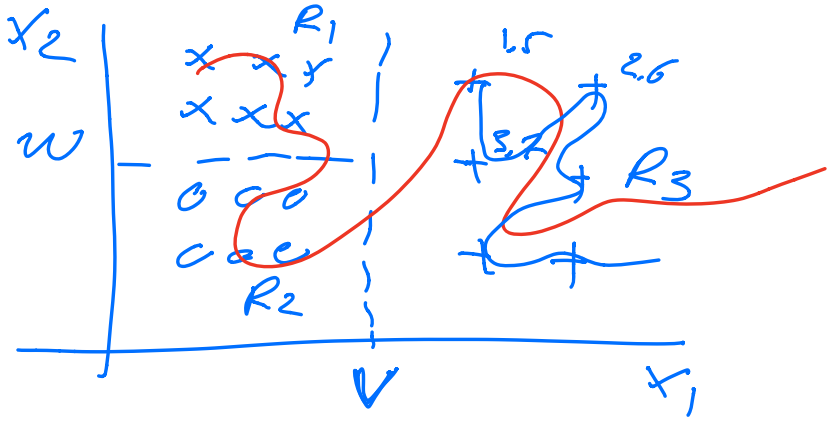
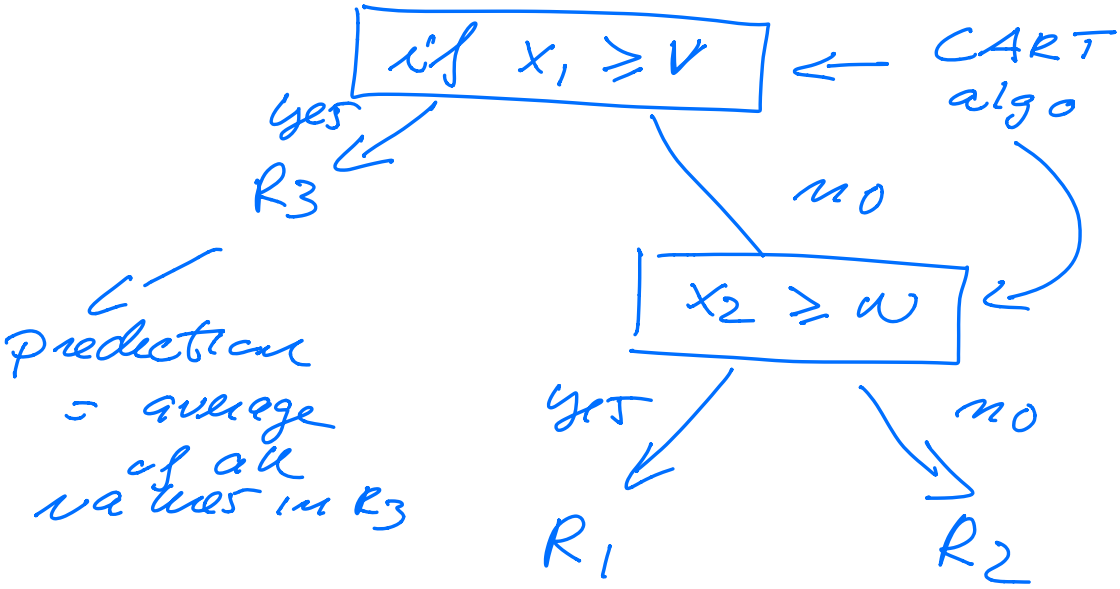
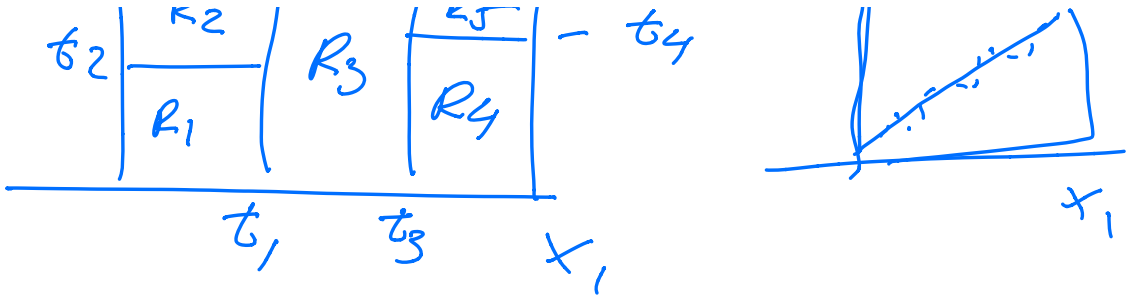
Lecture January 28

