Topic 2. The Bias-Variance Tradeoff

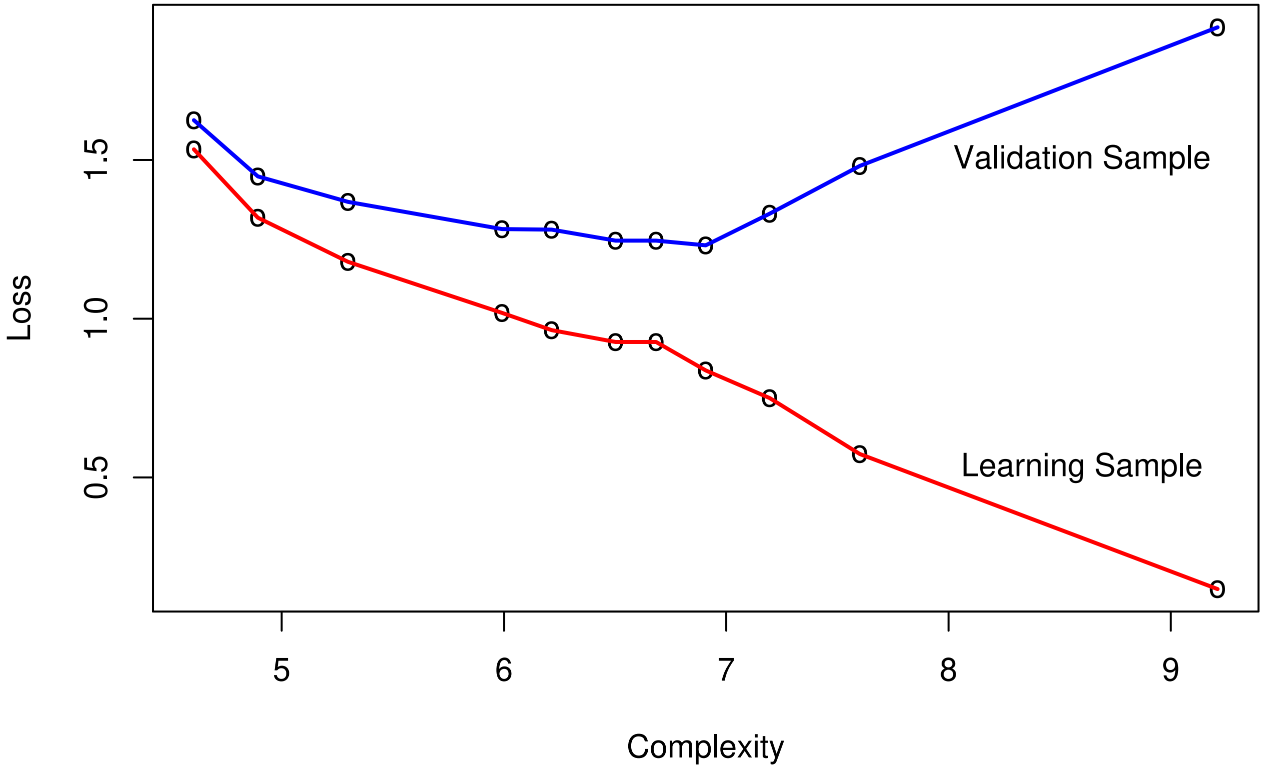
Case 2: Single Target Single Feature

Reading Assignment

Berry and Linoff (2000)

- Pages 111–120. Decision trees (review).
- Pages 119. Study figure carefully and compare to next slide.

Fig 22. Learning and Validation Loss



The purpose of this topic is to explain why every tool will produce a picture like this as complexity increases: Loss is underestimated in the learning sample and correctly estimated in the validation sample. See also Berry and Linoff p. 119. Complexity in regression increases when a variable is added; in nearest neighbors when the number of neighbors decreases, in a neural net when a hidden layer is added, in a decision tree when a leaf is added. Correct complexity is where validation loss is minimized.

The Bias-Variance Tradeoff

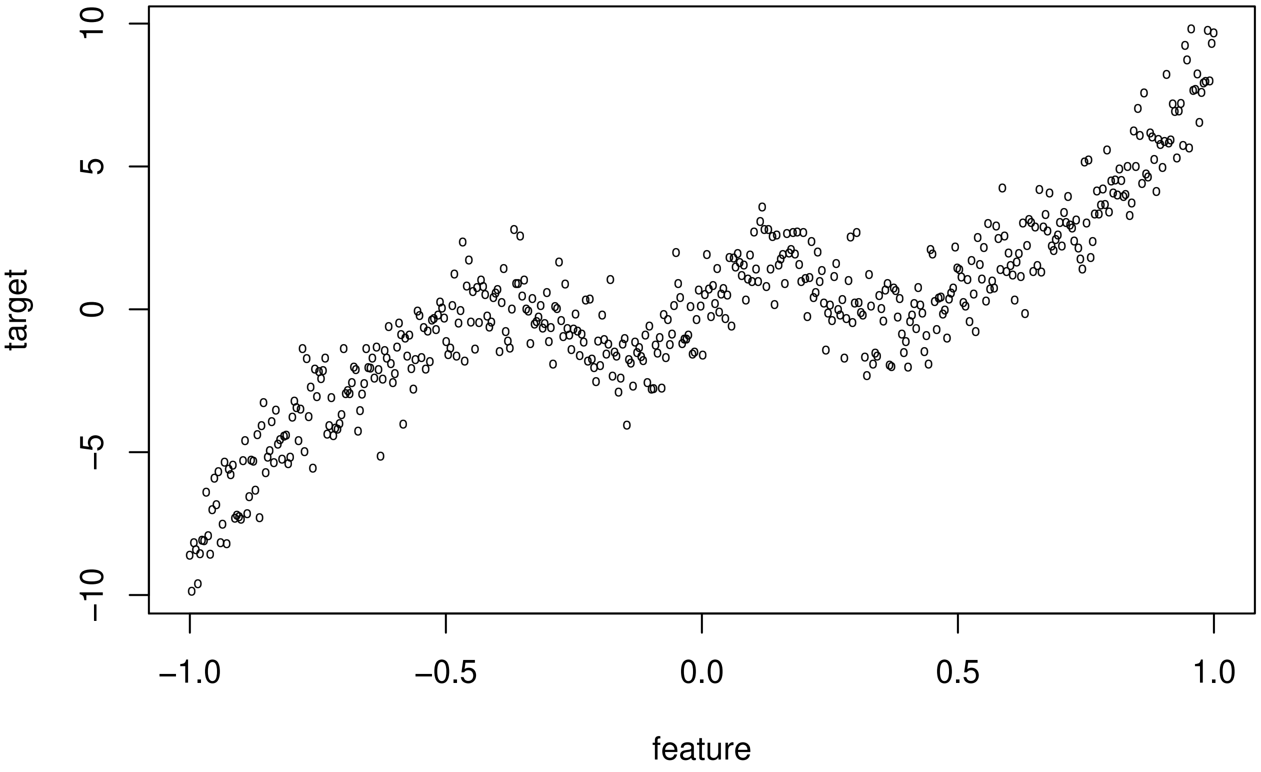
To understand Figure 22 we shall

- Examine the consequences of underfitting and overfitting.

- Show why validation samples are essential in data mining.

- Incidentally, become more familiar with trees.

Fig 23. Simulated Training Data



One numeric feature x and one numeric target y.
Sample size $n = 500$.

The Truth

For these data we could learn the truth from the data alone because:

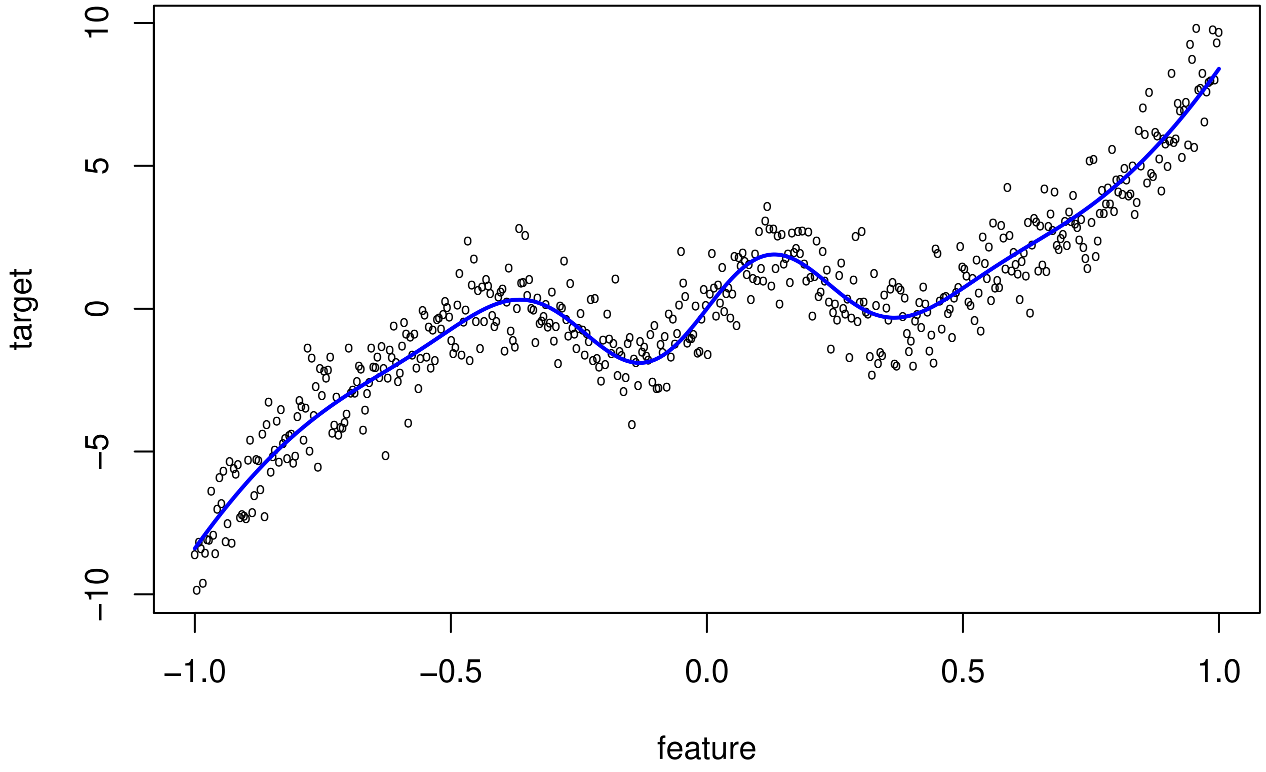
- The dimension is low. Graphical methods have a chance.

- The signal to noise ratio is high. Graphical methods will actually work, e.g. Fig 23.

- Even if graphs do not suggest useful derived features so that we can use regression; trees, neural nets, and localization methods will work because we have enough data and the signal to noise ratio is high.

- Because these data are simulated, we can actually plot the truth, here it is —

Fig 24. The Truth



The model is $y = f(x)+e$, where f is shown in blue and e is normally distributed with $\sigma = 1$. Sample size $n = 500$.

In Fig 24 notice that:

- The best prediction of y_i at feature x_i is $f(x_i)$, which is the ordinate of the blue line at abscissa x_i.
- No other prediction will generalize as well to future samples.
- The error made by this prediction is $e_i = y_i - f(x_i)$.
- The sample variance of this prediction is

$$s_e^2 = \frac{1}{n}\sum_{t=1}^{n} e_i^2$$

which converges to its population value of σ_e^2 as n increases.

Decision Trees

Decision trees work a little differently in prediction:

1. Start with feature x_1 and divide the sample into two groups depending on whether the feature is to the left, $x_1 < c$, or the right, $c < x_1$, of the cut c. In each group, the prediction is the average of the y_i in that group.

2. Compute the loss over the sample; try all possible cuts; remember the cut with smallest loss. Do this for every feature and choose the feature and cut with smallest loss.

