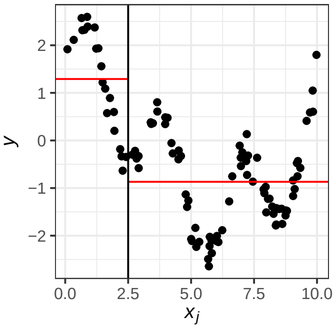
Introduction to Machine Learning

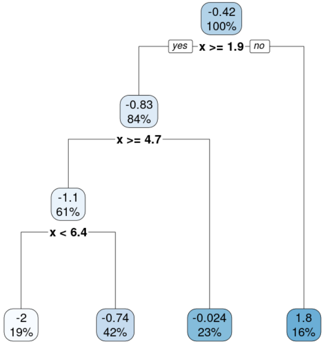
CART
Splitting Criteria for Regression

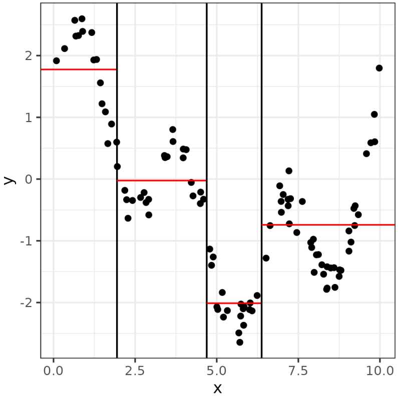


Learning goals

- Understand how to define split criteria via ERM
- Understand how to find splits in regression with L_2 loss

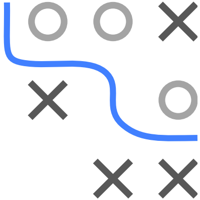
SPLITTING CRITERIA





How to find good splitting rules? $\Longrightarrow$ **Empirical Risk Minimization**

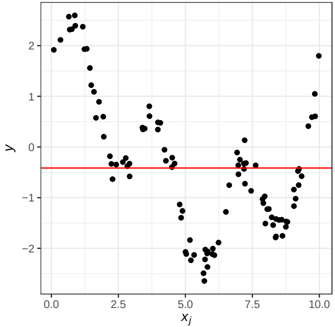
OPTIMAL CONSTANTS IN LEAVES



Idea: A split is good if each child's point predictor reflects its data well.

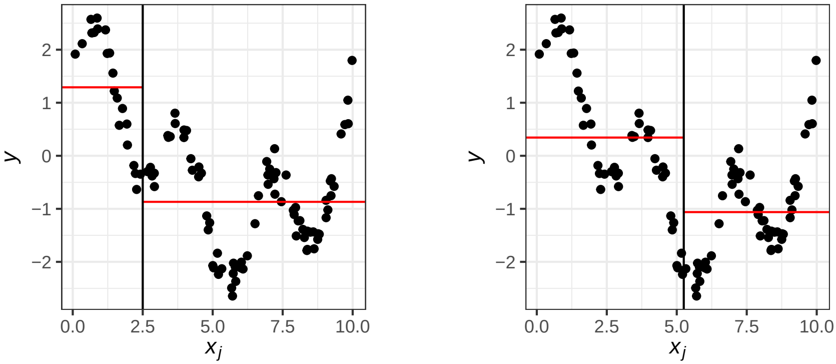
For each child $\mathcal{N}$, predict with optimal constant, e.g., the mean $c_\mathcal{N} = \frac{1}{|\mathcal{N}|} \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} y$ for the L_2 loss, i.e., $\mathcal{R}(\mathcal{N}) = \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} (y - c_\mathcal{N})^2$.

Root node:

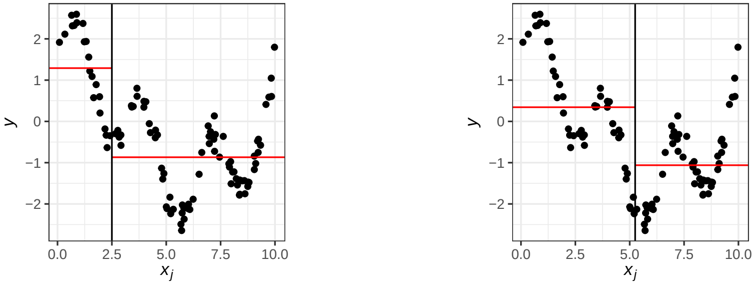


OPTIMAL CONSTANTS IN LEAVES

Which of these two splits is better?



RISK OF A SPLIT



$\mathcal{R}(\mathcal{N}_1) = 23.4$, $\mathcal{R}(\mathcal{N}_2) = 72.4$ $\quad$ $\mathcal{R}(\mathcal{N}_1) = 78.1$, $\mathcal{R}(\mathcal{N}_2) = 46.1$

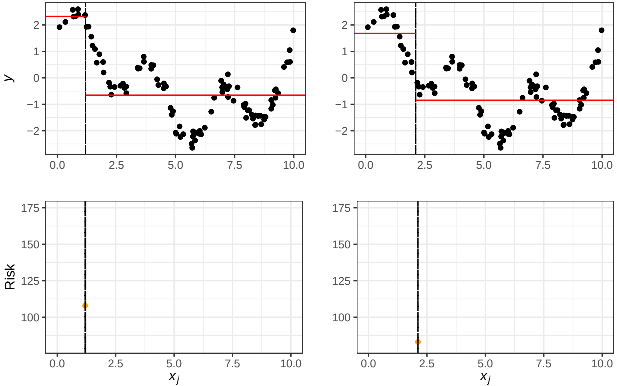
The total risk is the sum of the individual losses:

$23.4 + 72.4 = 95.8$ $\quad$ $78.0 + 46.1 = 124.1$

Based on the SSE, we prefer the first split.

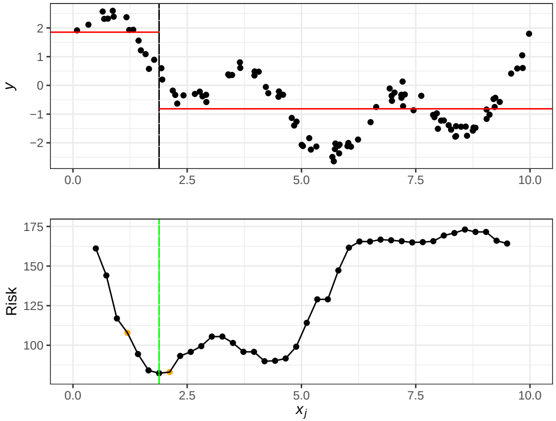
SEARCHING THE BEST SPLIT

Let’s find the best split for this feature by tabulating results.



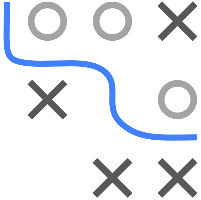
SEARCHING THE BEST SPLIT

Let's iterate – quantile-wise or over all points.



We have reduced the problem to a simple loop.

FORMALIZATION



- $\mathcal{N} \subseteq \mathcal{D}$ is the data contained in this node
- Let $c_\mathcal{N}$ be the predicted constant for $\mathcal{N}$
- The risk $\mathcal{R}(\mathcal{N})$ for a node is:

$$\mathcal{R}(\mathcal{N}) = \sum_{(\mathbf{x},y)\in\mathcal{N}} L(y, c_\mathcal{N})$$

- The optimal constant is $c_\mathcal{N} = \arg\min_c \sum_{(\mathbf{x},y)\in\mathcal{N}} L(y, c)$
- We often know what that is from theoretical considerations – or we can perform a simple univariate optimization