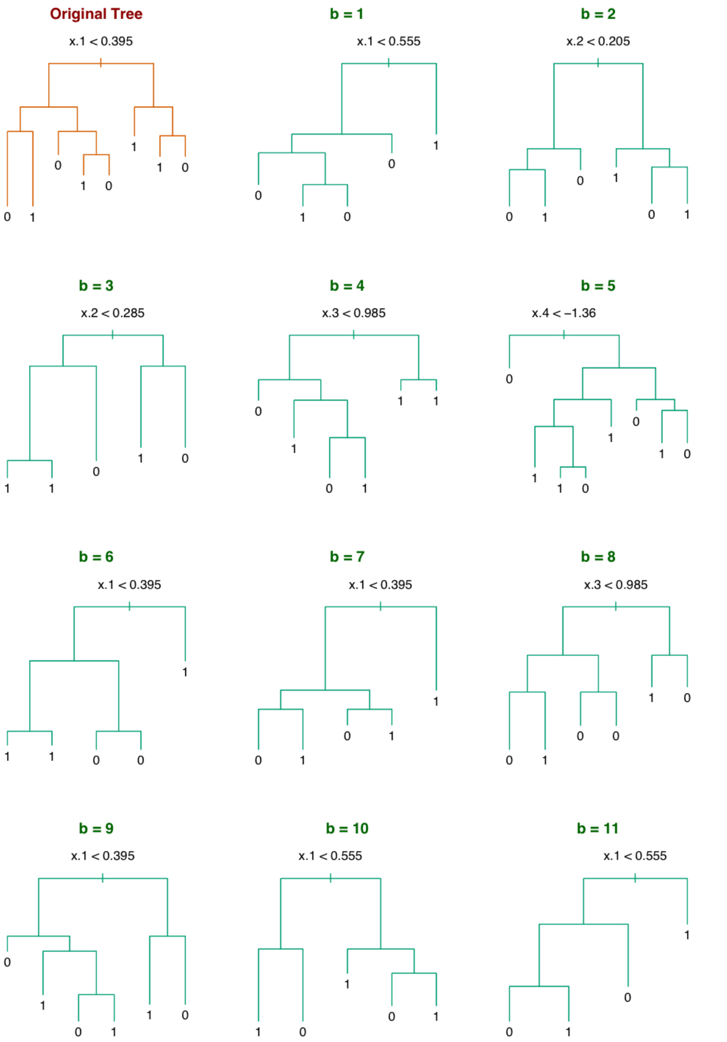
$\hat{f}^{*1}(x)$ $\underline{\hat{f}^{*b}(x)}$ $b = 1, ..., B$

Bootstrap samples and trees

For each bootstrap sample we construct a decision tree, $\hat{f}^{*b}(x)$ with $b = 1, ..., B$, and we then use information from all of the trees to draw inference.

Study Figure 8.9 to see variability of trees - observe that different covariates are present in each tree. All features are equally correlated 0.95. Observe

- trees variable (high variance)
- bagging will smooth out this variance to reduce the test error



$N = 30$

$p = 5$ covariates

$\text{Corr}(x_j, x_\ell) = 0.95$

FIGURE 8.9. *Bagging trees on simulated dataset. The top left panel shows the original tree. Eleven trees grown on bootstrap samples are shown. For each tree, the top split is annotated.*

Bagging regression trees

For regression trees, we take the average of all of the predictions and use this as the final result:

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

- Bagging works well under squared loss.
- It can be shown that true population aggregation never increases the MSE (ELS Eq 8.52).
- Thus, we expect a decrease in MSE with bagging.
- However, a “similar result” is not true for classification trees.

$Z = \{(x_i, y_i)\}_1^N$ independently drawn from $\mathcal{P}$

$Z^* = \{(x_i^*, y_i^*)\}_1^N$ "bootstrap" sample from $\mathcal{P}$

$\uparrow$ NB not from $\hat{\mathcal{P}}$

at some fixed x

$f_{ag}(x) = E_{\mathcal{P}}(\hat{f}^*(x))$ $\leftarrow$ will be est by $\frac{1}{B}\sum \hat{f}^{*b}(x)$ if bagging

an estimate based on a Z^*

expected value over $\mathcal{P}$

aggregation

one tree

$$E_{\mathcal{P}}[(Y - \hat{f}^*(x))^2] = E_{\mathcal{P}}\left[\left(Y - f_{ag}(x) + f_{ag}(x) - \hat{f}^*(x)\right)^2\right]$$

$$= E_{\mathcal{P}}\left[(Y - f_{ag}(x))^2 + (f_{ag}(x) - \hat{f}^*(x))^2 + \underbrace{2(Y - f_{ag}(x))(f_{ag}(x) - \hat{f}^*(x))}_{0}\right]$$

since $f_{ag} = E_{\mathcal{P}}(\hat{f}^*(x))$

$$= E_{\mathcal{P}}((Y - f_{ag}(x))^2) + \underbrace{E_{\mathcal{P}}((f_{ag}(x) - \hat{f}^*(x))^2)}_{\geq 0}$$

