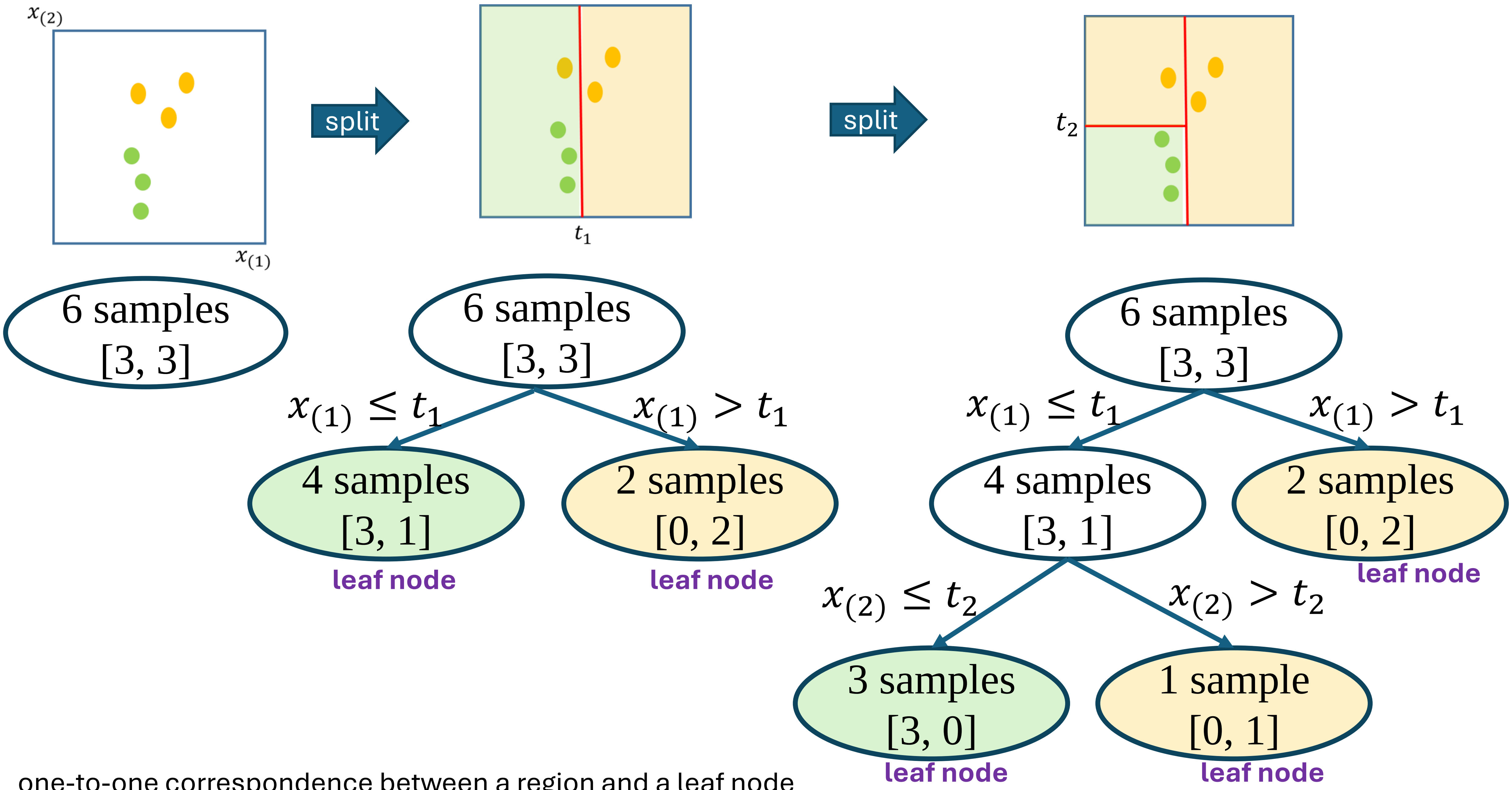
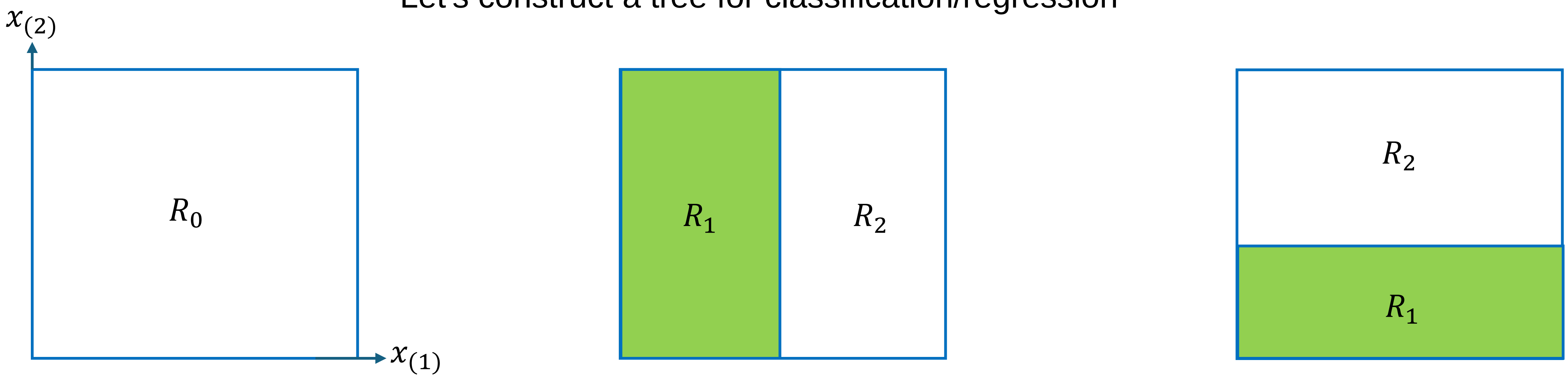


