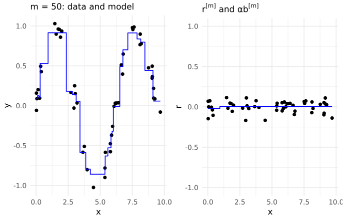
Introduction to Machine Learning

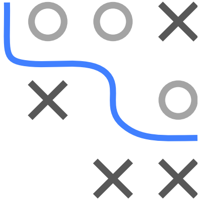
Boosting
Gradient Boosting with Trees 1



Learning goals

- Examples for GB with trees
- Understand relationship between model structure and interaction depth

GRADIENT BOOSTING WITH TREES



Trees are most popular BLs in GB.

Reminder: advantages of trees

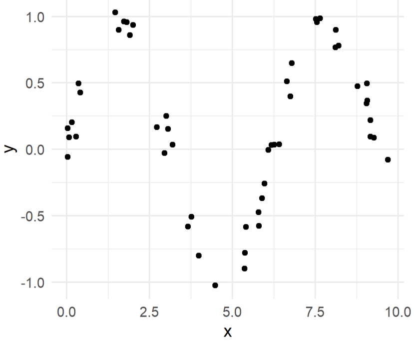
- No problems with categorical features.
- No problems with outliers in feature values.
- No problems with missing values.
- No problems with monotone transformations of features.
- Trees (and stumps!) can be fitted quickly, even for large n.
- Trees have a simple, built-in type of variable selection.

GB with trees inherits these, and strongly improves predictive power.

EXAMPLE 1

Simulation setting:

- Given: one feature x and one numeric target variable y of 50 observations.
- x is uniformly distributed between 0 and 10.
- y depends on x as follows: $y^{(i)} = \sin\left(x^{(i)}\right) + \epsilon^{(i)}$ with $\epsilon^{(i)} \sim \mathcal{N}(0, 0.01)$, $\forall i \in \{1, \dots, 50\}$.



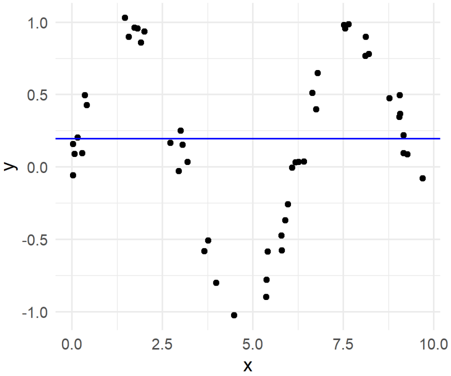
Aim: we want to fit a gradient boosting model to the data by using stumps as base learners.

Since we are facing a regression problem, we use $L2$ loss.

EXAMPLE 1 / 2

Iteration 0: initialization by optimal constant (mean) prediction $\hat{f}^{[0](i)}(x) = \bar{y} \approx 0.2$.

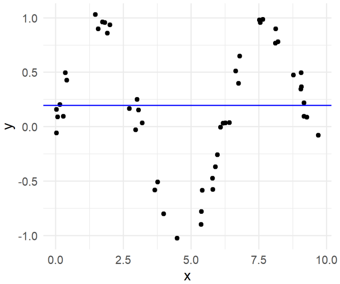
i	$x^{(i)}$	$y^{(i)}$	$\hat{f}^{[0]}$
1	0.03	0.16	0.20
2	0.03	-0.06	0.20
3	0.07	0.09	0.20
⋮	⋮	⋮	⋮
50	9.69	-0.08	0.20



EXAMPLE 1 / 3

Iteration 1: (1) Calculate pseudo-residuals $\tilde{r}^{[m](i)}$ and (2) fit a regression stump $b^{[m]}$.