3. Work down the tree, subdividing groups using the same procedure.

4. Stop when some measure of complexity such as the depth of the tree or the number of leaves is reached.

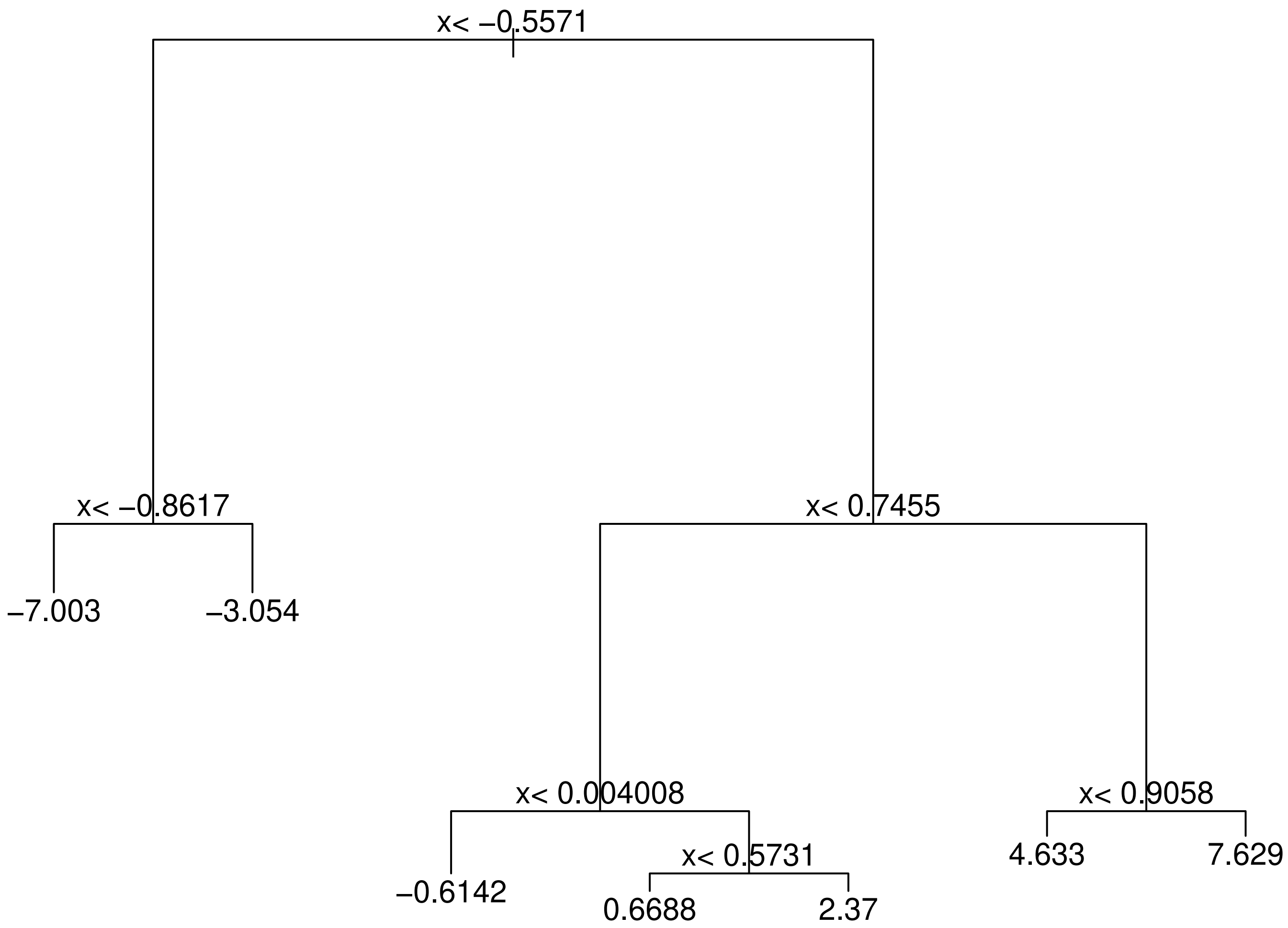
For prediction, MSE, which is the mean of the squared prediction errors, is the usual loss function.

In this particular case there is only one feature x_1 so we are just cutting the x-axis into pieces.

The Default Tree

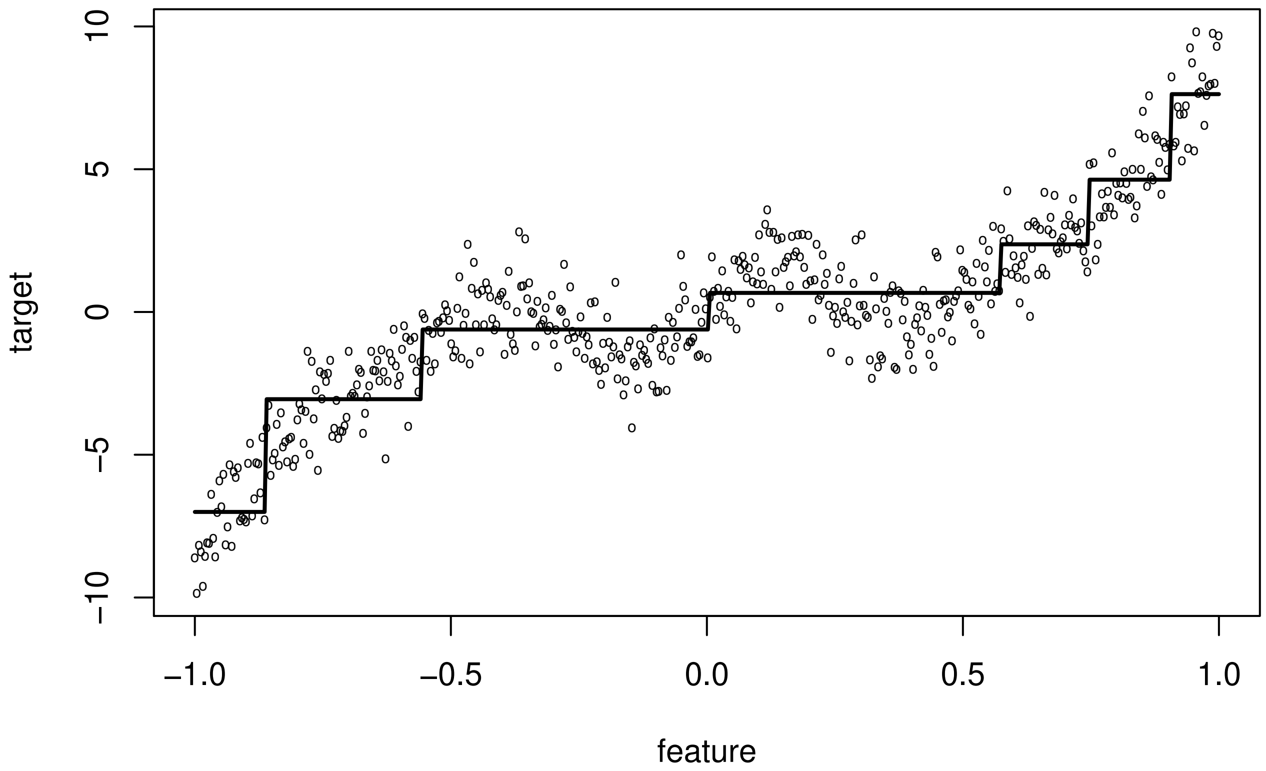
First we shall fit the default tree.

Fig 25. The Default Tree



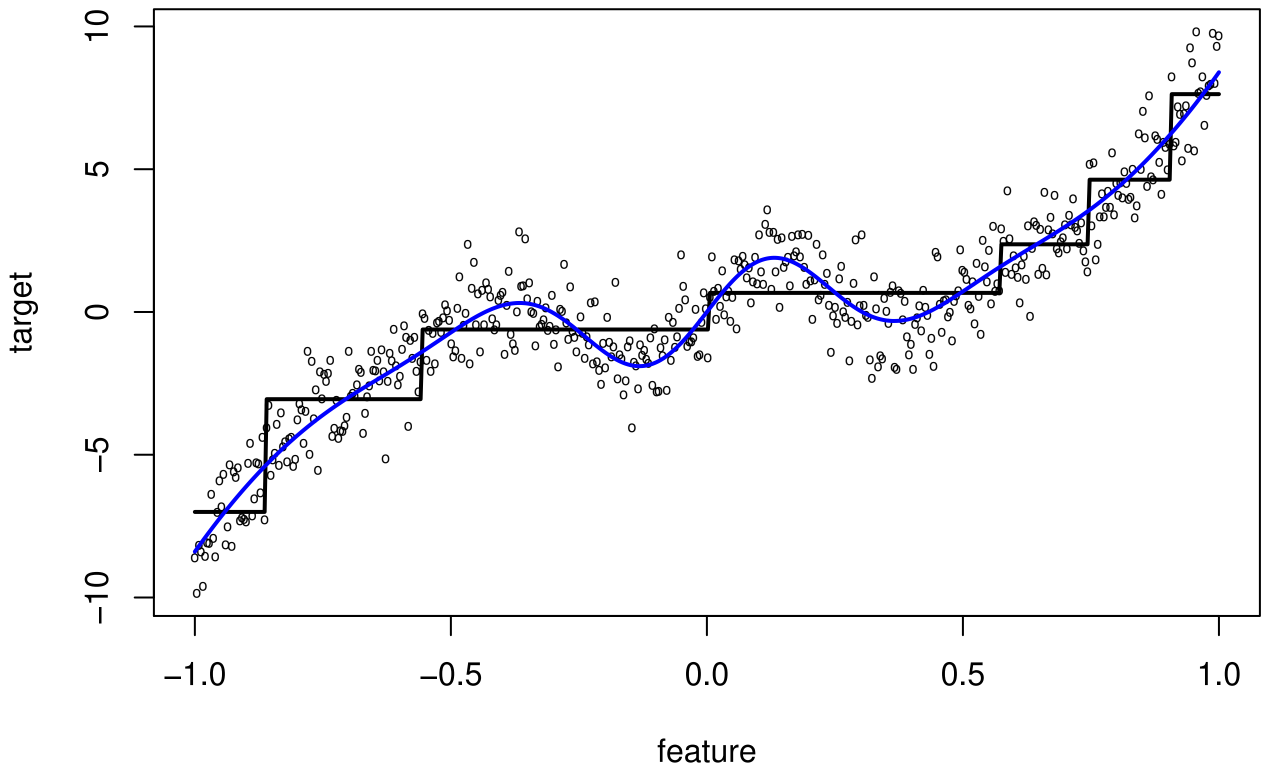
The default tree has seven terminal nodes. The left branch corresponds to a feature x that satisfies the inequality. The prediction at a leaf is the average of the y_i at that leaf.

Fig 26. Default Tree Compared to Data



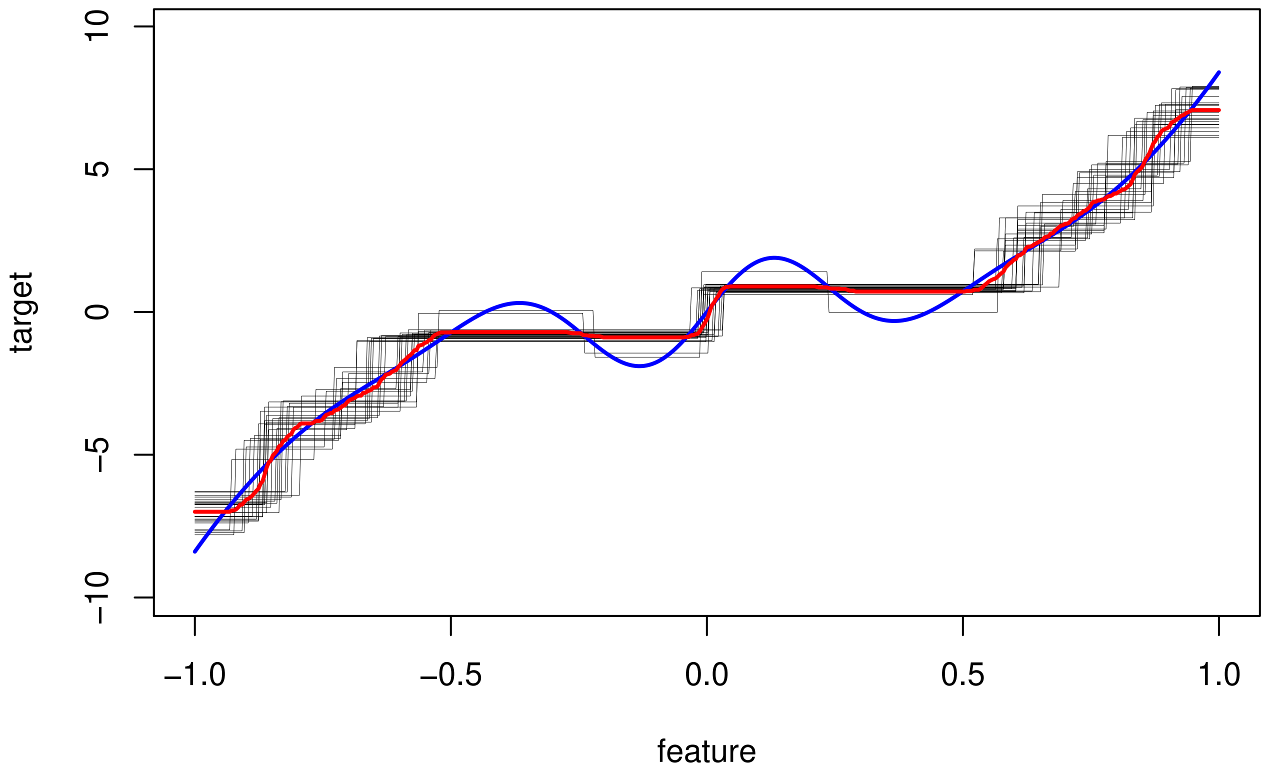
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles. The black line is the average of the y_i that have abscissae within the flat segment.

Fig 27. Default Tree Compared to Truth



The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles. Same as Fig 26 but with blue line added.

Fig 28. Variability in Default Tree



Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees. The black lines are the predictions of the first 25 trees.

Sources of Variability

We have already seen that there is an unavoidable error in making a prediction, which is the deviation, $y_i - f(x_i)$, of the actual occurrence of an observation y_i from the truth $f(x_i)$. Fig 28 shows two additional sources of prediction variability.

