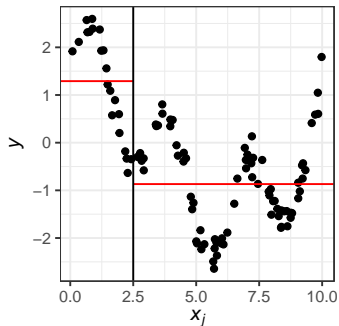


Introduction to Machine Learning

CART

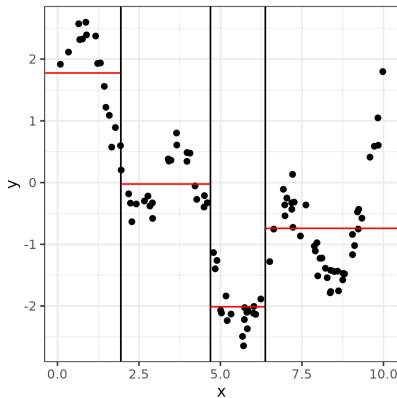
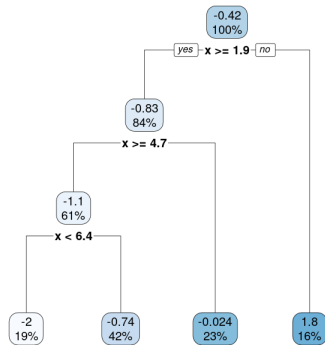
Splitting Criteria for Regression



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? \Rightarrow **Empirical Risk Minimization**

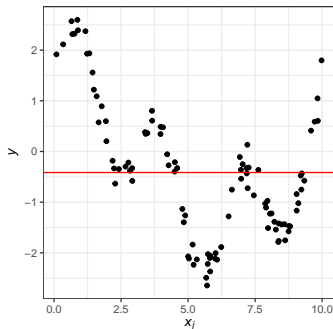
OPTIMAL CONSTANTS IN LEAVES

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

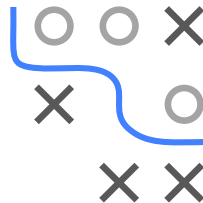
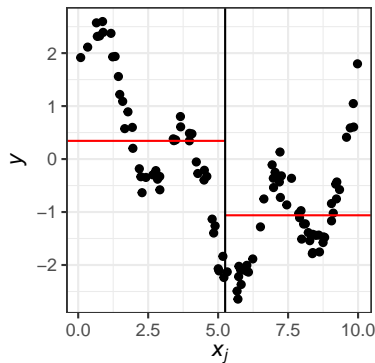
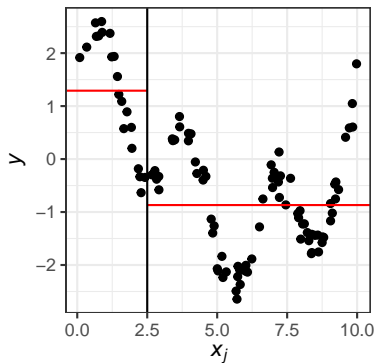
$$c_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{(\mathbf{x}, y) \in \mathcal{N}} y \text{ for the } L_2 \text{ loss, i.e., } \mathcal{R}(\mathcal{N}) = \sum_{(\mathbf{x}, y) \in \mathcal{N}} (y - c_{\mathcal{N}})^2.$$

Root node:

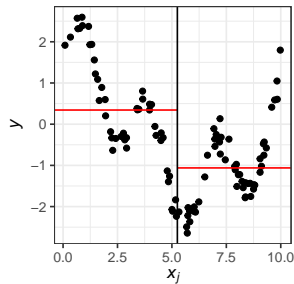
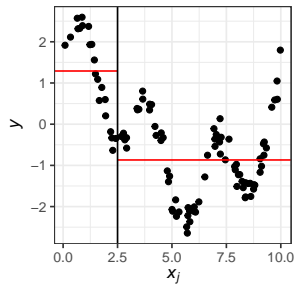


OPTIMAL CONSTANTS IN LEAVES

Which of these two splits is better?



