Bagging

CS534 - Machine Learning

Yubin Park, PhD

Recall the bias-variance decomposition

for Squared Loss

$$\mathrm{E}[(y-f)^2] = \mathrm{Bias}[f]^2 + \mathrm{Var}[f] + \sigma^2$$

This time, we will decompose a bit differently:

$$\begin{aligned} \mathrm{E}[(y - f)^2] &= \mathrm{E}[((y - \mathrm{E}[f]) + (\mathrm{E}[f] - f))^2] \\ &= \mathrm{E}[(y - \mathrm{E}[f])^2] + \mathrm{E}[(\mathrm{E}[f] - f)^2] \\ &\geq \mathrm{E}[(y - \mathrm{E}[f])^2] \end{aligned}$$

Maybe too obvious.

If we can make f close to $\mathrm{E}[f]$

the expected loss will be less.

But, how?

Bagging (1)

Imagine that we know the "real" distribution for the samples: $(\mathbf{x}_i, y_i)$

To estimate $\mathrm{E}[f]$, we would repeat:

1. draw a set of samples
2. estimate f
3. repeat the above as many as possible
4. then aggregate all estimated f

The only caveat is that we do not know the "real" distribution.

Perhaps, we can "simulate" the real distribution with the samples we have?

BTW, What's the difference between 1) sampling from the distribution and 2) sampling from the samples?

Bagging (2)

Bagging (Bootstrap Aggregation) works as follows:

1. draw random samples from the data with "replacement"
2. estimate f
3. repeat B number of times

then get the final model by averaging as follows:

$$f_{\text{bagged}} = \frac{1}{B} \sum_{b=1}^{B} f_b$$

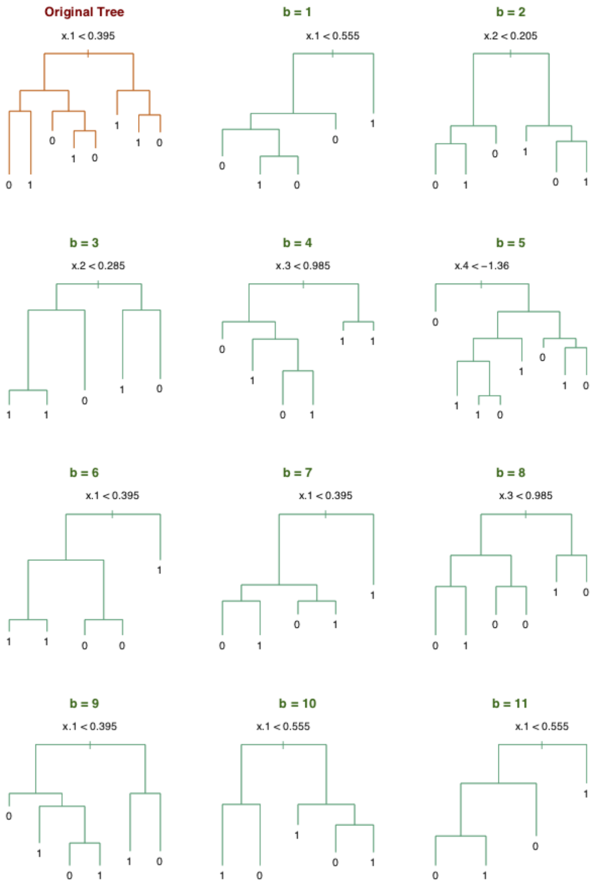

FIGURE 8.9. *Bagging trees on simulated dataset. The top left panel shows the original tree. Eleven trees grown on bootstrap samples are shown. For each tree, the top split is annotated.*

Chapter 8 of ESLII

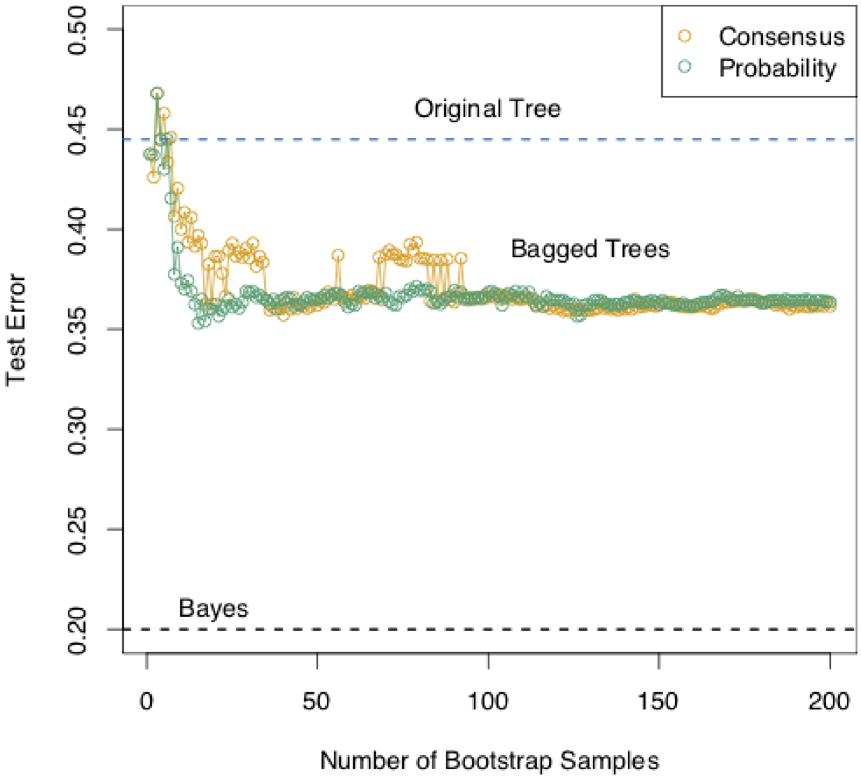

FIGURE 8.10. *Error curves for the bagging example of Figure 8.9. Shown is the test error of the original tree and bagged trees as a function of the number of bootstrap samples. The orange points correspond to the consensus vote, while the green points average the probabilities.*

Chapter 8 of ESLII

Model Averaging and Stacking

One step beyond the simple averaging models:

$$\mathrm{E}[(y - \sum_{b=1}^{B} w_b f_b)^2] \leq \mathrm{E}[(y - \frac{1}{B}\sum_{b=1}^{B} f_b)^2]$$

If $w_b = \frac{1}{B}$, then the both sides are equal.

How do we estimate the weights, w_b?

We will divide the training set into two parts:

- for fitting f_b
- for estimating w_b

This can be viewed as stacking two layers of models, thus Model Stacking.

Questions?