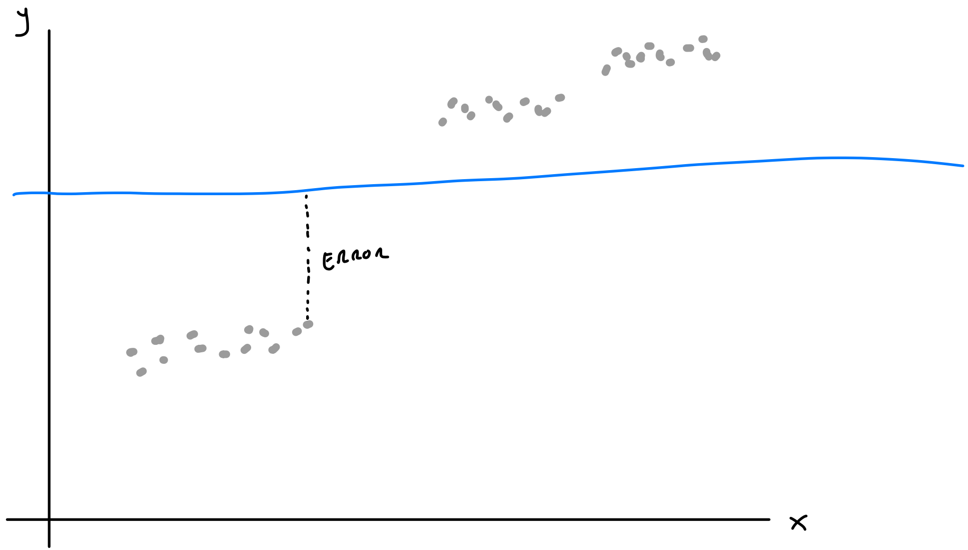
CS 307

Spring 2024

Dalpiaz

Decision Trees

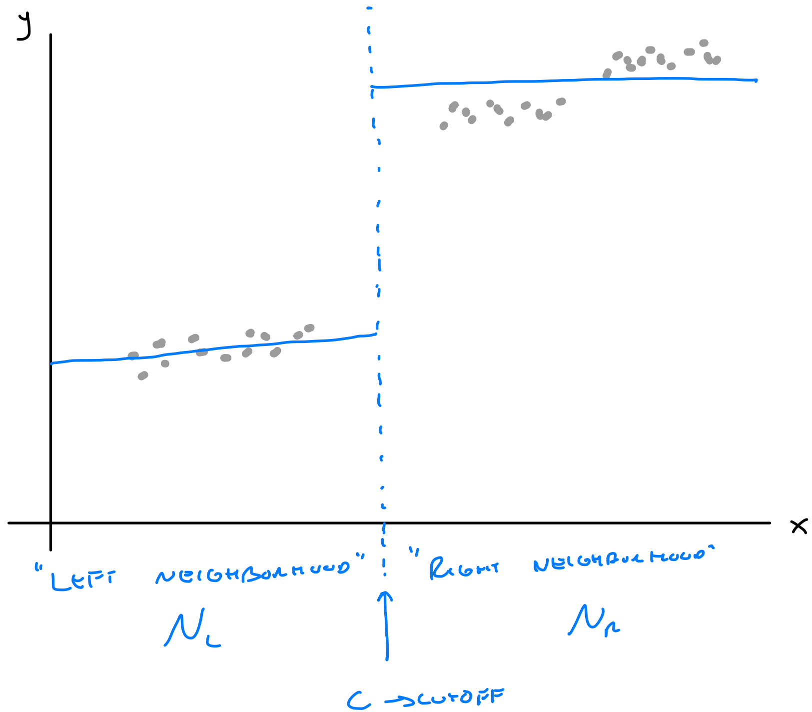
Regression

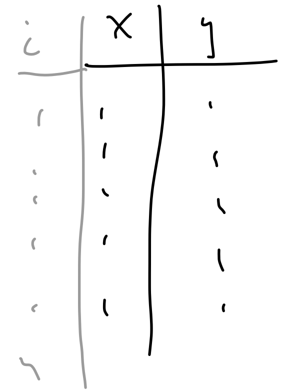


$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

SUM SQUARES TOTAL

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$





IDEA: CREATE NEIGHBORHOODS, PREDICT MEAN IN NEIGHBORHOOD

Decision Tree

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

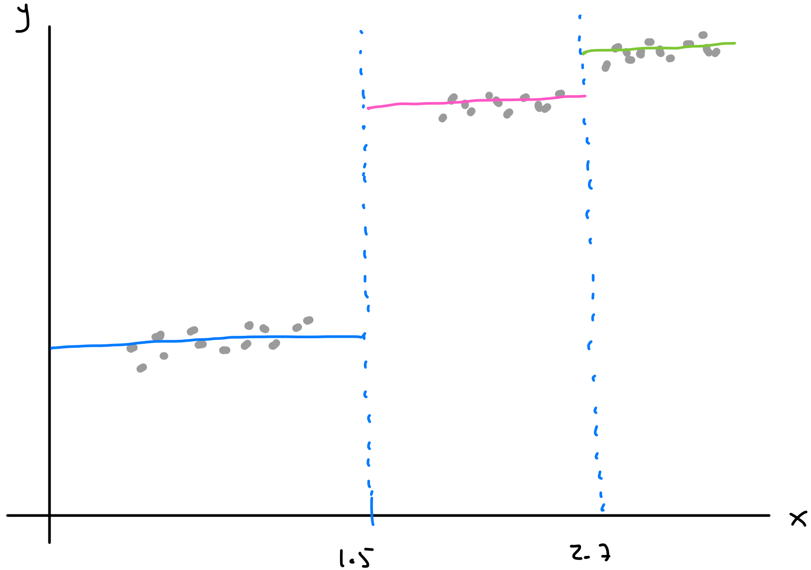
$$SSE = \sum_{i \in N_L} (y_i - \bar{y}_L)^2 + \sum_{i \in N_R} (y_i - \bar{y}_R)^2$$