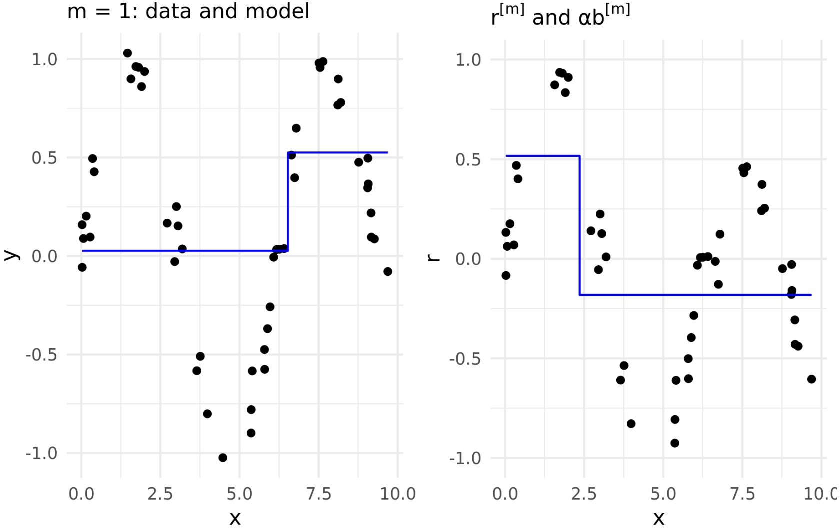
i	$x^{(i)}$	$y^{(i)}$	$\hat{f}^{[0]}$	$\tilde{r}^{[1](i)}$	$\hat{b}^{[1](i)}$
1	0.03	0.16	0.20	-0.04	-0.17
2	0.03	-0.06	0.20	-0.25	-0.17
3	0.07	0.09	0.20	-0.11	-0.17
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
50	9.69	-0.08	0.20	-0.27	0.33



(3) Update model by $\hat{f}^{[1]}(x) = \hat{f}^{[0]}(x) + \hat{b}^{[1]}$.

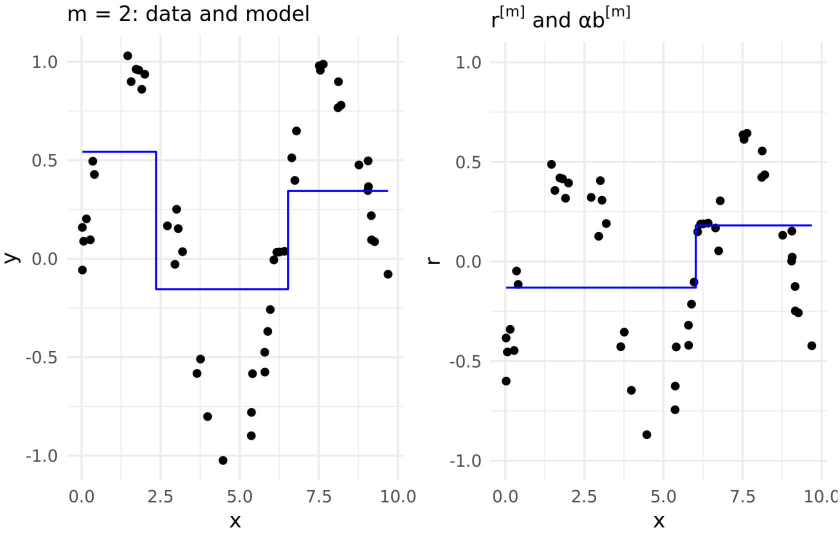
EXAMPLE 1

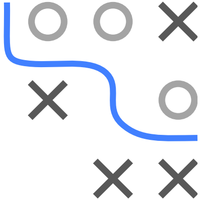
Repeat step (1) to (3):



EXAMPLE 1

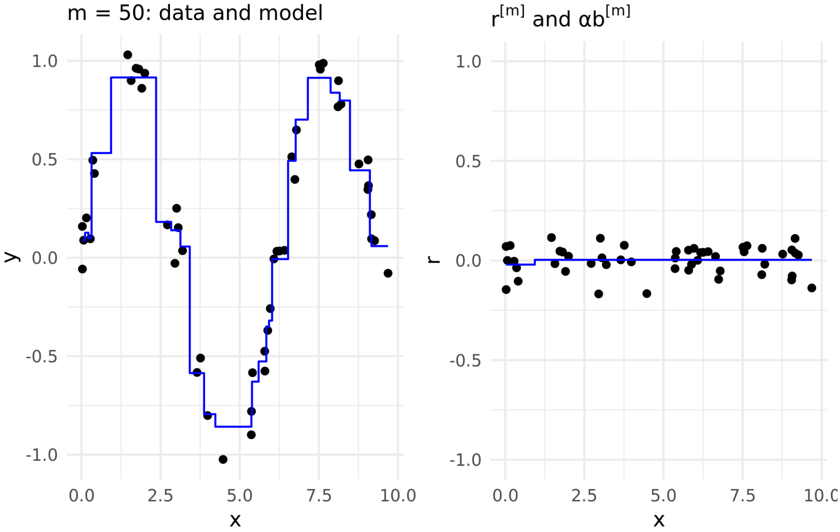
Repeat step (1) to (3):