The first is the red line that we shall denote by $\bar{f}(x_i)$. It is the average of all possible decision trees that we could have gotten from learning samples the same size as ours. The difference from the truth, $\bar{f}(x_i) - f(x_i)$, is called bias and is a source of prediction error.

The second source of error is represented by the black lines. Each black line is the tree $\hat{f}(x)$ that we get from a particular learning sample. The variation is due to the fact that learning samples differ from one another due to randomness. The difference, $\hat{f}(x) - \bar{f}(x)$, is called estimation error.

Controllable Variation

The two sources of variation under our control are the bias and the estimation error.

As we increase complexity, i.e. allow the tree to have more leaves, the estimation error increases and the bias decreases.

Conversely, as we decrease complexity, the estimation error decreases and the bias increases.

We use the validation sample to try to find a balance between these two sources of error.

But first we need a measure of complexity.

Complexity of a Tree

A tree has many tuning parameters: the minimum number of cases allowed to the right or left of a cut, the minimum number of cases in a leaf, the maximum depth of the tree, the maximum number of leaves, etc.

These are crude structural controls. A finer control is the complexity parameter cp, which is the smallest decrease in $Loss$ in the learning sample for which a split is permitted; $cp = 0.01$ for the default tree.

We shall plot $-\log(cp)$ on the horizontal axis rather than cp because it scales the plot better. We change its sign because tree complexity increases as cp decreases.

Loss for a Prediction Tree

$Loss$ for prediction problems is mean squared error:

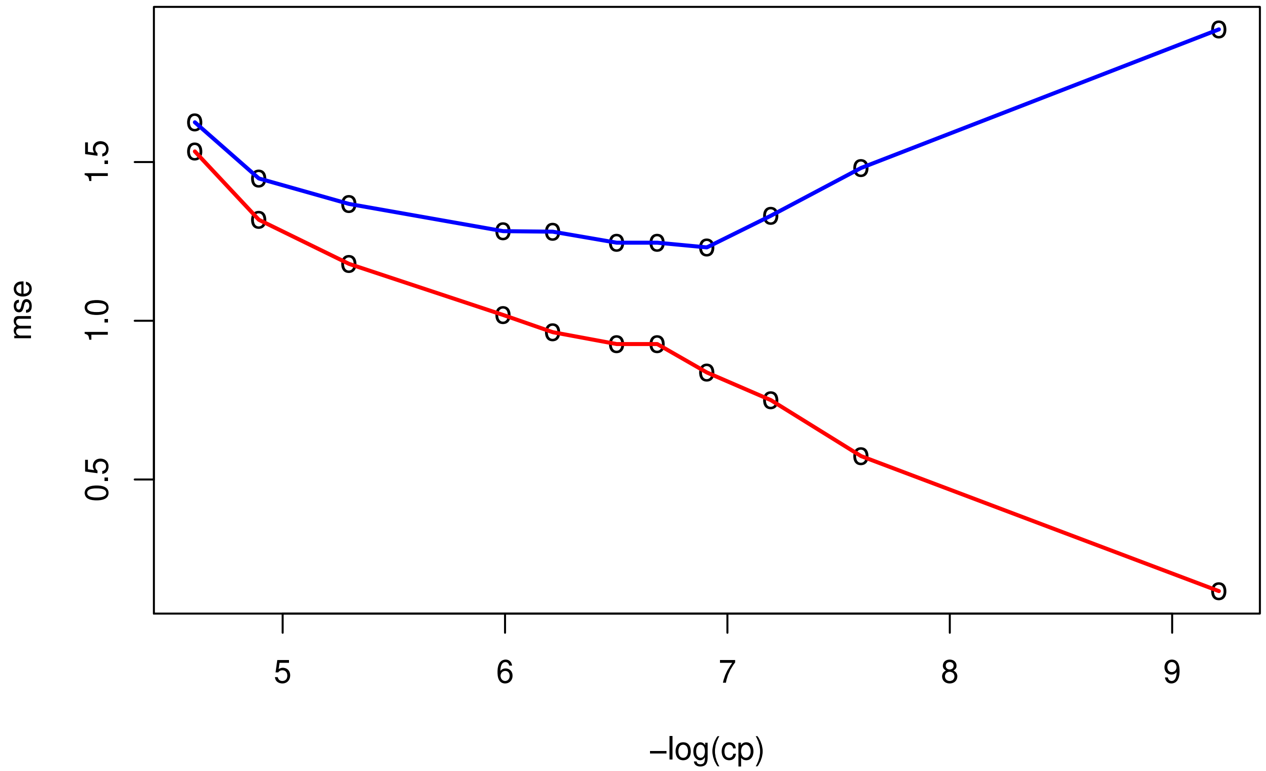
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}(x_i)\right]^2$$

As we shall see in a few slides, in the validation sample MSE correctly measures the sum of the three sources of variability – noise, bias, and estimation error – and indicates the optimal complexity.

As we shall also see, in the learning sample MSE underestimates variability and fails to indicate the optimal complexity.

For our problem, $Loss$ is minimized at complexity $cp = 0.001$ for which $-\log(cp) = 6.9077$ as seen next.

Fig 29. Learning and Validation MSE



The MSE in the learning sample is shown in red. The MSE in the validation sample is shown in blue; its minimum is at $cp = 0.001$. The default tree is the leftmost point on the graph. The MSE in the leaning sample underestimates the bias plus variance and does not indicate optimal complexity.

The Math

Add and subtract like terms to get

$$y_i - \hat{f}(x_i) = y_i - f(x_i) + \bar{f}(x_i) - \hat{f}(x_i) + f(x_i) - \bar{f}(x_i)$$

.

Square to get

$$\begin{aligned}\left[y_i - \hat{f}(x_i)\right]^2 &= e_i^2 + \left[\bar{f}(x_i) - \hat{f}(x_i)\right]^2 + \left[f(x_i) - \bar{f}(x_i)\right]^2 \\ &+ \text{ cross product terms.}\end{aligned}$$

recalling that $y_i = f(x_i) + e_i$.

The Cross Product Terms

The two terms $\bar{f}(x_i)$ and $f(x_i)$ have no sampling variation in them because the first is a population mean and the second is a nonrandom function.

The product of a random variable that has mean zero and a quantity that is not random has mean zero. Therefore, the following two cross product terms have mean zero in both the training sample and the validation sample:

$$e_i \left[\bar{f}(x_i) - f(x_i) \right] \qquad \left[\hat{f}(x_i) - \bar{f}(x_i) \right] \left[\bar{f}(x_i) - f(x_i) \right]$$

The third cross product term is

$$e_i \left[\bar{f}(x_i) - \hat{f}(x_i) \right]$$

The Third Cross Product Term

When considering

$$e_i \left[\bar{f}(x_i) - \hat{f}(x_i)\right]$$

it matters whether we are discussing the training sample or the validation sample.

First let's consider what happens in the training sample. Look at Fig 26 and recall that $y_i = f(x_i) + e_i$ and that $\hat{f}(x_i)$ is a mean that involves y_i. When e_i increases y_i increases and therefore $\hat{f}(x_i)$ increases. Consequently $\left[\bar{f}(x_i) - \hat{f}^{(}x_i)\right]$ decreases when e_i increases and the correlation between them must be negative.

Therefore the mean of the cross product term, which is the numerator of the correlation, is negative in the training sample.

The Third Cross Product Term

Now consider what happens to the third cross product term

$$e_i \left[\bar{f}(x_i) - \widehat{f}(x_i) \right]$$

in the validation sample.

In the validation sample e_i and $\widehat{f}(x_i)$ have nothing to do with each other because cases in the validation sample were not used to compute $\widehat{f}(x_i)$. Therefore e_i and $\widehat{f}(x_i)$ are independent random variables. The mean of the product of independent random variables is the product of their means. Recall that e_i has mean zero.

Therefore the mean of the third cross product term is zero in the validation sample.

The Relevant Quantities

The within sample variance due to (1) uncontrollable noise, (2) controllable estimation error, and (3) controllable bias are

$$s_e^2 = \frac{1}{n} \sum_{t=1}^{n} e_i^2$$

$$s_f^2 = \frac{1}{n} \sum_{t=1}^{n} \left[\bar{f}(x_i) - \hat{f}(x_i) \right]^2$$

$$s_b^2 = \frac{1}{n} \sum_{t=1}^{n} \left[f(x_i) - \bar{f}(x_i) \right]^2$$

And (4), the covariance between noise and estimation error is

$$s_{ef} = \frac{1}{n} \sum_{t=1}^{n} e_i \left[\bar{f}(x_i) - \hat{f}(x_i) \right]$$

The corresponding population quantities are σ_e^2, σ_f^2, σ_b^2, and σ_{ef}.

The Main Facts

In the validation sample, MSE is estimating the quantity

$$\sigma_e^2 + \sigma_f^2 + \sigma_b^2$$

As complexity increases, σ_f^2 increases and σ_b^2 decreases. The desired complexity is where MSE is at a minimum.

In the learning sample, MSE is estimating the quantity

$$\sigma_e^2 + \sigma_f^2 + \sigma_b^2 + 2\sigma_{ef}$$

where σ_{ef} is negative. Moreover, $2\sigma_{ef}$ overwhelms σ_f^2 so MSE decreases as complexity increases.

More Features

Adding additional features also increases complexity. For example, in the Loan Case, the regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 (x_1)^2 + b_4 (x_2)^2 + b_5 (x_1)(x_2)$$

has higher complexity than the regression

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

The principle is the same: Use a validation sample to choose the correct degree of complexity.

More Plots

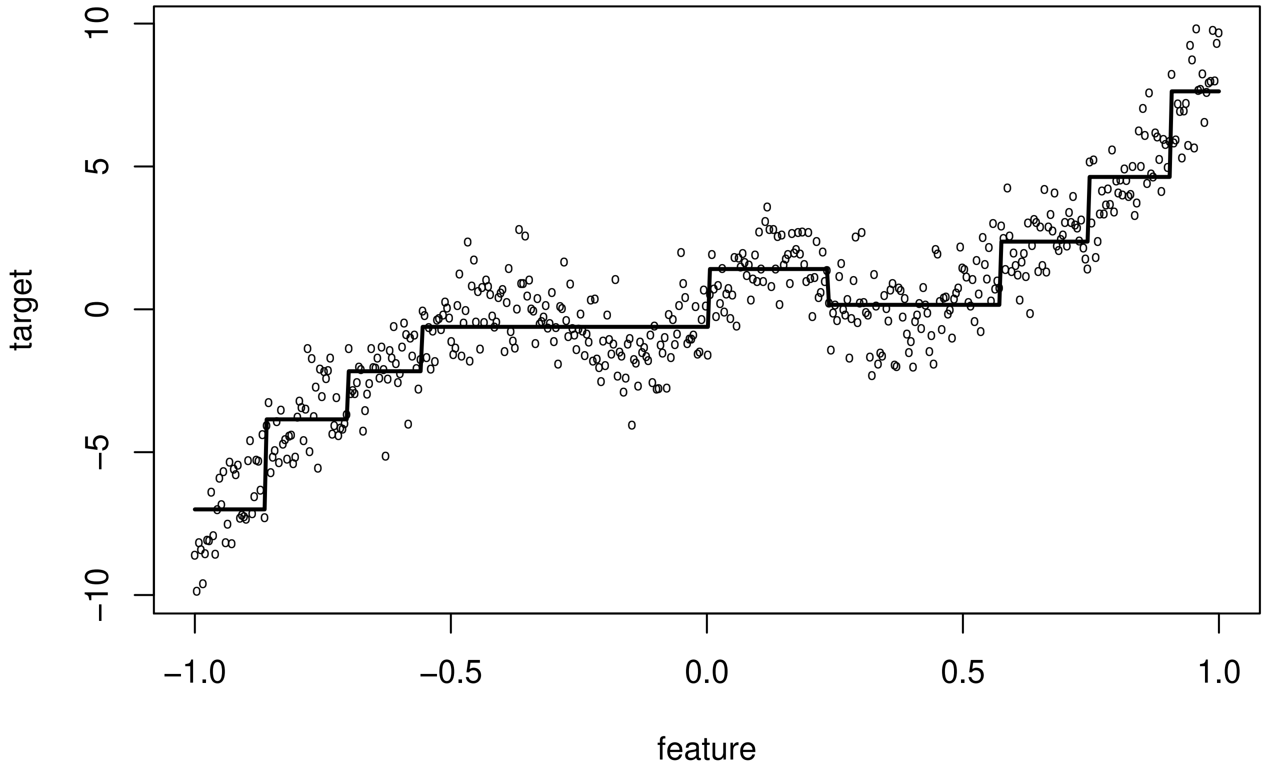
The next slides show what happens as complexity increases.

Pay special attention to Figs 32, 35, 39, 42, 45, and 49.

In these plots, the black fuzz represents σ_f^2 and the difference between the blue and red lines represents σ_b^2.