- one-to-one correspondence between a region and a leaf node
- splitting a region = splitting a node

Let's construct a tree for classification/regression

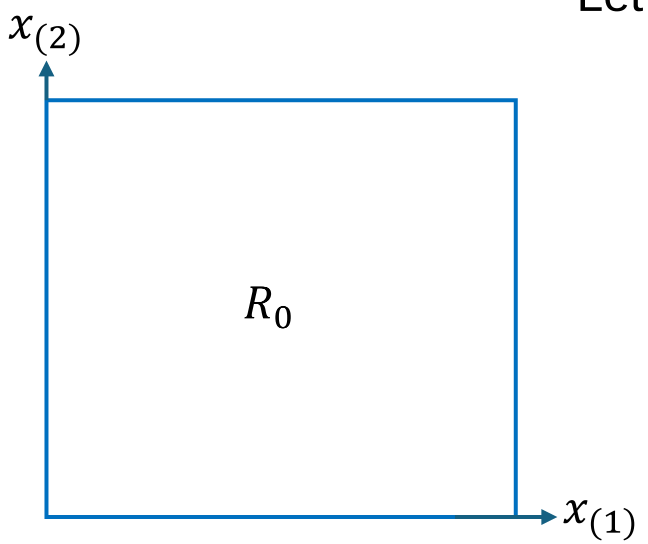


option1: split the region using $x_{(1)}$

option2: split the region using $x_{(2)}$

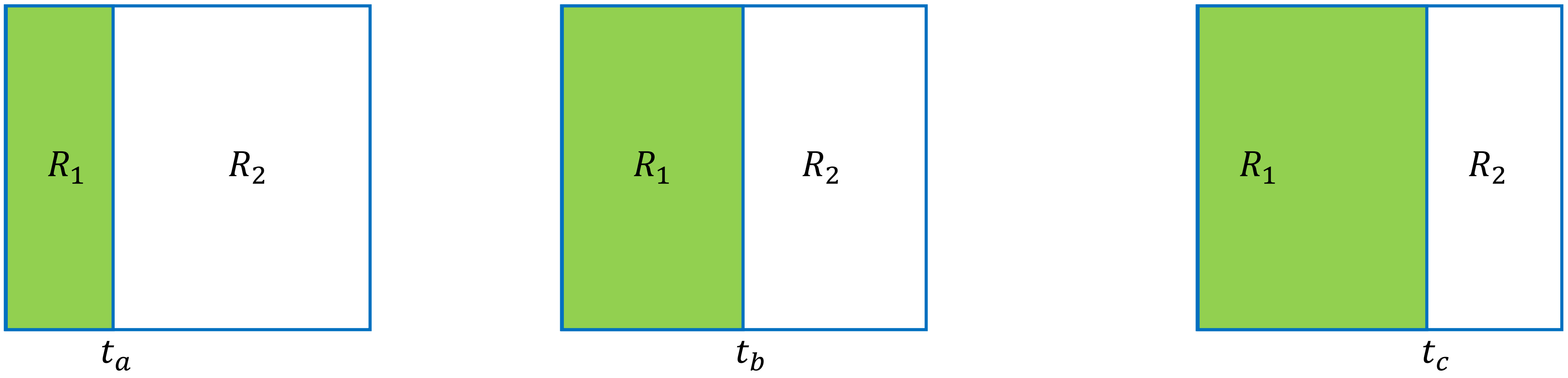
- **randomly** select **<u>a subset of features</u>.**
  - For example, $\{x_{(1)}, x_{(2)}\}$ from $\{x_{(1)}, x_{(2)}, \ldots, x_{(100)}\}$
- For each feature $x_{(s)}$ in the subset (e.g., $\{x_{(1)}, x_{(2)}\}$), find the best split/threshold $t_{(s)}$, such that the objective $E(t_{(s)})$ is maximized: $E(t_{(s)}) = Q(R_0) - \frac{N_1}{N_0} Q(R_1) - \frac{N_2}{N_0} Q(R_2)$
- Choose the feature that has the maximum objective
  - for example, if $E\left(t_{(1)}\right) > E(t_{(2)})$, then $x_{(1)}$ with $t_{(1)}$ is selected for splitting the region R_0