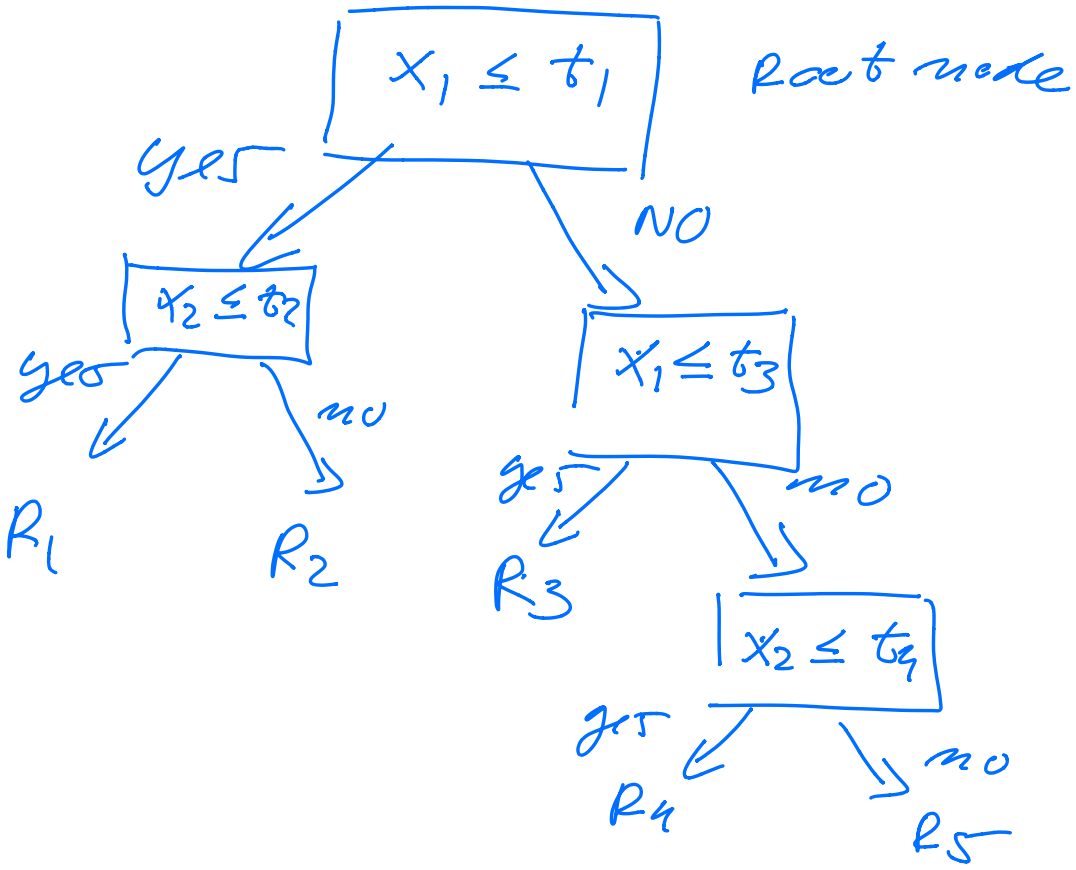




prediction = average of all values in R_3

prediction is the mean value of all data points in R_1





Algorithm for Regression case (CART)

(y_i, X_i) $i = 0, 1, \ldots n-1$

$X_i = (x_{i0}, x_{i1}, \ldots x_{ip-1})$

- Partition data into M-Regions $R_1, R_2, \ldots R_M$
- Model Response as a constant c_m in each region

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

$$\hat{c}_m = \text{ave}(y_i \mid x_i \in R_m)$$

- Minimize

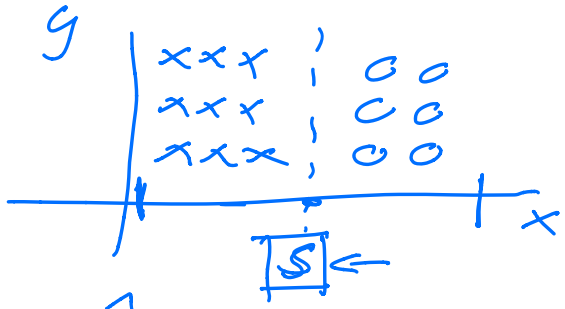
$$\sum_i (y_i - f(x_i))^2$$

- $R_1(j, s) = \{X \mid X_j \le s\}$
- $R_2(j, s) = \{X \mid X_j > s\}$

- Minimize

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2} (y_i - c_2)^2\right]$$

$$= I(j, s)$$



$$\hat{c}_1 = \text{ave}(y_i \mid x_i \in R_1(j,s))$$

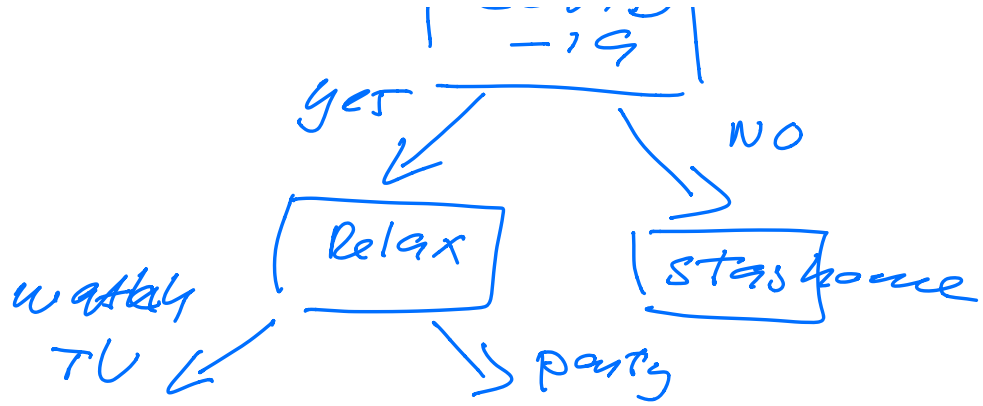
$$\hat{c}_2 = \text{ave}(y_i \mid x_i \in R_2(j,s))$$

$$J(j,s) = \frac{N_{left}}{N} MSE_{left} + \frac{N_{right}}{N} MSE_{right}$$

Classification algo (CART)

$$J(j,s) = \frac{N_{left}}{N} G_{left} + \frac{N_{right}}{N} G_{right}$$

G_{left}/G_{right} ? — Gini



Define a node $-m-$ (Represents a region R_m with N_m observations)

Proportion of class/category $-k-$ observations in node m

$$P_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Classify observations in node $-m-$

$$k(m) = \arg\max_k P_{mk}$$

Misclassification

(i) $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m))$

$1 - p_{mk}$

(ii) Gini index (Default in sci-kit-learn)

K = classes

$\sum_{k=1} p_{mk}(1 - p_{mk})$

(iii) Entropy

$-\sum_{k=1}^{K} p_{mk} \log p_{mk}$