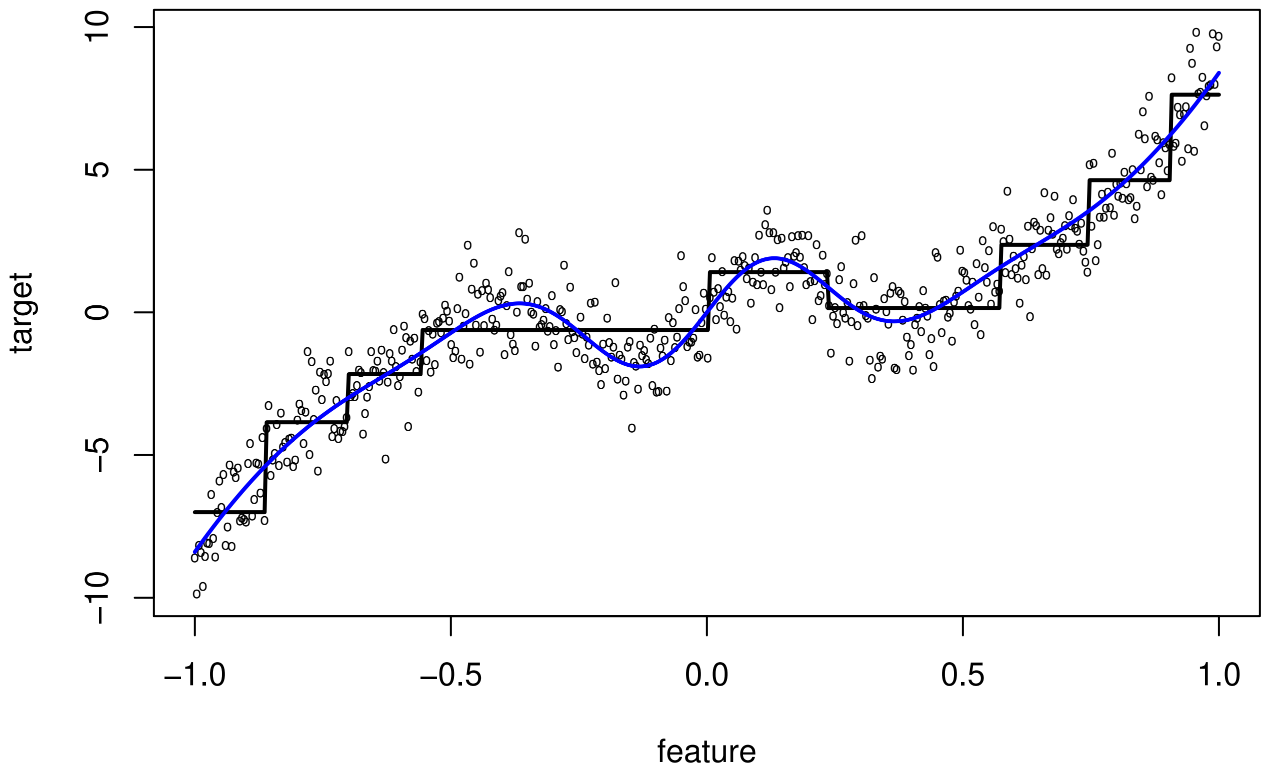
Notice that the red and blue lines eventually coincide to within plotting accuracy so that σ_b^2 becomes nearly zero.

Fig 30. Tree Compared to Data, $cp = 0.0075$



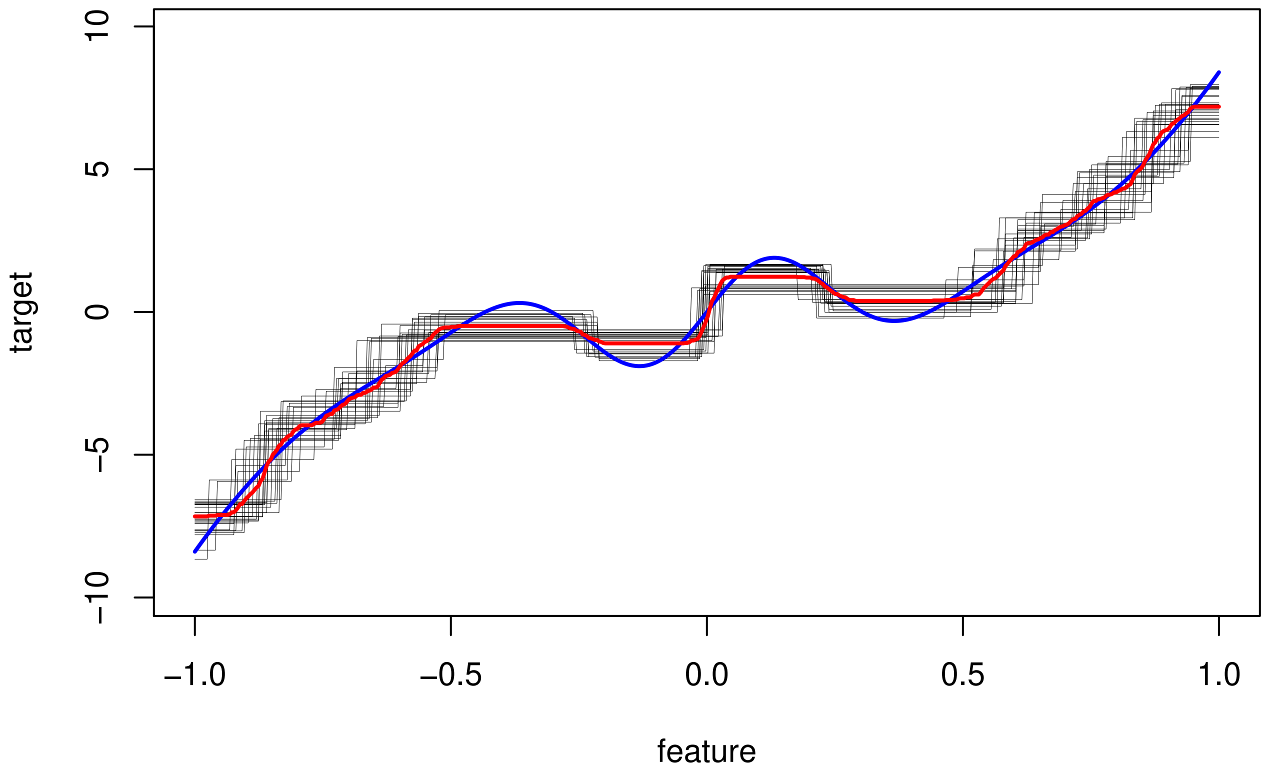
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 31. Tree Compared to Truth, $cp = 0.0075$



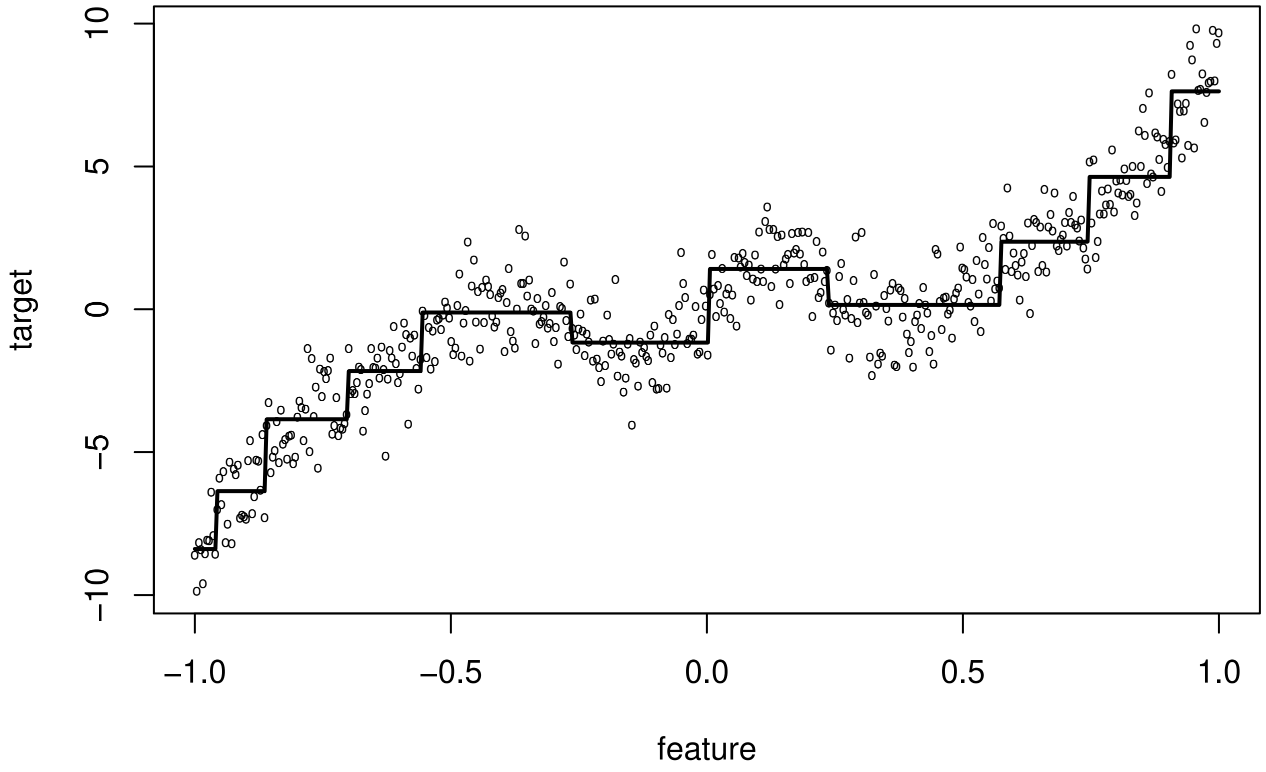
The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 32. Variability in Tree, $cp = 0.0075$



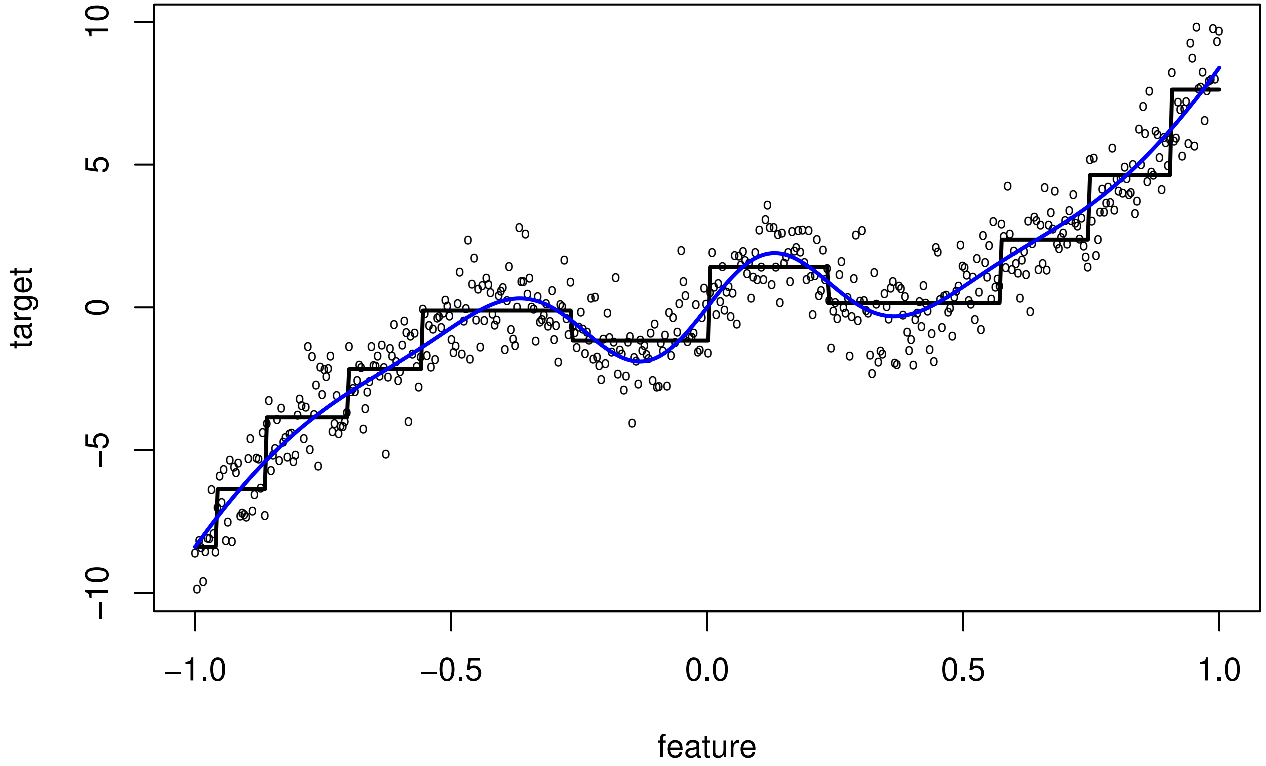
Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees.