$\uparrow$ variance of $\hat{f}^*(x)$ since $E((\hat{f}^*(x) - E(\hat{f}^*(x)))^2 = \text{var}$.

$$\geq E_{\mathcal{P}}((Y - f_{ag}(x))^2)$$

$\Rightarrow$ the true pop. aggregation never increase MSE.

Bagging classification trees

For classification trees there are two possible ways to use the tree ensemble of B trees:

consensus

- we record the predicted class (for a given observation x) for each of the B trees and
- use the most occurring classification (majority vote) as the final prediction.
- (It is not wise to use the voting proportions to get posterior probabilites. For example is P(class 1)=0.75 but all B trees votes for 1.)

Let $\hat{G}(x)$ be the estimated class, and

$$\hat{G}(x) = \text{argmax}_k q_k(x)$$

where $q_k(x)$ is the proportion of the trees voting k, $k = 1, \ldots, K$.

probability

- alternatively average ~~posterior~~ probabilities for each class $\hat{p}^b_m$,
- and then choose the class with the largest probability.

$$\hat{G}(x) = \text{argmax}_k \frac{1}{B} \sum_{b=1}^{B} p^b_k(x)$$

where $p^b_k(x)$ is estimated probability for class k for tree b at x.

We examine this by an example: $K=2$ $p(\text{class } 1)$

Suppose we have $B = 5$ (no, B should be higher - this is only for illustration) classification tree and have obtained the following 5 estimated probabilities: $\{0.4, 0.4, 0.4, 0.4, 0.9\}$. If we average the probabilities, we get 0.5, and if we use a cut-off value of 0.5, our predicted class is 1. However, if we take a majority vote, using the same cut off value, the predicted classes will be $\{0, 0, 0, 0, 1\}$. The predicted class, based on a majority vote, would accordingly be 0.

The two procedures thus have their pros and cons: * By averaging the predictions no information is lost. We do not only get the final classification, but the probability for belonging to the class 0 or 1. * However, this method is not robust to outliers. By taking a majority vote, outliers have a smaller influence on the result.

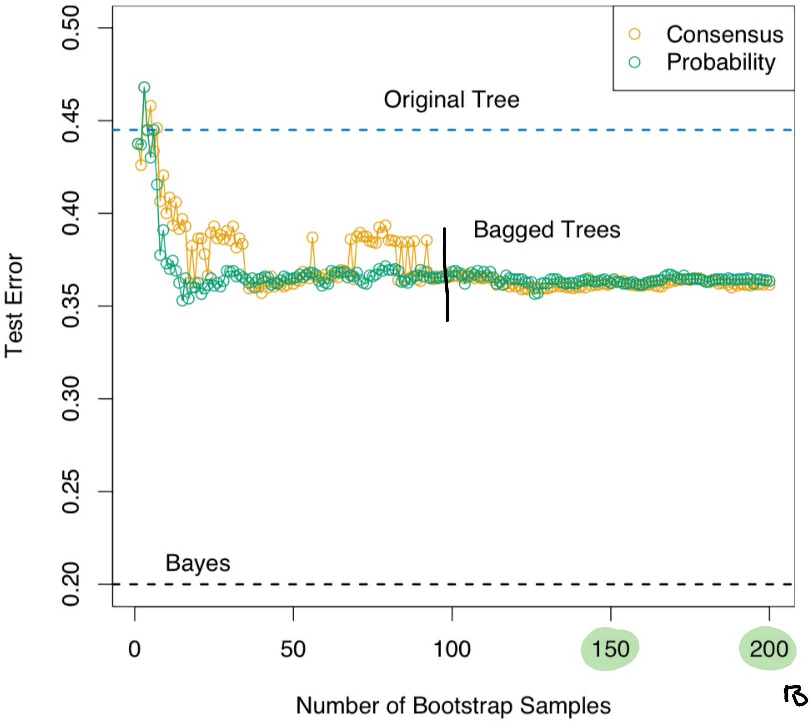
- According to ELS (page 283) the probability method will give bagged classifiers with lower variance that with the consensus method.

- Bagging a good classifier can make it better, but bagging a bad classifier can make it worse. (Not as for MSE for regression tree.)