error

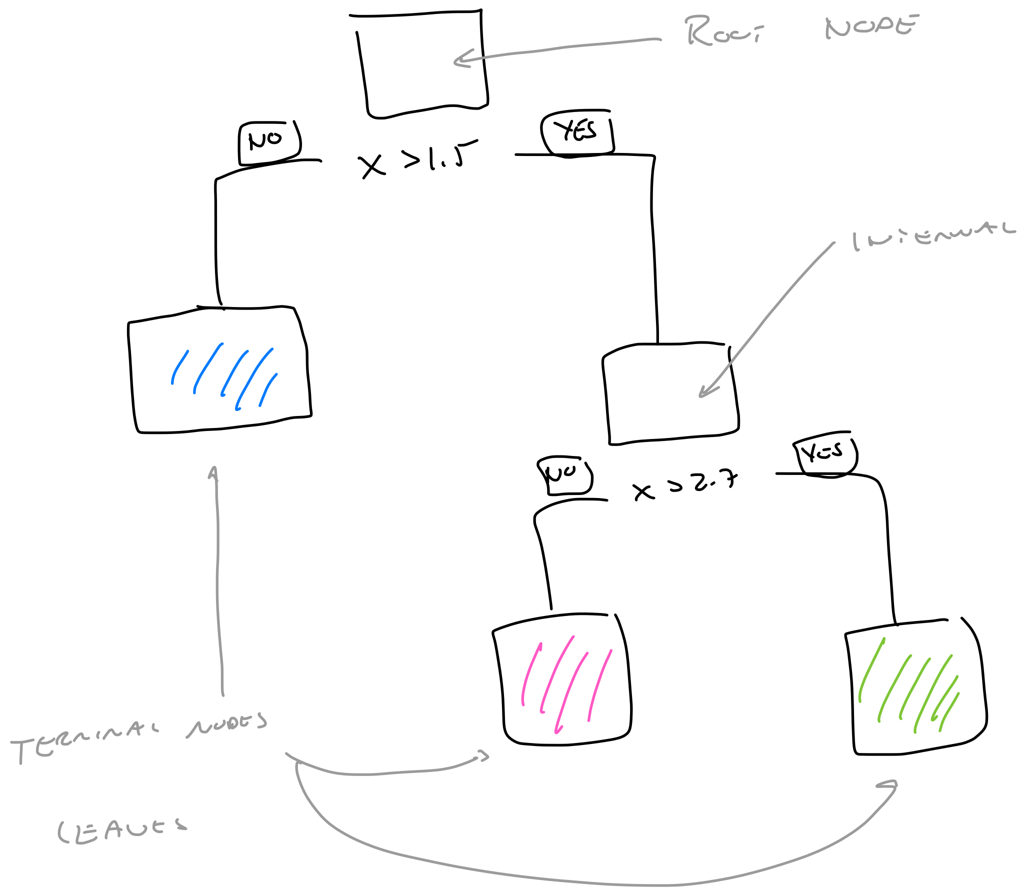
mean of y in N_L

Find c that minimizes SSE

considering all features



$$SSE = \sum_{i \in N_L} (y_i - \bar{y}_L)^2 + \sum_{i \in N_{R_L}} (y_i - \bar{y}_{R_L})^2 + \sum_{i \in N_{R_R}} (y_i - \bar{y}_{R_R})^2$$


ROOT NODE
NO
x > 1.5
YES
INTERNAL
NO
x > 2.7
YES
TERMINAL NODES
LEAVES

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

ONE SPLIT

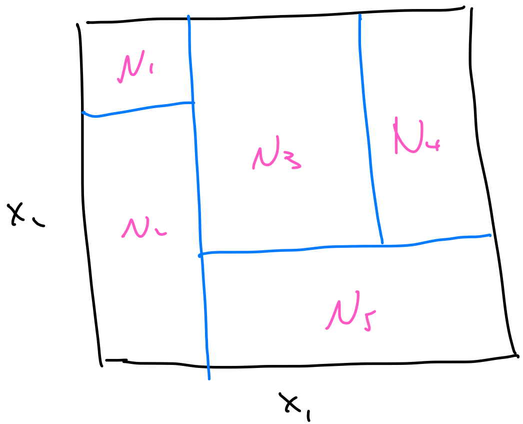
$$SSE = \sum_{i \in N_L} (y_i - \bar{y}_L)^2 + \sum_{i \in N_R} (y_i - \bar{y}_R)^2$$

TWO SPLITS

$$SSE = \sum_{i \in N_L} (y_i - \bar{y}_L)^2 + \sum_{i \in N_{R_L}} (y_i - \bar{y}_{R_L})^2 + \sum_{i \in N_{R_R}} (y_i - \bar{y}_{R_R})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

RECURSIVE PARTITIONING



$$SSE = \sum_{j=1}^{J} \sum_{i \in N_j} \left(y_i - \bar{y}_j \right)^2$$

MEAN OF POINTS IN NODE J, N_j

WHEN TO STOP?

MIN SAMPLES SPLIT

DEPTH