RISK OF A SPLIT



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

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

The total risk is the sum of the individual losses:

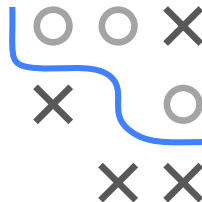
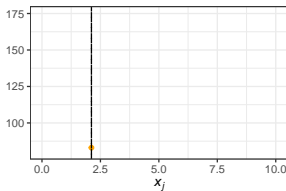
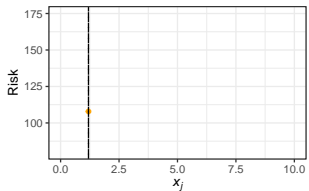
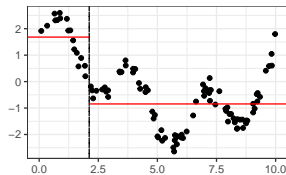
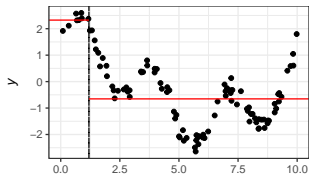
$$23.4 + 72.4 = 95.8$$

$$78.0 + 46.1 = 124.1$$

Based on the SSE, we prefer the first split.

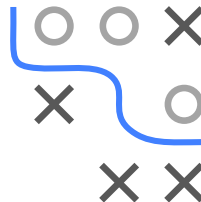
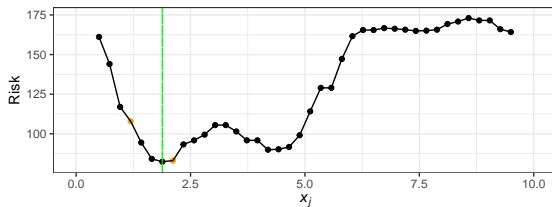
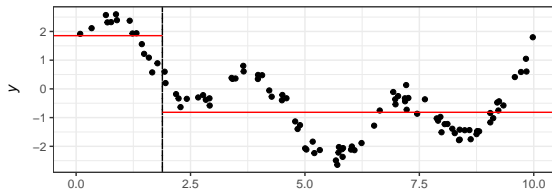
SEARCHING THE BEST SPLIT

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



FORMALIZATION

- A split w.r.t. **feature** x_j **at split point** t divides a parent node \mathcal{N} into

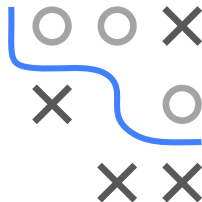
$$\mathcal{N}_1 = \{(\mathbf{x}, y) \in \mathcal{N} : x_j < t\} \text{ and } \mathcal{N}_2 = \{(\mathbf{x}, y) \in \mathcal{N} : x_j \geq t\}.$$

- To evaluate its quality, we compute the risk of our new, finer model

$$\begin{aligned}\mathcal{R}(\mathcal{N}, j, t) &= \mathcal{R}(\mathcal{N}_1) + \mathcal{R}(\mathcal{N}_2) \\ &= \left(\sum_{(\mathbf{x}, y) \in \mathcal{N}_1} L(y, c_{\mathcal{N}_1}) + \sum_{(\mathbf{x}, y) \in \mathcal{N}_2} L(y, c_{\mathcal{N}_2}) \right)\end{aligned}$$

- Finding the best way to split \mathcal{N} into $\mathcal{N}_1, \mathcal{N}_2$ means solving

$$\arg \min_{j, t} \mathcal{R}(\mathcal{N}, j, t)$$



FORMALIZATION

- $\mathcal{R}(\mathcal{N}, j, t) = \mathcal{R}(\mathcal{N}_1) + \mathcal{R}(\mathcal{N}_2)$, makes sense if \mathcal{R} is a simple sum
- If we use averages, we have to reweight the terms to obtain a global average w.r.t. \mathcal{N} as the children have different sizes

$$\bar{\mathcal{R}}(\mathcal{N}, j, t) = \frac{|\mathcal{N}_1|}{|\mathcal{N}|} \bar{\mathcal{R}}(\mathcal{N}_1) + \frac{|\mathcal{N}_2|}{|\mathcal{N}|} \bar{\mathcal{R}}(\mathcal{N}_2)$$

- We mention this for clarity, as quite a few texts contain only the (more complicated) weighted formula without clear explanation