Fig 33. Tree Compared to Data, $cp = 0.005$



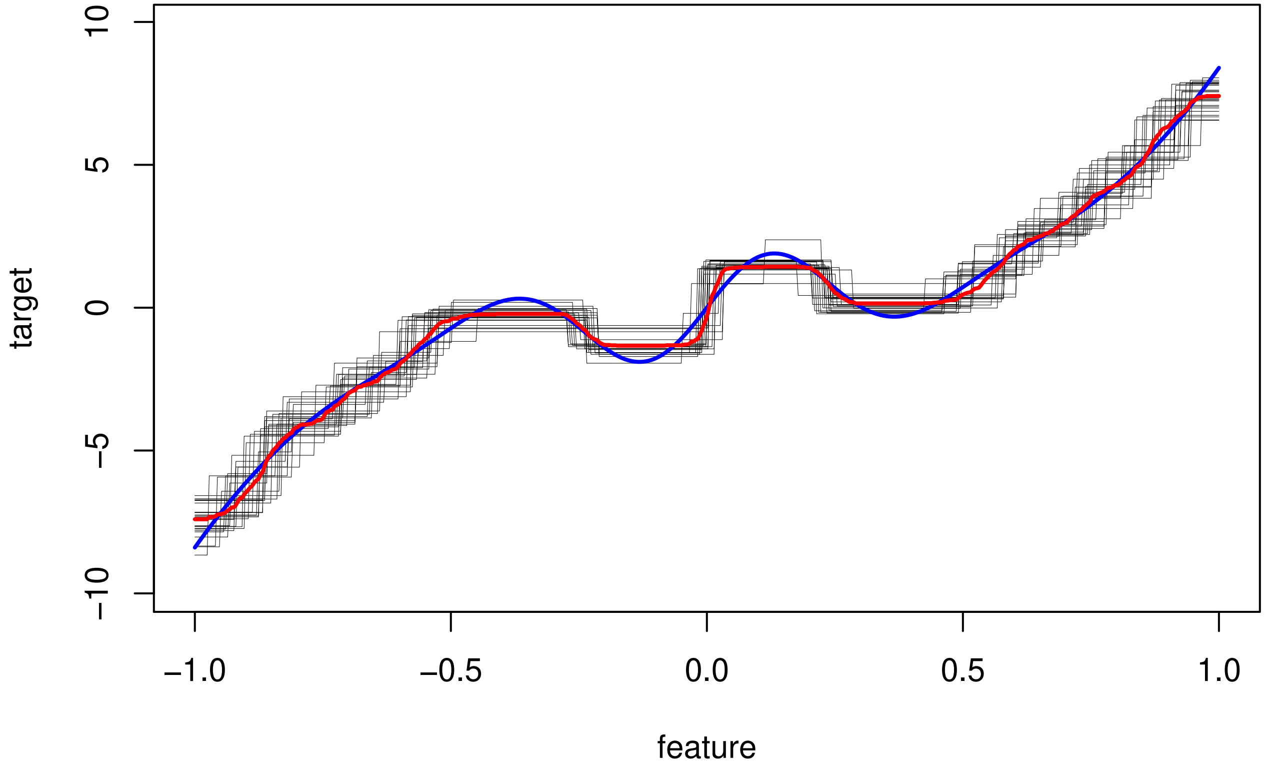
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 34. Tree Compared to Truth, $cp = 0.005$



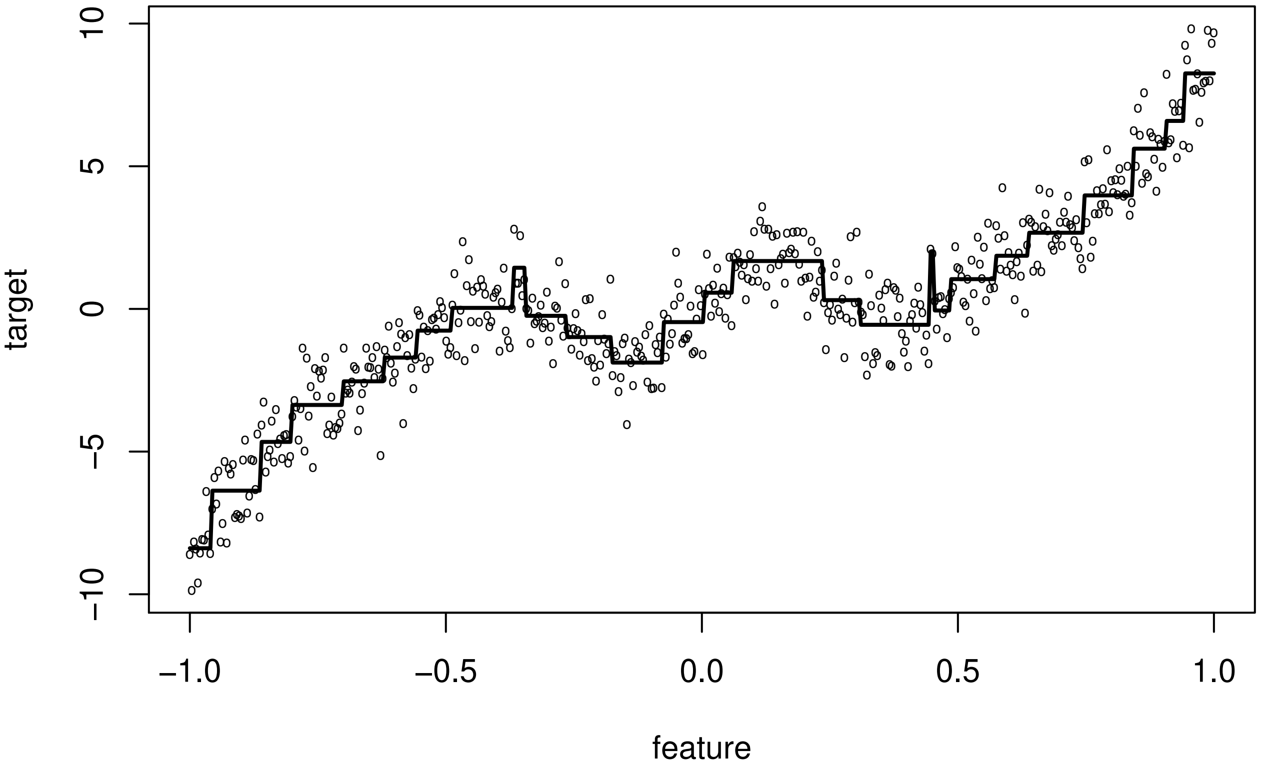
The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 35. Variability in Tree, $cp = 0.005$



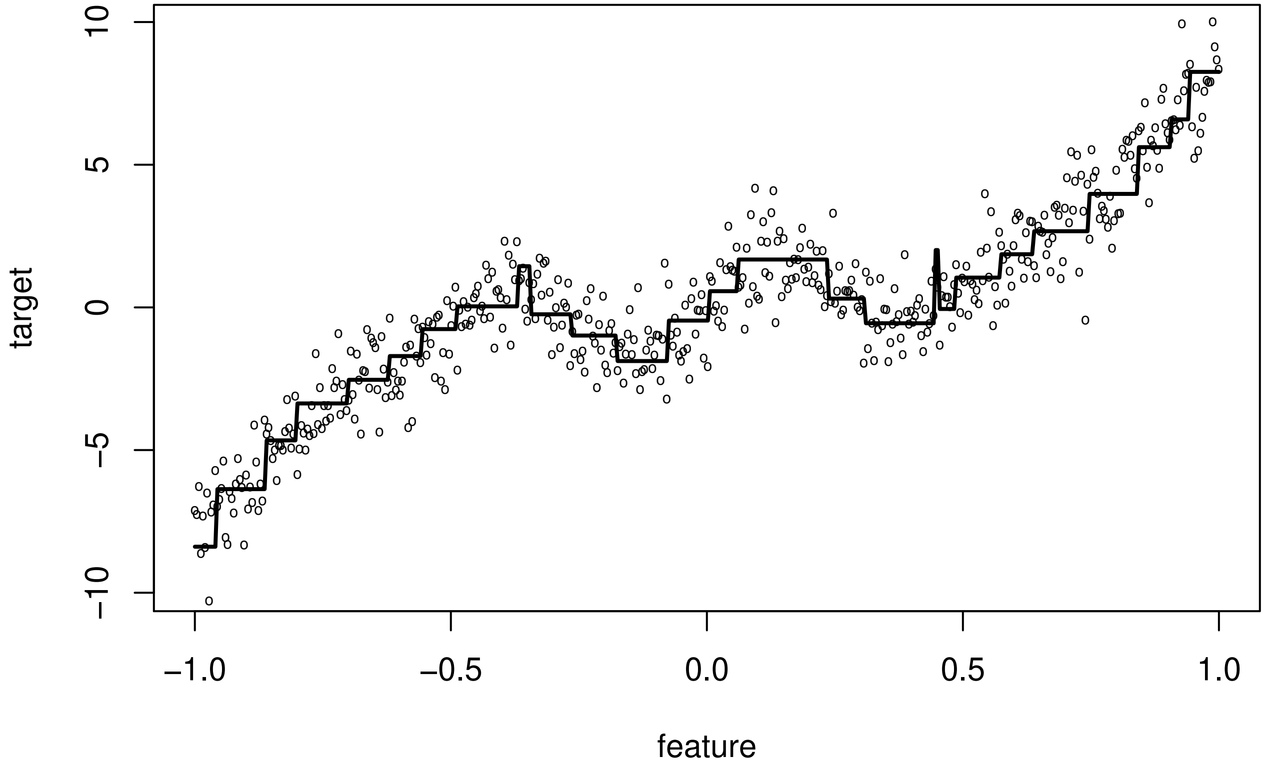
Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees.

Fig 36. Tree Compared to Data, $cp = 0.001$



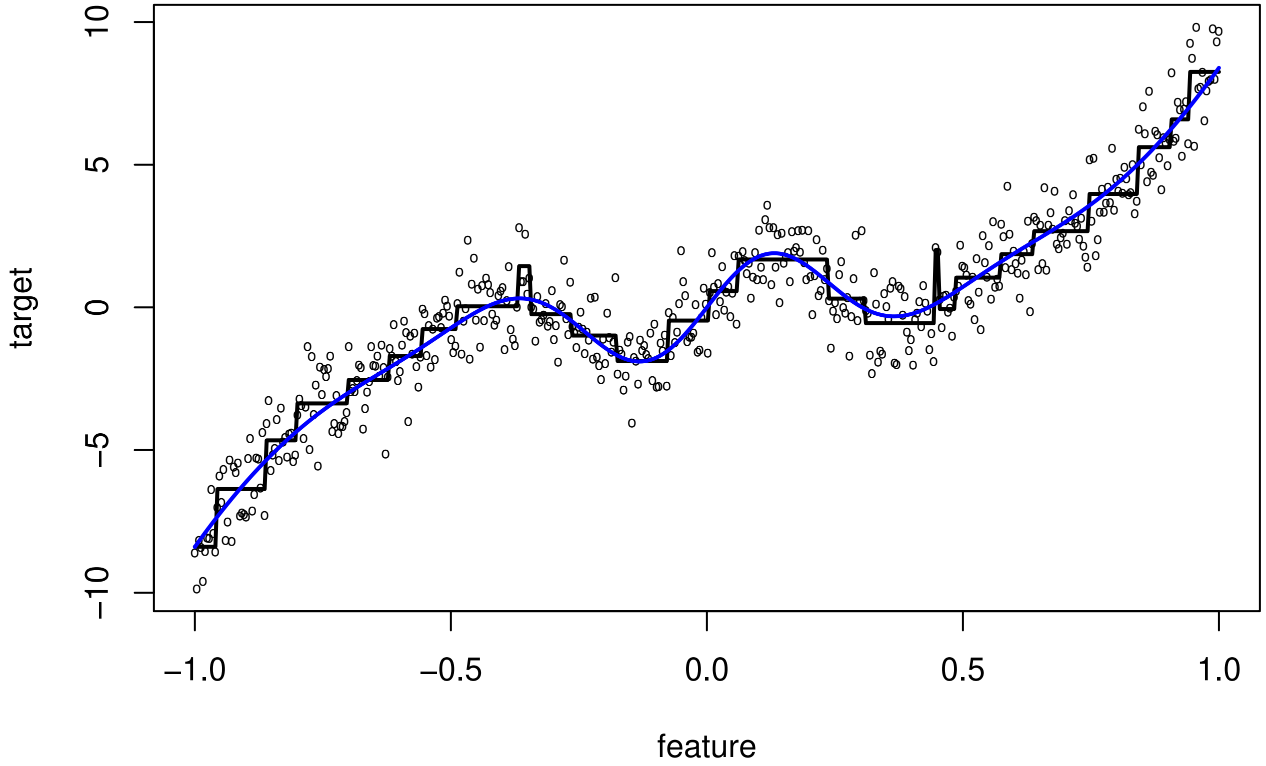
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles. This is the best tree.