Choosing B

- The number B is chosen to be as large as "necessary".
- An increase in B will not lead to overfitting, and B is not regarded as a tuning parameter.
- If a goodness of fit measure is plotted as a function of B (soon) we see that (given that B is large enough) increasing B will not change the goodness of fit measure.

(Study Figure 8.10 to see effect of B and the two strategies.)



↖ ok with B ≥ 100

B

FIGURE 8.10. *Error curves for the bagging example of Figure 8.9. Shown is the test error of the original tree and bagged trees as a function of the number of bootstrap samples. The orange points correspond to the consensus vote, while the green points average the probabilities.*

Pruning trees?

Originally, Breiman (1996) suggested to prune each tree, but later research has found that it is better to leave the trees at maximal size (a bushy tree), to make the trees as different from each other as possible.

↑

stopping criterion

node size

Regression: $\frac{1}{B} \sum f^{*b}(x)$

Classification: majority vote (consensus)

probability $\frac{1}{B} \sum^{B} p_k^{*b}(x)$

Bagging algorithm

(We write out in class.)

1) Generate B bootstrap samples

2) Fit a tree to each bootstrap sample

3) Aggregate results:

Wisdom of the crowd

(ELS page 286)

the collective knowledge of a diverse and independent body of people typically exceeds the knowledge of any single individual, and can be harnessed by voting

Study Figure 8.11 in ELS - and what about https://tv.nrk.no/serie/alle-mot-1

Bagging: sadly not independent bodies.

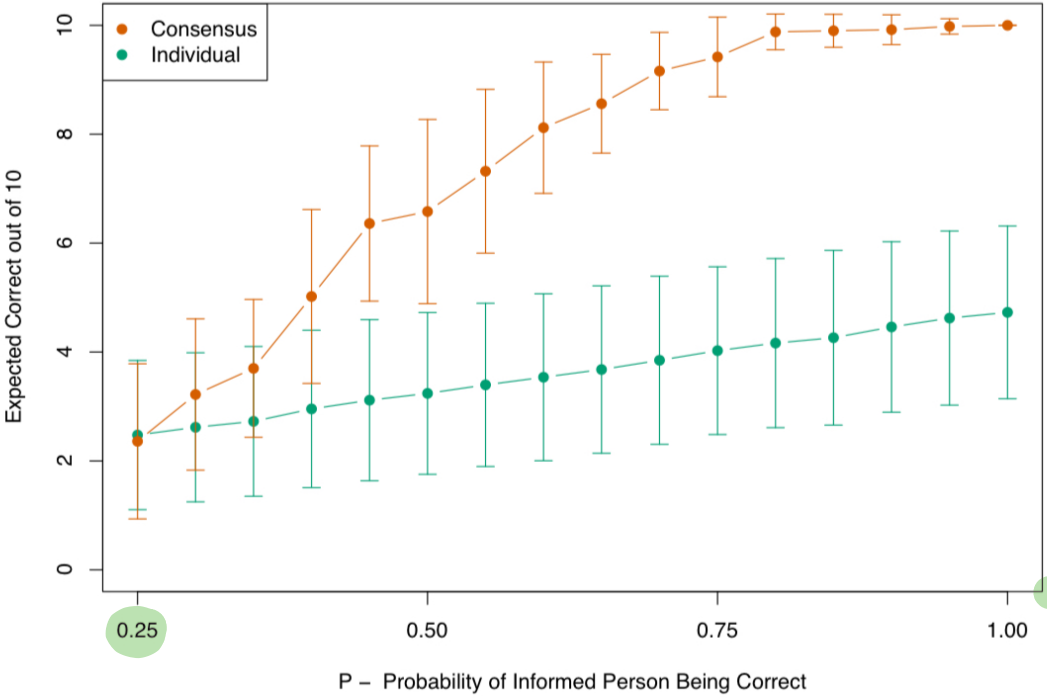


FIGURE 8.11. *Simulated academy awards voting.* 50 *members vote in 10 categories, each with* 4 *nominations. For any category, only* 15 *voters have some knowledge, represented by their probability of selecting the "correct" candidate in that category (so $P = 0.25$ means they have no knowledge). For each category, the* 15 *experts are chosen at random from the* 50*. Results show the expected correct (based on* 50 *simulations) for the consensus, as well as for the individuals. The error bars indicate one standard deviation. We see, for example, that if the* 15 *informed for a category have a* 50% *chance of selecting the correct candidate, the consensus doubles the expected performance of an individual.*

27.02.2021 new season - probably

tv.nrk.no/serie/alle-mot-1

↑

→ recognise wisdom of crowd?