Let's construct a tree for classification/regression



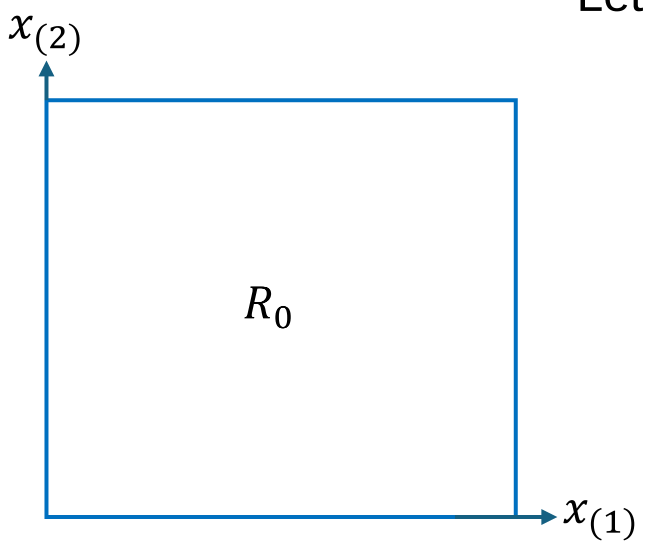
option1: split the region using $x_{(1)}$

We can try different threshold values and find the best one



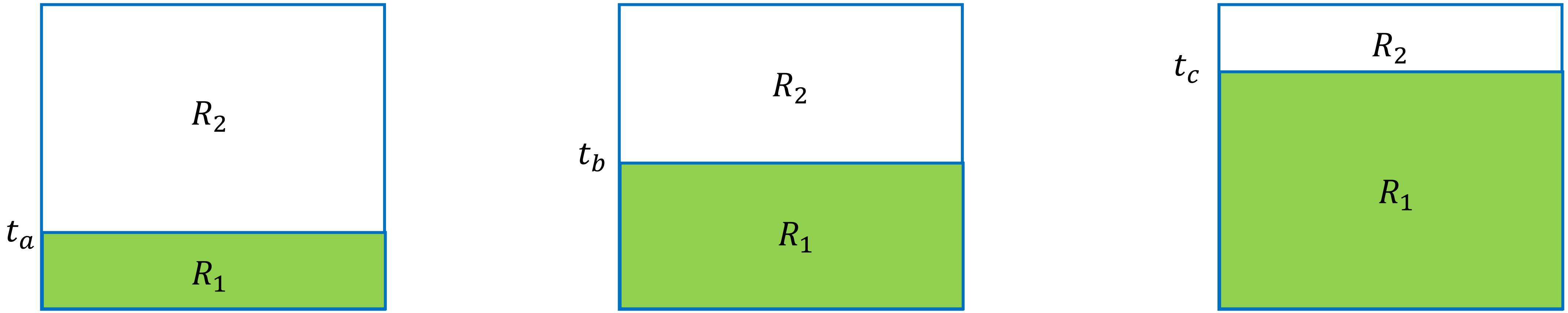
- For each threshold t, the value of the objective function is $E(t) = Q(R_0) - \frac{N_1}{N_0} Q(R_1) - \frac{N_2}{N_0} Q(R_2)$
- Assuming $E(t_a) > E(t_b) > E(t_c) > \cdots$, then the best threshold $x_{(1)} = t_a$

Let's construct a tree for classification/regression



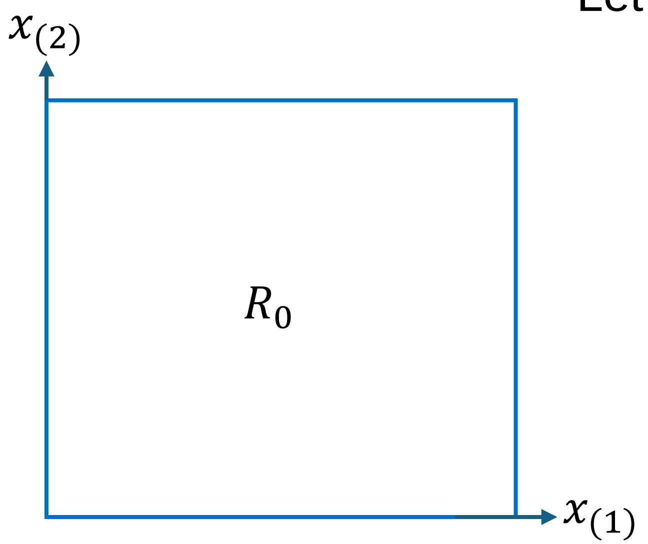
option2: split the region using $x_{(2)}$