Fig 37. Tree Compared to Validation Data, $cp = 0.001$



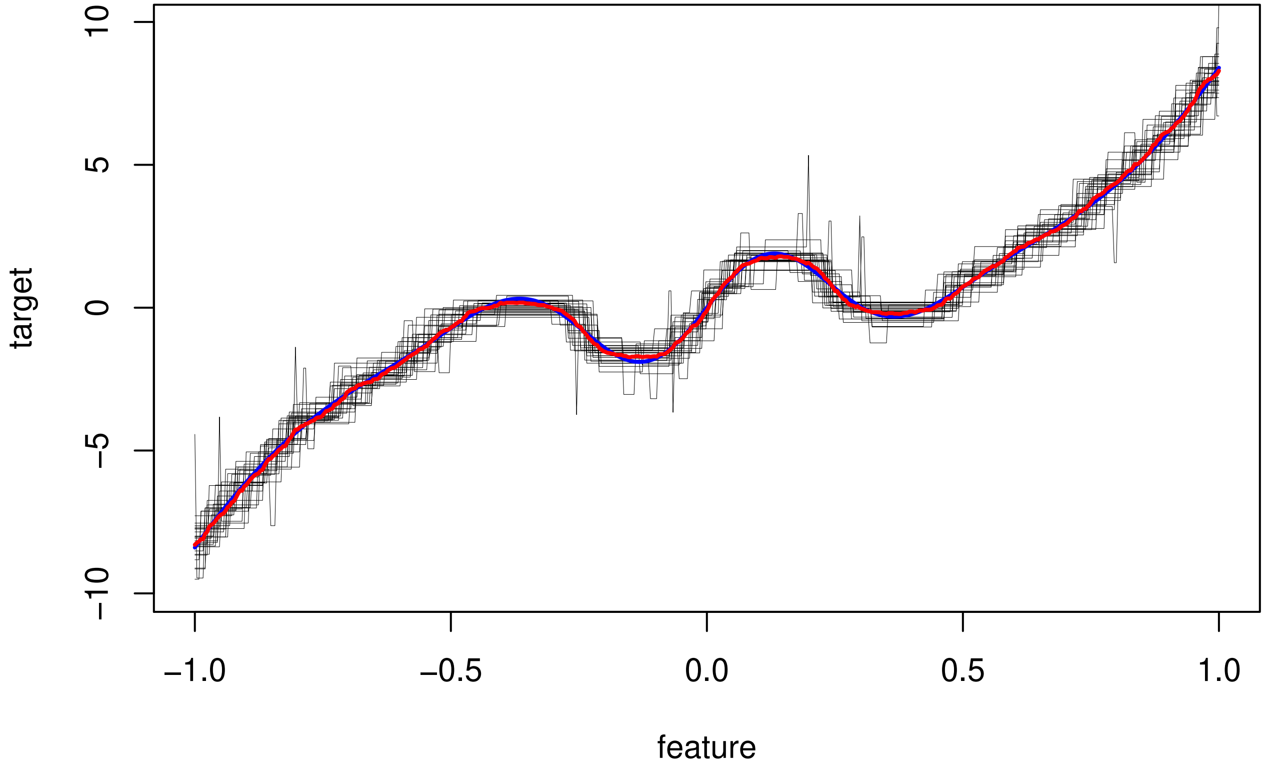
The prediction of the fitted tree is shown as the black line; the validation sample is shown as black circles. This tree generalizes well.

Fig 38. Tree Compared to Truth, $cp = 0.001$



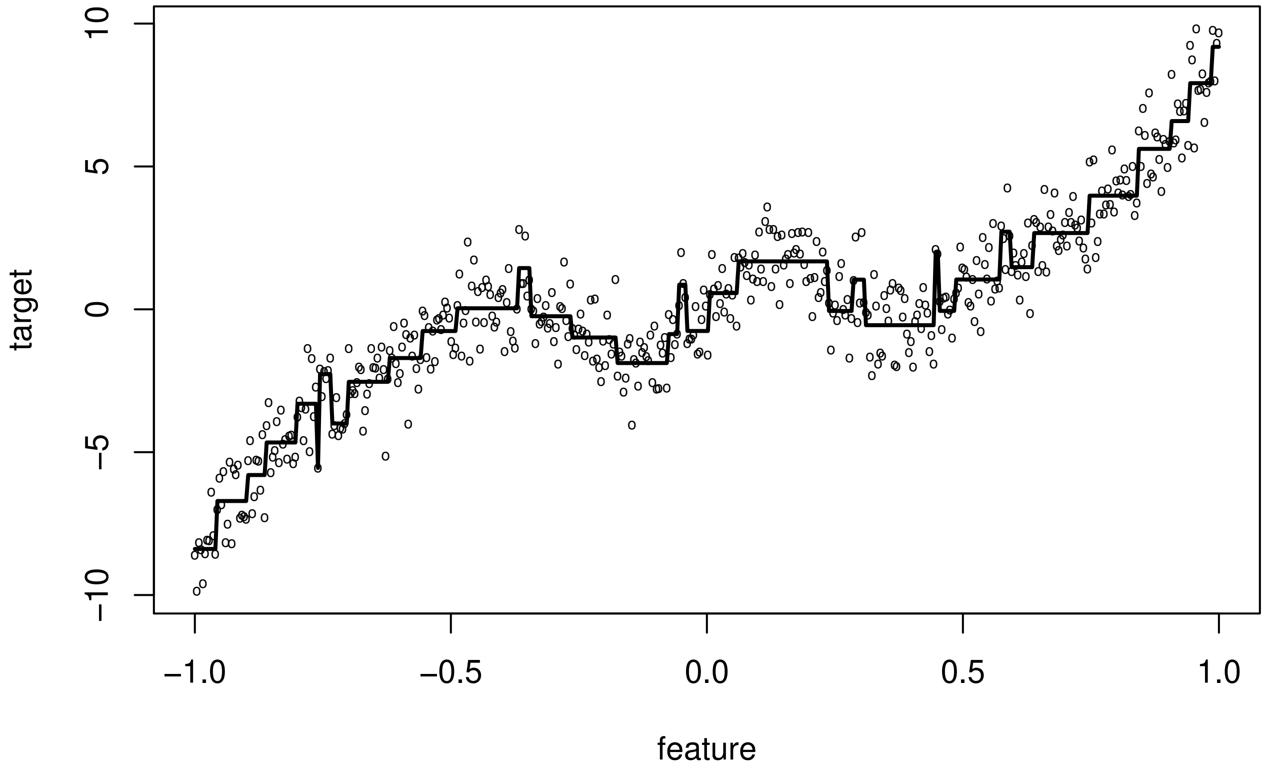
The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles. This is the best tree.

Fig 39. Variability in Tree, $cp = 0.001$



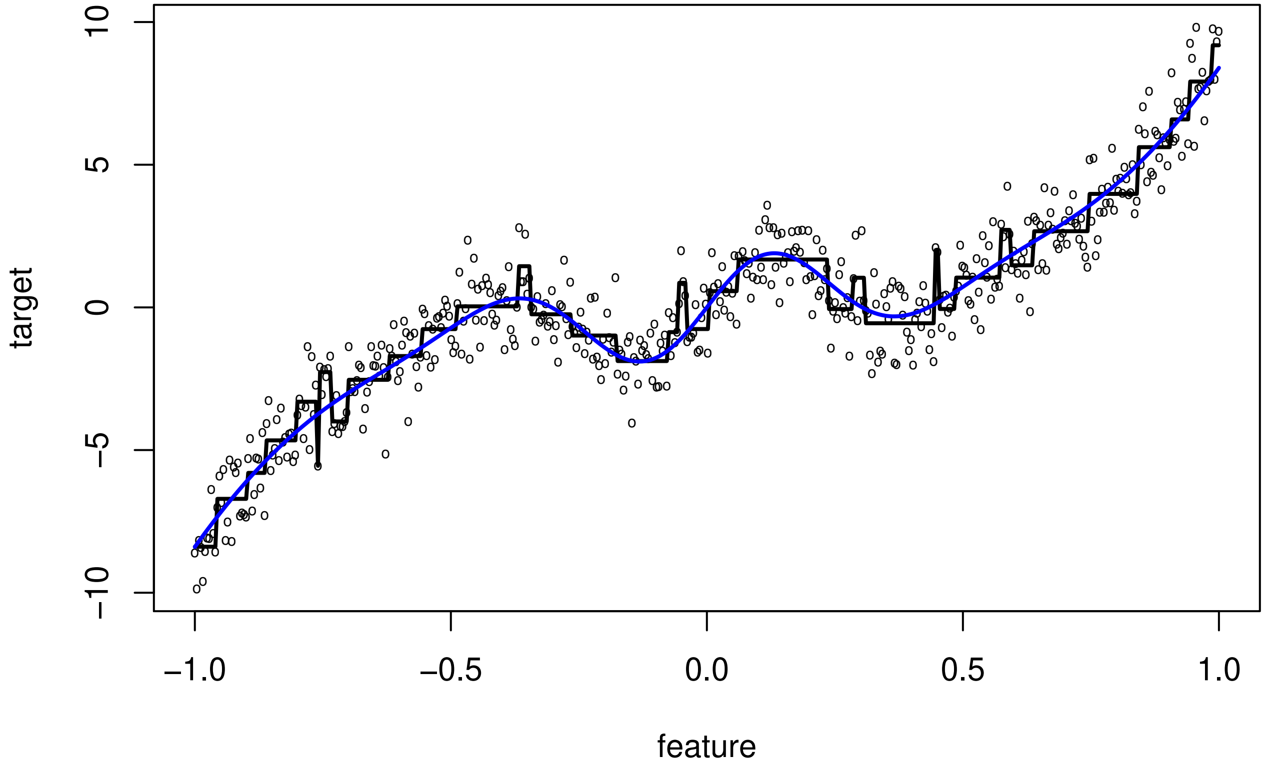
Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees. This is the best tree.

Fig 40. Tree Compared to Data, $cp = 0.00075$



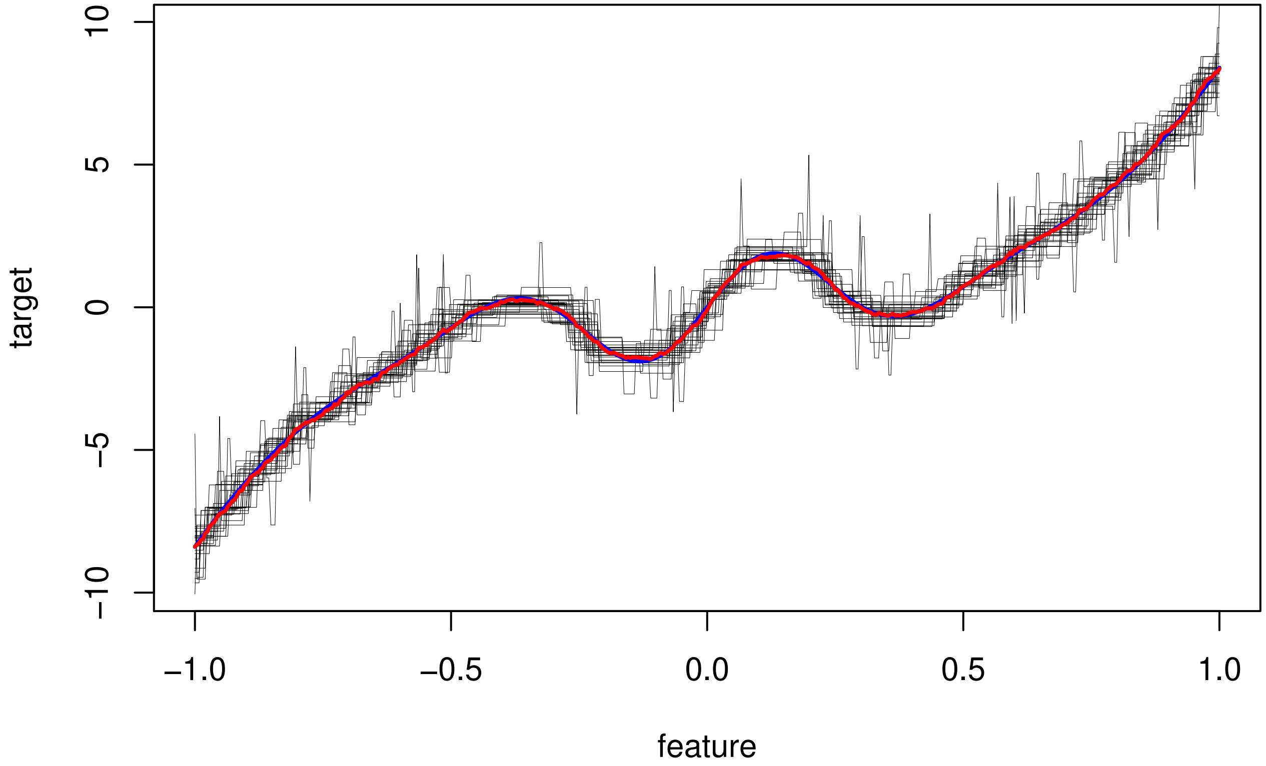
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 41. Tree Compared to Truth, $cp = 0.00075$



