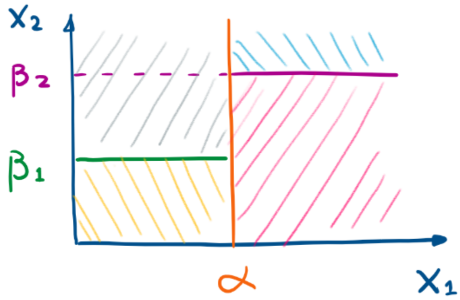
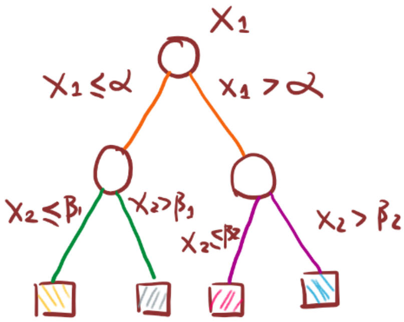
CART Note

Zengchang Qin

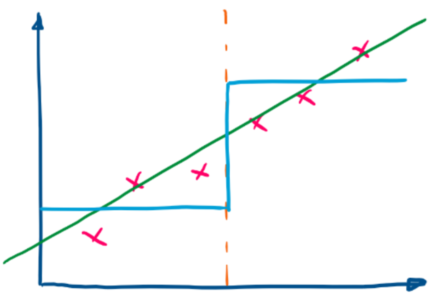
CART - classification and Regression Tree

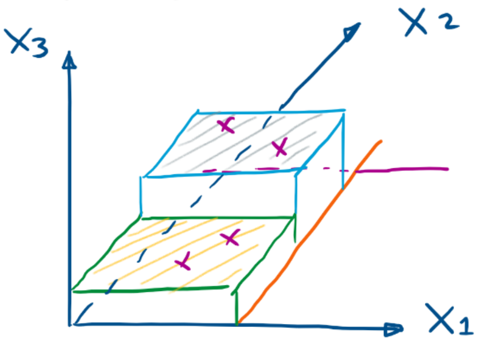
1) We have learned that a decision tree is a model that divides the input space into a few regions.





2) If we only consider one-dimensional variable x. how can we use a tree model for regression?





3) Given training data $D = \{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \cdots (x^{(N)}, y$

We hope to learn a CART tree that minimize the following cost function:

$$Loss = \min_{j,s} \left[\min_{C_1} L(y^{(i)}, C_1) + \min_{C_2} L(y^{(i)}, C_2) \right]$$

where $C_m = \text{ave}(y_i \mid x_i \in R_m)$

In the CART Model, we need to find the best variable x_j and cut point s in order to minimize the loss function. We can rewrite the loss by:

$$Loss = \min_{j,s} \left[\min_{C_1} \sum_{x \in R_1(j,s)} (y^{(i)} - C_1)^2 + \min_{C_2} \sum_{x \in R_2(j,s)} (y^{(i)} - C_2)^2 \right]$$

where $R_1(j,s) = \{ x \mid x_j \le s \}$

$R_2(j,s) = \{ x \mid x_j > s \}$

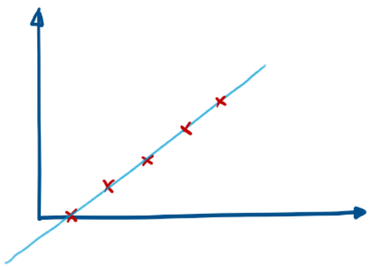
$C_m = \frac{1}{N_m} \sum_{x \in R_m(j,s)} y^{(i)}, \quad m = 1, 2$

4) An example:

x	1	2	3	4	5	6	7	8	9	10
y	0	1	2	3	4	5	6	7	8	9

$(j=1)$

1.5   4.5



Given $S = 1.5$

$R_1 = \{0\}$ $\quad C_1 = 0$

$R_2 = \{1, 2, 3, \cdots 9\}$ $\quad C_2 = 5$

$Loss(1.5) = 0^2 + \frac{1}{9} \sum_{i=1}^{9} (i-5)^2$

$\vdots$

when $S = 4.5$

$R_1 = \{0, 1, 2, 3\}$ $\quad C_1 = 1.5$

$R_2 = \{4, 5, 6 \cdots 9\}$ $\quad C_2 = 6.5$

$Loss(4.5) = \sum_{i=0}^{3} (i-1.5)^2 + \sum_{j=4}^{9} (j-6.5)^2$

we need to calculate Loss (s) for $s = 1.5, 2.5 \cdots 9.5$ to find the best cut point s. If we have more than one variable, we need to search $(j = 1, 2 \cdots n)$ with best cut point s_j.