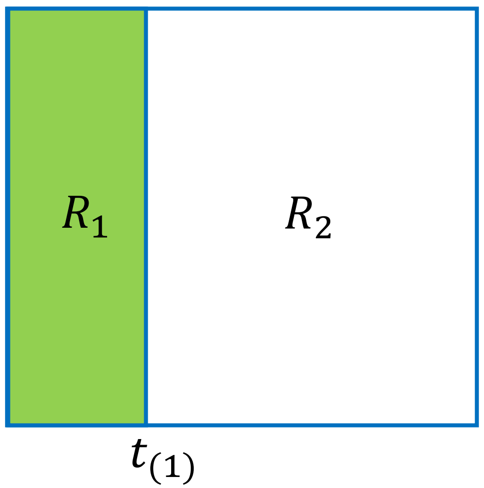
We can try different threshold values and find the best one



- For each threshold t, the value of the objective function is $E(t) = Q(R_0) - \frac{N_1}{N_0} Q(R_1) - \frac{N_2}{N_0} Q(R_2)$
- Assuming $E(t_a) > E(t_b) > E(t_c) > \cdots$, then the best threshold $x_{(2)} = t_a$

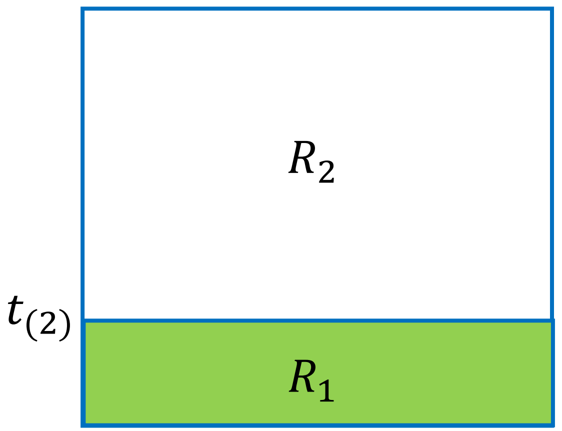
Let's construct a tree for classification/regression





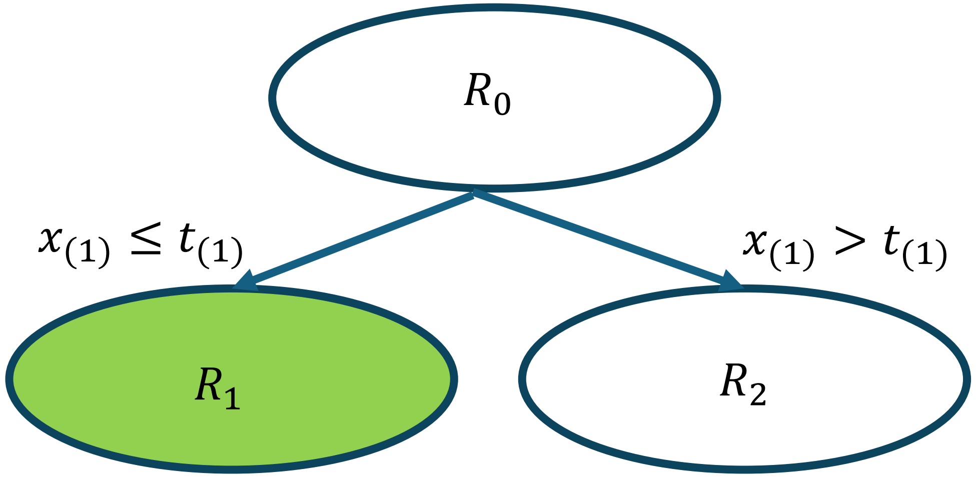
option1: split the region using $x_{(1)}$

or



option2: split the region using $x_{(2)}$

- Assume that $E\left(t_{(1)}\right) > E(t_{(2)})$, then $x_{(1)}$ with $t_{(1)}$ is selected for splitting the region R_0



$Q(R_j)$ measures the impurity of the node/region *j*

goal: minimize $Q(R_j)$ to get a node/region as pure as possible

For classification:

- entropy $H = -\sum_{k=1}^{K} \hat{p}_{(j,k)} log \hat{p}_{(j,k)}$
- Gini index: $\sum_{k=1}^{K} \hat{p}_{(j,k)}(1 - \hat{p}_{(j,k)})$ very similar to entropy

For Regression:

- $Q(R_j) = MSE = \frac{1}{N_j} \sum_{x_n \in R_j} (y_n - c_j)^2$

- A node is pure if the MSE is 0