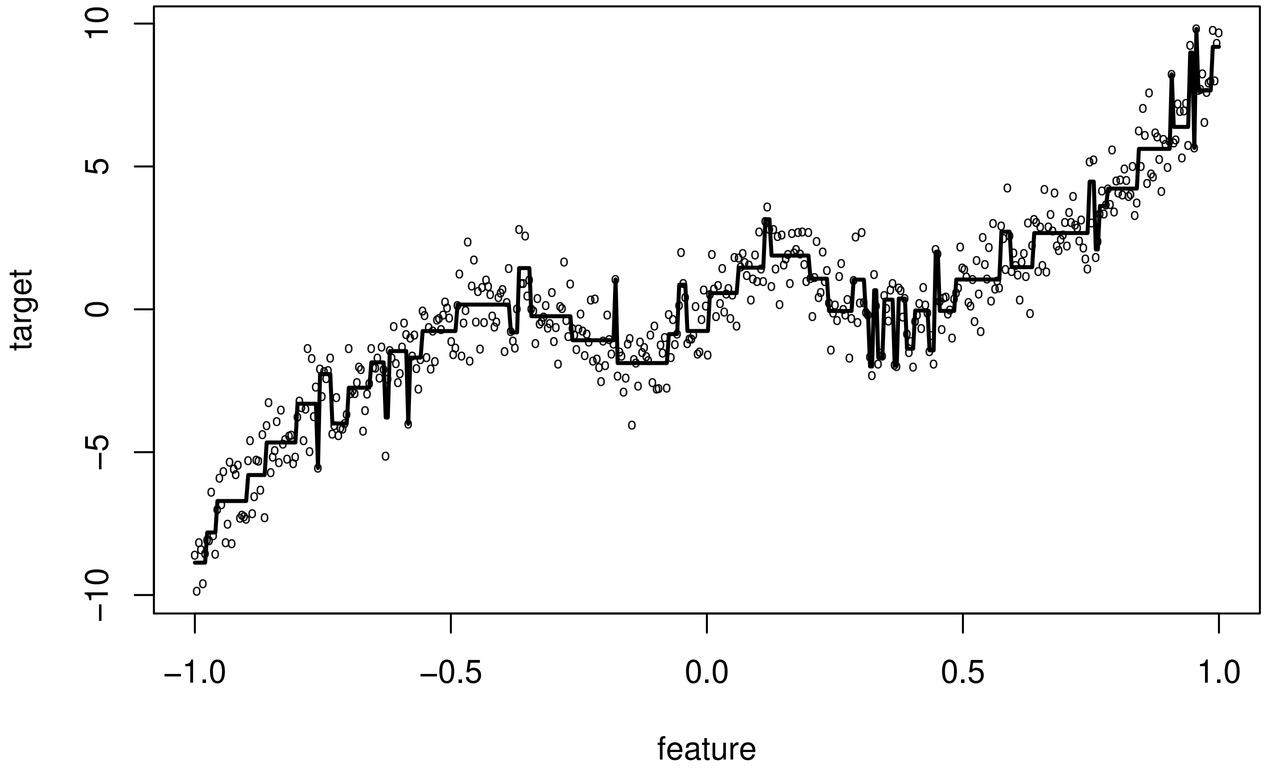
The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 42. Variability in Tree, $cp = 0.00075$



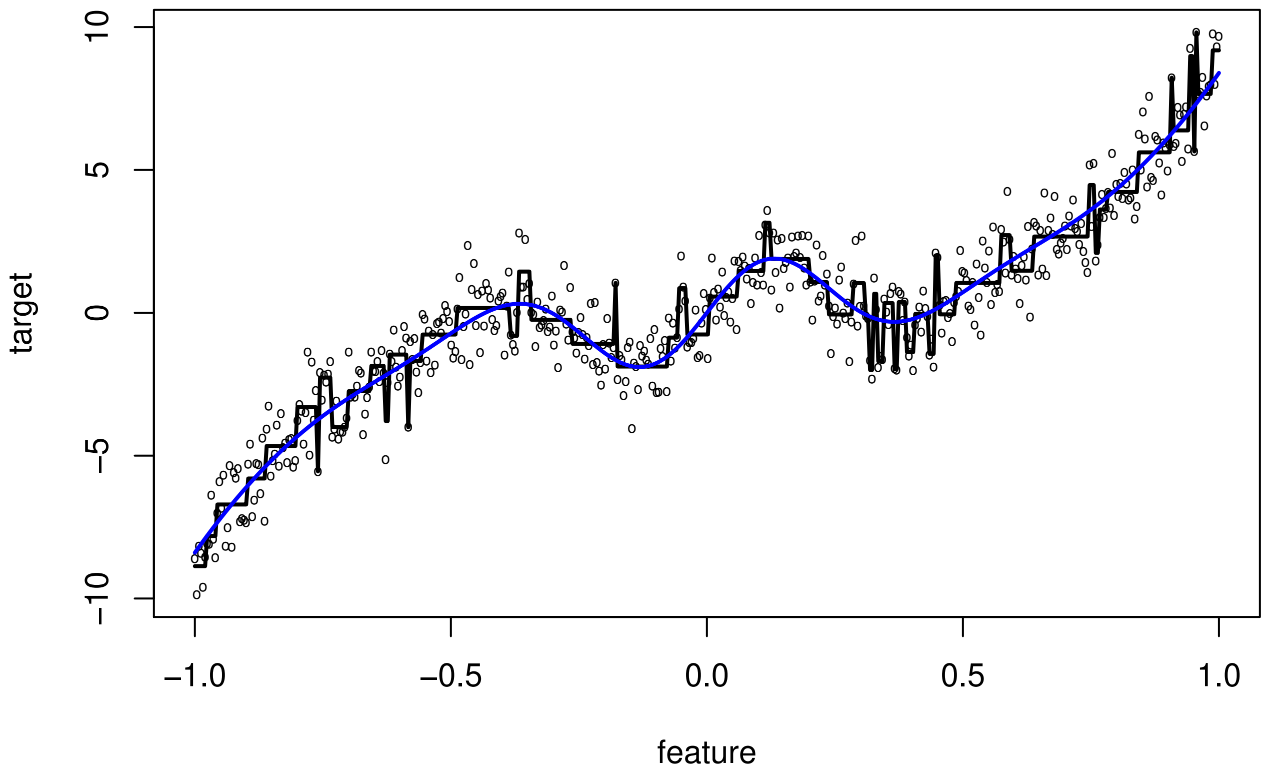
Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees.

Fig 43. Tree Compared to Data, $cp = 0.0005$



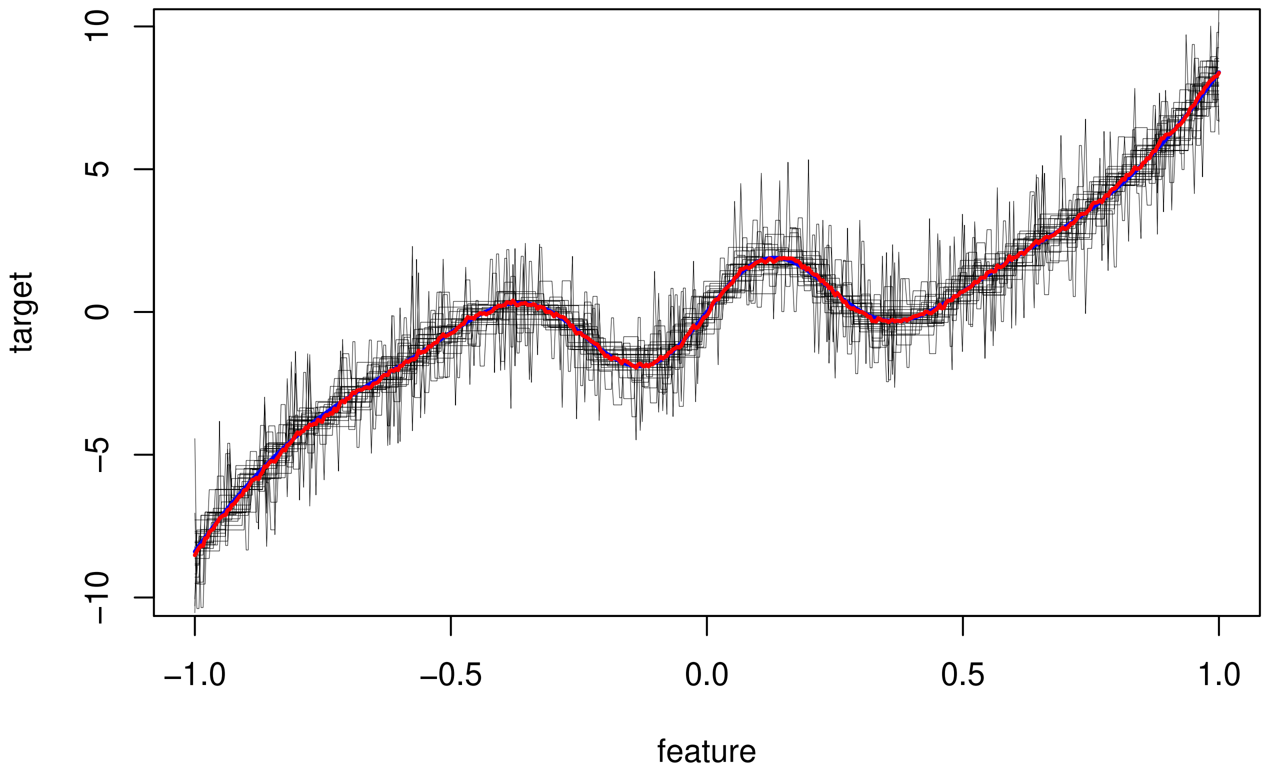
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 44. Tree Compared to Truth, $cp = 0.0005$



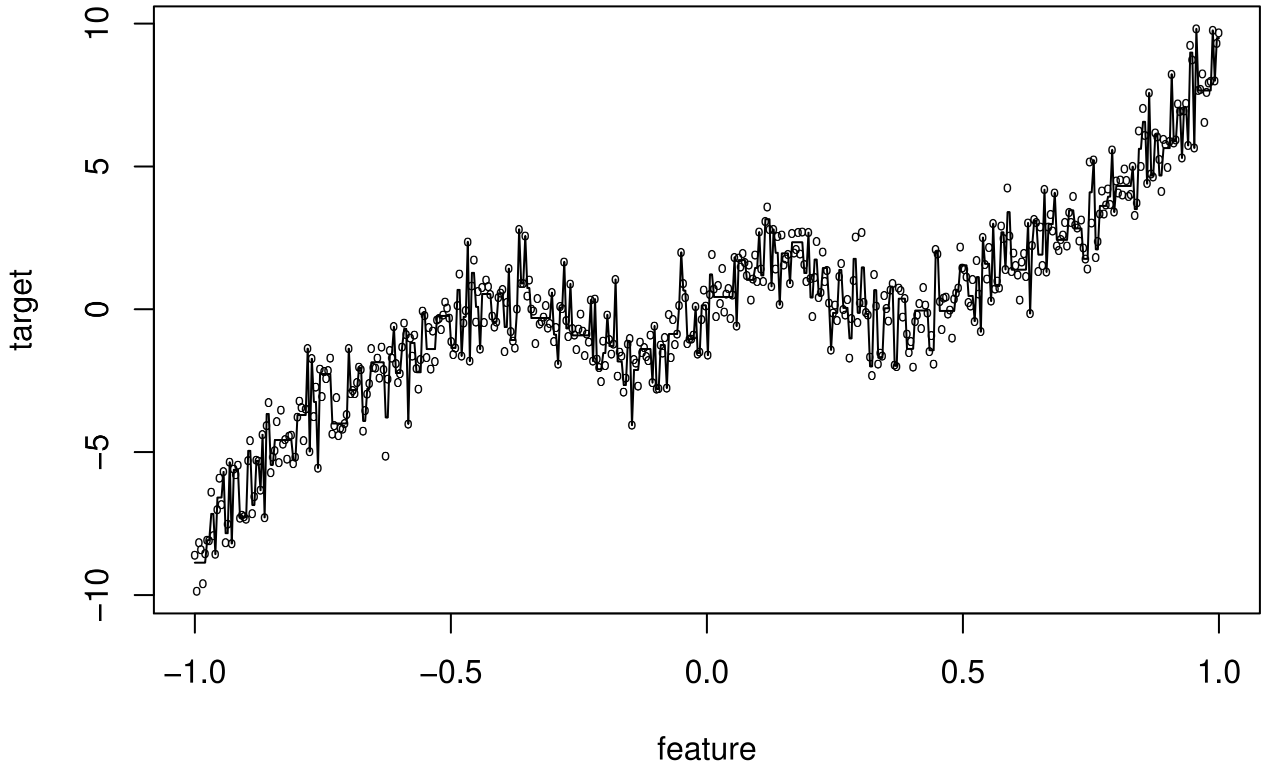
The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 45. Variability in Tree, $cp = 0.0005$