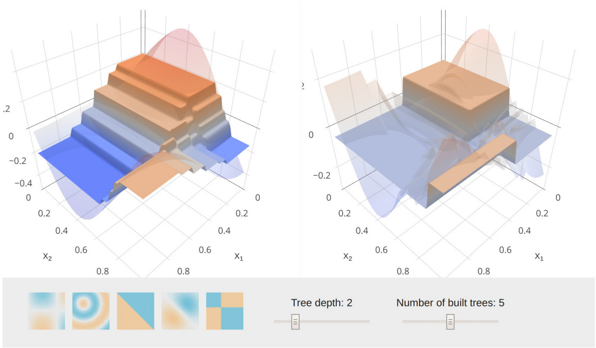
EXAMPLE 1

Repeat step (1) to (3):



EXAMPLE 2

This website shows on various 3D examples how tree depth and number of iterations influence the model fit of a GBM with trees.



MODEL STRUCTURE AND INTERACTION DEPTH

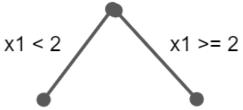
Model structure directly influenced by depth of $b^{[m]}(\mathbf{x})$.

$$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha^{[m]} b^{[m]}(\mathbf{x})$$

Remember how we can write trees as additive model over paths to leafs.

With stumps (depth = 1), $f(\mathbf{x})$ is additive model (GAM) without interactions:

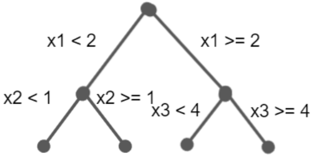
$$f(\mathbf{x}) = f_0 + \sum_{j=1}^{p} f_j(x_j)$$



With trees of depth 2, we have two-way interactions:

$$f(\mathbf{x}) = f_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{j \neq k} f_{j,k}(x_j, x_k)$$

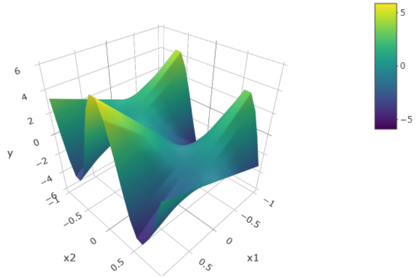
with f_0 being a constant intercept.



MODEL STRUCTURE AND INTERACTION DEPTH / 2

Simulation setting:

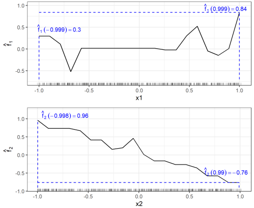
- Features x_1 and x_2 and numeric y; with $n = 500$
- x_1 and x_2 are uniformly distributed between -1 and 1
- $y^{(i)} = x_1^{(i)} - x_2^{(i)} + 5\cos(5x_2^{(i)}) \cdot x_1^{(i)} + \epsilon^{(i)}$ with $\epsilon^{(i)} \sim \mathcal{N}(0, 1)$
- We fit 2 GB models, with tree depth 1 and 2, respectively.

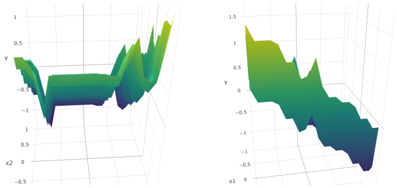


MODEL STRUCTURE AND INTERACTION DEPTH / 3

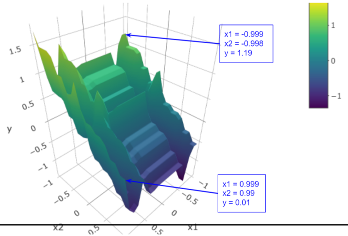
GBM with interaction depth of 1 (GAM)

No interactions are modelled: Marginal effects of x_1 and x_2 add up to joint effect (plus the constant intercept $\hat{f}_0 = -0.07$).

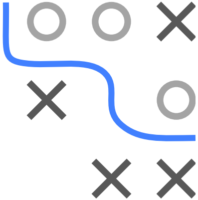




$$\hat{f}(-0.999, -0.998)$$
$$= \hat{f}_0 + \hat{f}_1(-0.999) + \hat{f}_2(-0.998)$$
$$= -0.07 + 0.3 + 0.96 = 1.19$$



MODEL STRUCTURE AND INTERACTION DEPTH / 4



GBM with interaction depth of 2

Interactions between x_1 and x_2 are modelled: Marginal effects of x_1 and x_2 do NOT add up to joint effect due to interaction effects.