Out-of-bag error estimation

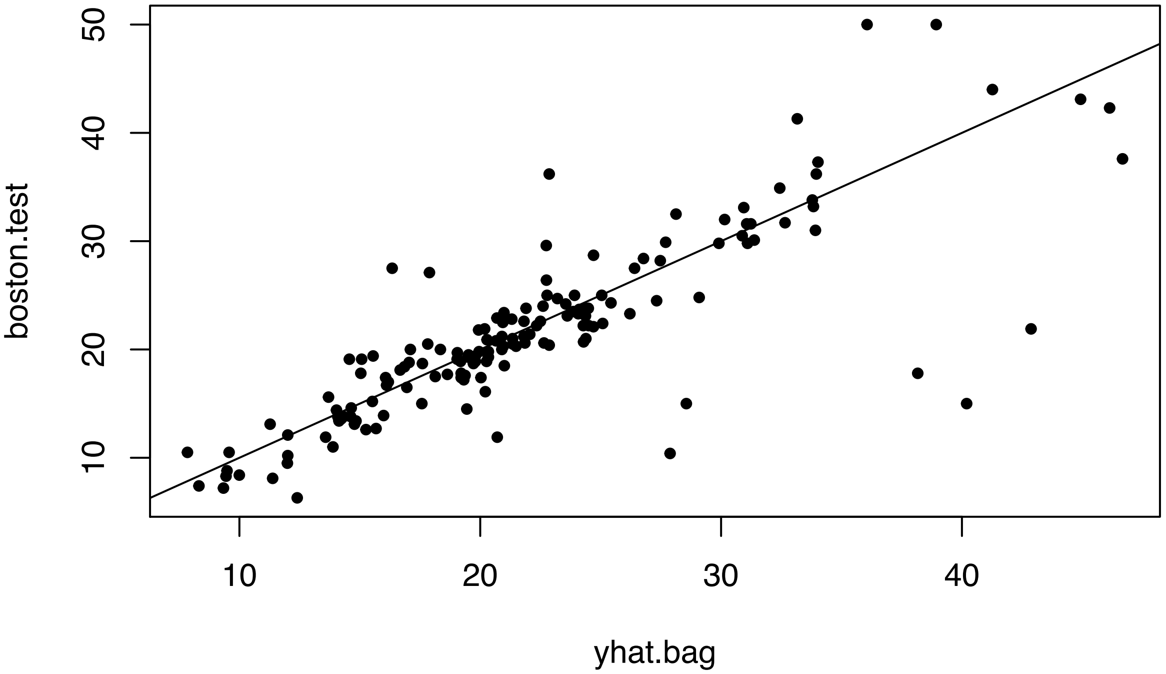
- We use a subset of the observations in each bootstrap sample. We know that the probability that an observation is in the bootstrap sample is approximately $1 - e^{-1}$=0.6321206 (0.63212).
- when an observation is left out of the bootstrap sample it is not used to build the tree, and we can use this observation as a part of a “test set” to measure the predictive performance and error of the fitted model, $f^{*b}(x)$.

In other words: Since each observation i has a probability of approximately 2/3 to be in a bootstrap sample, and we make B bootstrap samples, then observation i will be outside the bootstrap sample in approximately $B/3$ of the fitted trees.
The observations left out are referred to as the *out-of-bag* observations, and the measured error of the $B/3$ predictions is called the *out-of-bag error*.

```
##                      Number of trees: 500
## No. of variables tried at each split: 13
##
##           Mean of squared residuals: 12.00879
##                     % Var explained: 86.83
```

R-package
randomForest
mtry=p

Plotting predicted test values vs true values.



```
## [1] 23.12171
```

Error rate on test set for bagging

Remember that the error rate on the test set for a single tree was: 36.2318993.

Pima indians

Here the misclassification rate for the OOB is reported.