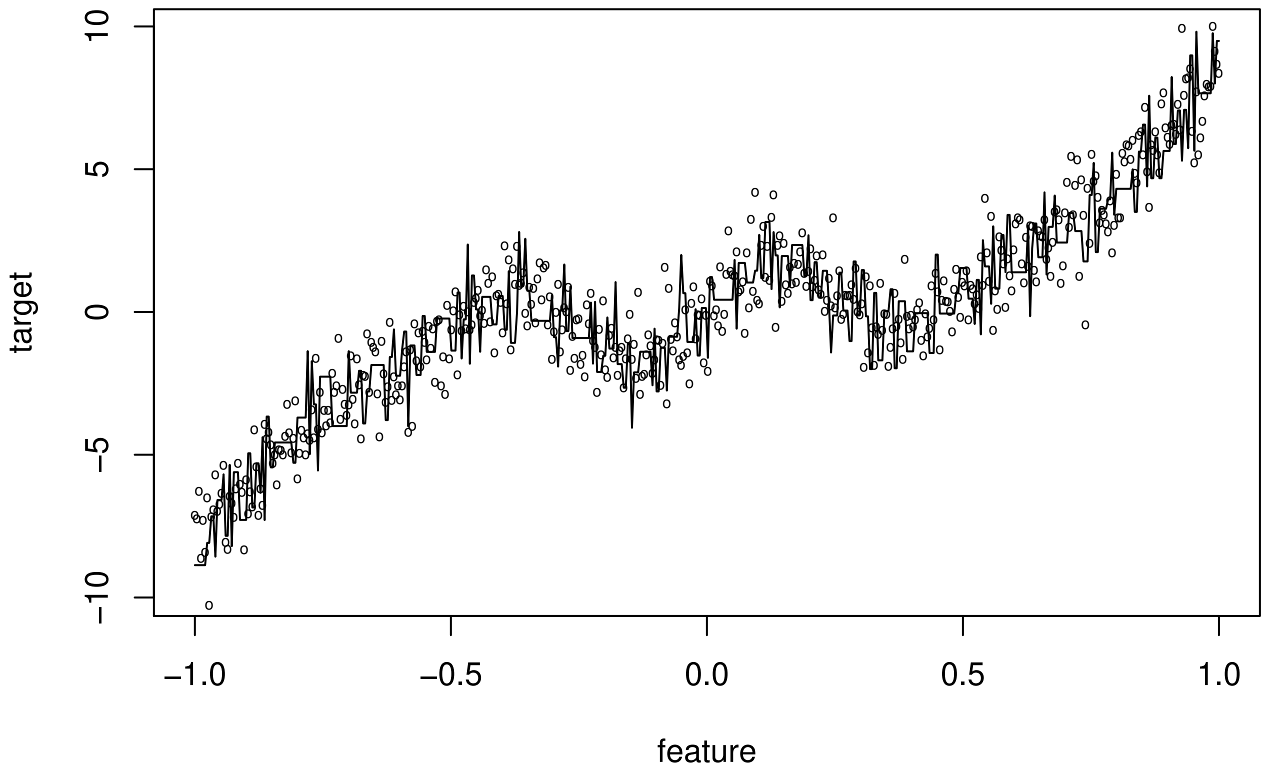
Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees.

Fig 46. Tree Compared to Data, $cp = 0.0001$



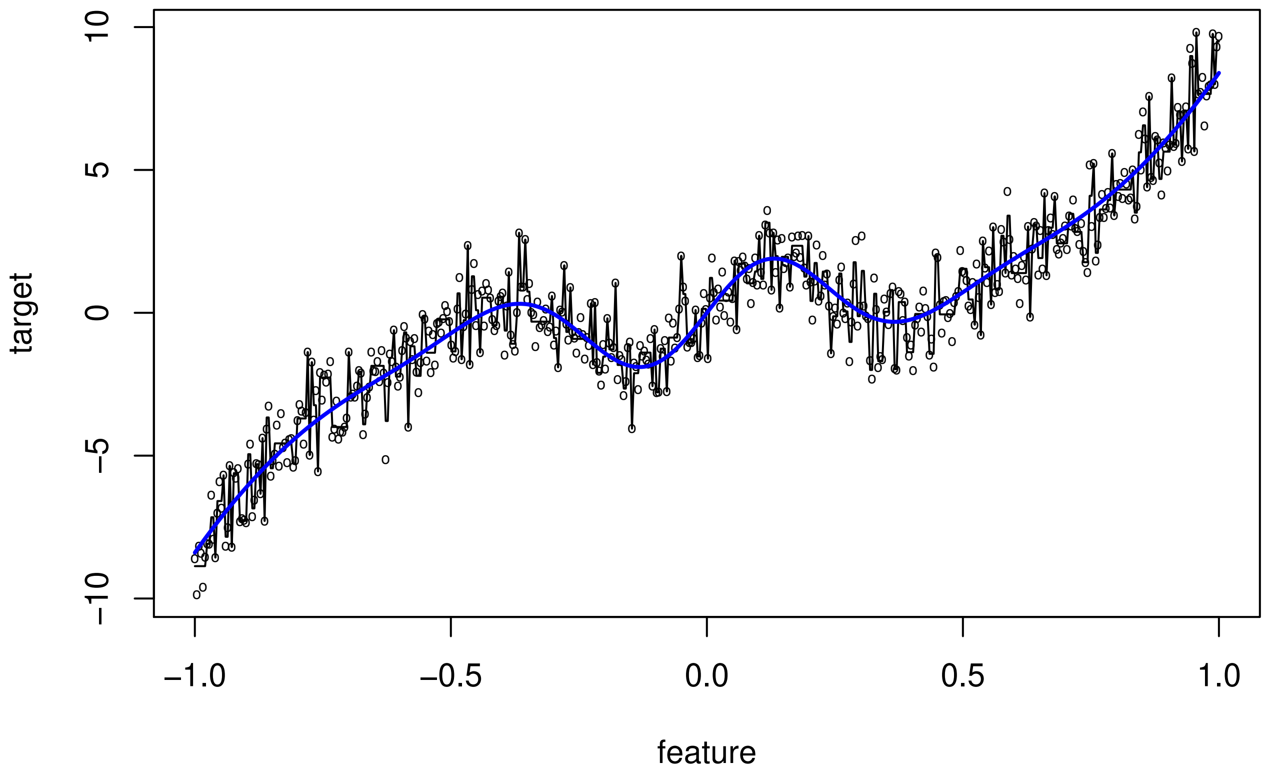
The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 47. Tree Compared to Validation Data, $cp = 0.0001$



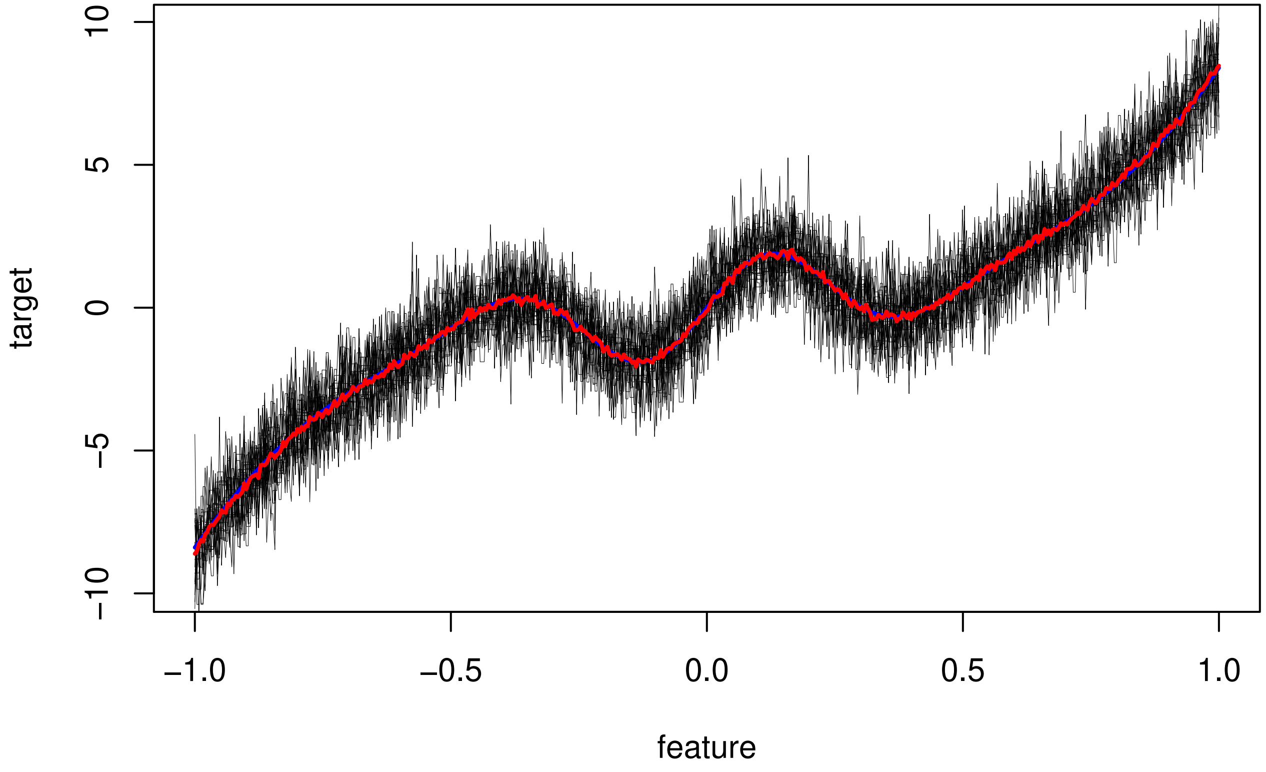
The prediction of the fitted tree is shown as the black line; the validation sample is shown as black circles. One sees that an overfitted tree generalizes poorly.

Fig 48. Tree Compared to Truth, $cp = 0.0001$



The model is $y = f + e$ where f is shown in blue. The prediction of the fitted tree is shown as the black line; the learning sample is shown as black circles.

Fig 49. Variability in Tree, $cp = 0.0001$



Trees were fitted to 100 samples from $y = f + e$ where f is shown in blue. The red line is the average of the 100 trees and the black lines are the predictions of the first 25 trees.

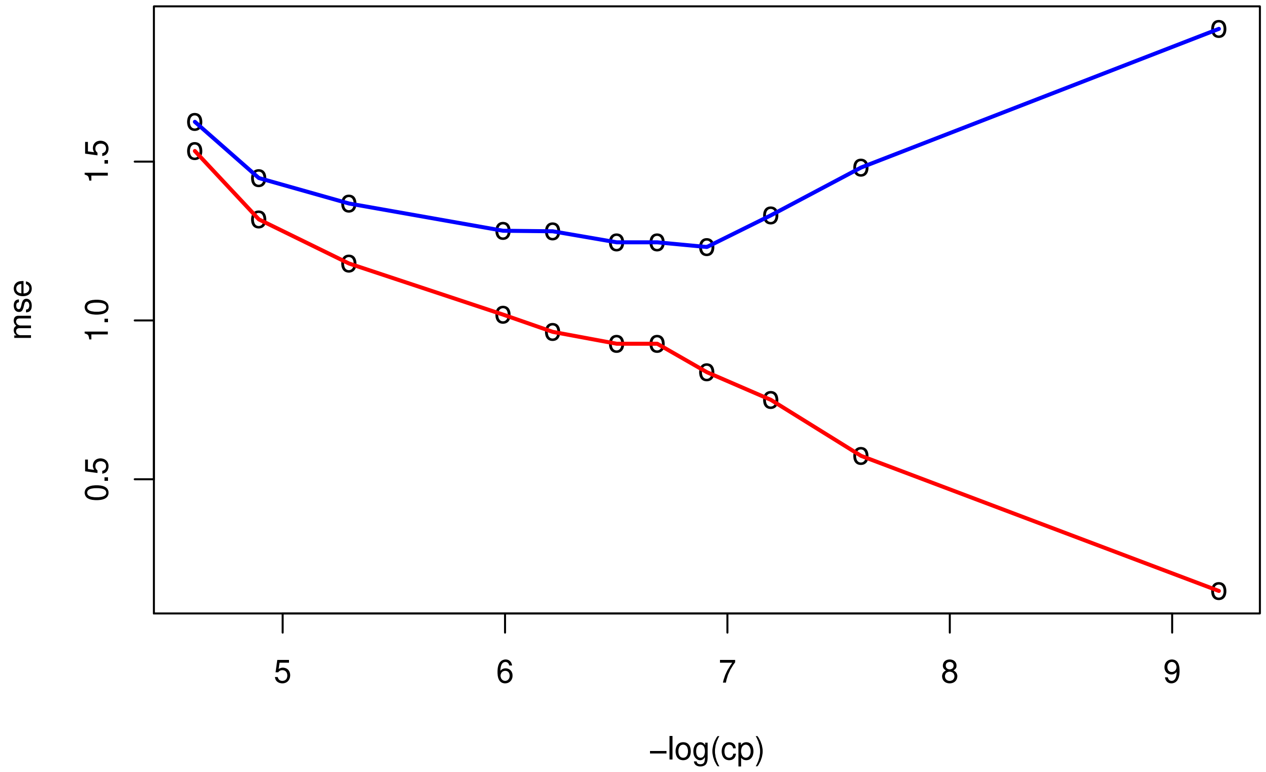
The Main Points

1. The correct complexity for any tool is at the point at which the MSE in the validation sample is minimized. Generalization is best at that point.

2. The MSE in the learning sample is an under estimate of the relevant MSE because σ_{ef} is negative. For most tools, the learning sample MSE curve will not have a minimum.

These two points can be verified visually by the following plot, which repeats Fig 28.

Fig 50. Learning and Validation MSE



The MSE in the learning sample is shown in red. The MSE in the validation sample is shown in blue; its minimum is at $cp = 0.001$. The MSE in the leaning sample underestimates the bias plus variance. The tree with $cp = 0.0001$ is the rightmost point on the graph.

Blank page

Blank page

Blank page