

# Statistics 360: Advanced R for Data Science

## MARS, part II

Brad McNeney



## Recursive partitioning

- ▶ In the week 4 exercises we will implement the forward part of an algorithm called recursive partitioning.
- ▶ Friedman (1991) recasts recursive partitioning in terms of a forward stepwise regression procedure that selects products of mirror-image step functions  $I((x_v - t) \geq 0) = I(x_v \geq t)$  and  $I(-(x_v - t) \geq 0) = I(x_v \leq t)$ .
- ▶ Exercise: draw these two step functions to see why they are mirror image.
- ▶ Note: I think the first mirror-image step function should be  $I(x_v > t)$ , but will not make any changes to the discussion in the paper.

# Region splitting as a product of step functions

- ▶ Take the case of two covariates and a region  
 $R = [a_1, b_1] \times [a_2, b_2]$ .
- ▶ Claim:  $R$  is the set of points  $(x_1, x_2)$  such that the basis function

$$B_R(x) = I(x_1 \geq a_1) \times I(x_1 \leq b_1) \times I(x_2 \geq a_2) \times I(x_2 \leq b_2) > 0$$

- ▶ Splitting  $R$  on variable  $x_v$  at point  $t \in [a_v, b_v]$  means removing a “parent” basis function  $B_R(x)$  from the model and replacing it with two “children” basis functions  $B_R(x)I(x_v \leq t)$  and  $B_R(x)I(x_v \geq t)$ .

# Recursive partitioning forward algorithm

- ▶ See page 11 of the paper.
- ▶ Outer loop 1 over the number of model terms, from 1 to some maximum number
  - ▶ Outer loop 2 over parent basis functions to replace by splitting
    - ▶ Inner loops to choose variables  $v$ , splits  $t$  in the region where  $B_R(x) > 0$ , and coefficients for  $B_R(x)I(x_v \leq t)$  and  $B_R(x)I(x_v \geq t)$  to minimize a LOF criterion. (This is like our recursive partitioning.)

## Week 5/lab 3

- ▶ Re-implement recursive partitioning with the forward stepwise regression algorithm outlined above.
- ▶ A more detailed breakdown of the tasks will follow.

# MARS generalization

- ▶ Replace the step functions with hinge functions  $h(t - x_v)$  and  $h(x_v - t)$ , where  $h(x) = \max(0, x)$ .
- ▶ Do not remove a parent basis function, just add pairs of children.
- ▶ Restrict the product that defines a basis function to distinct variables; i.e., no variable appears twice in the product.

# MARS forward algorithm

- ▶ See page 17 of the paper.
- ▶ Outer loop 1 over the number of model terms,  $M$ , from 1 to some maximum number
  - ▶ Outer loop 2 over parent basis functions  $B_m$  to generate children
    - ▶ Inner loops over variables  $v$  not part of  $B_m$ , splits  $t$  such that  $B_m$  is positive for  $x_v = t$  and coefficients for the child basis functions to minimize a LOF criterion.