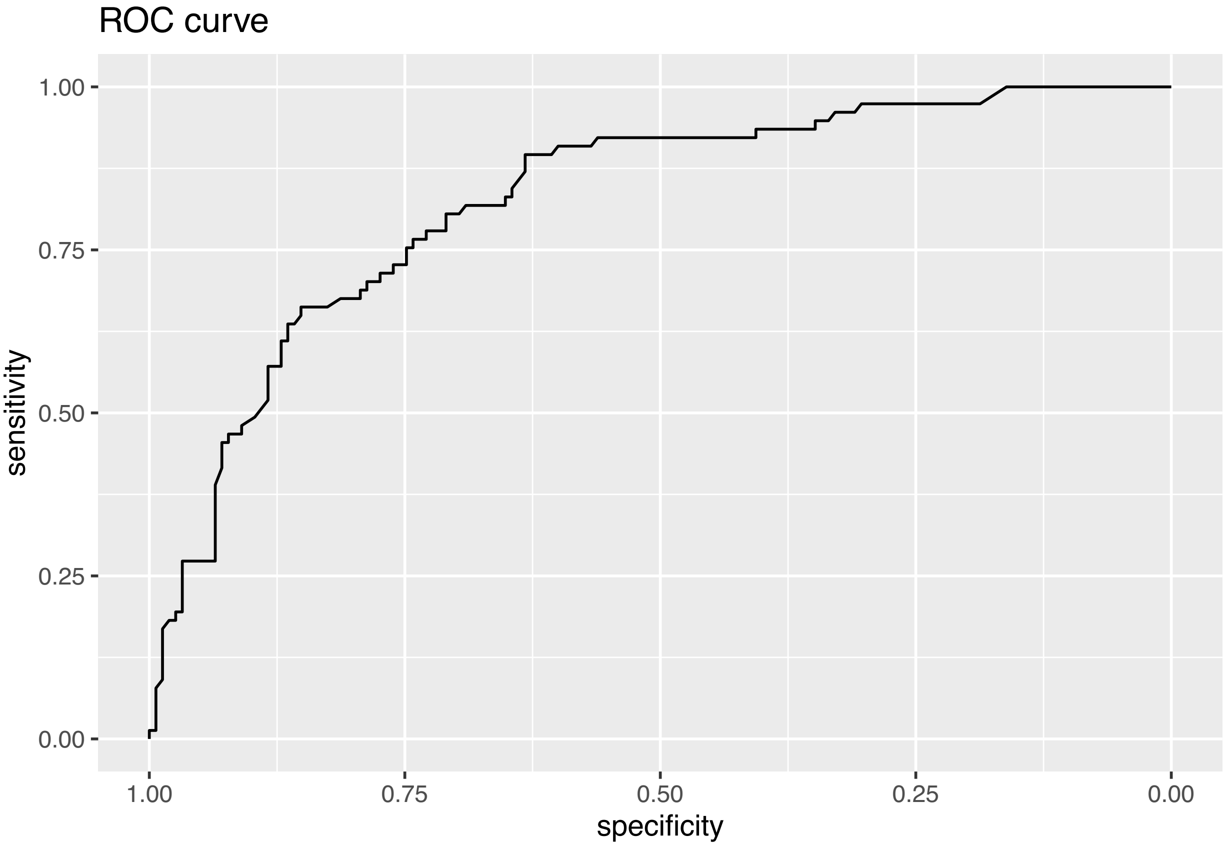
```
##
## Call:
##  randomForest(formula = as.factor(diabetes) ~ npreg + glu + bp +      skin + bmi + ped + age, data =
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 23.67%
## Confusion matrix:
##     0  1 class.error
## 0 167 33       0.165
## 1  38 62       0.380
```

```
## [1] "Evaluation on training data"
```

```
##  Accuracy
## 0.7633333
## [1] "Evaluation on test data"

##  Accuracy
## 0.7844828

## Area under the curve: 0.8257
```



Remember that the misclassification error rate on the test set for a single tree (after pruning) was: $1 - 0.75 = 0.25$.

When should we use bagging?

Bagging can be used for predictors (regression and classification) that are not trees, and according to Breiman (1996):

- the vital element is the instability of the prediction method
- if perturbing the learning set can cause significant changes in the predictor constructed, then bagging can improve accuracy.

Breiman (1996) suggests that these methods should be suitable for bagging:

- neural nets, classification and regression trees, subset selection in linear regression

however not nearest neighbours - since

- the stability of nearest neighbour classification methods with respect to perturbations of the data distinguishes them from competitors such as trees and neural nets.

Would you think that much is gained by bootstrapping MLR, logistic regression, or a lasso version of the two?

When should we use bagging?

Bagging can be used for predictors (regression and classification) that are not trees, and according to Breiman (1996):

- the vital element is the instability of the prediction method
- if perturbing the learning set can cause significant changes in the predictor constructed, then bagging can improve accuracy.

Breiman (1996) suggests that these methods should be suitable for bagging:

1996 version?

CART

- neural nets, classification and regression trees, subset selection in linear regression

however not nearest neighbours - since

- the stability of nearest neighbour classification methods with respect to perturbations of the data distinguishes them from competitors such as trees and neural nets. ?

Would you think that much is gained by bootstrapping MLR, logistic regression, or a lasso version of the two?

BUT: making many trees destroys the interpretability of the estimator.