CART : Classification And Regression Trees

the context:

we have data $\mathcal{L} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$

where $X_i = (X_i^1, \dots, X_i^p)$: the vector that contains the observation

of the p explanatory variables for the individual i.

y_i : the observation of the response variable for the individual i.

In regression: $\forall i \in \{1, \dots, n\}$,

$Y_i \in \mathbb{R}$

In classification: $\forall i \in \{1, \dots, n\}$,

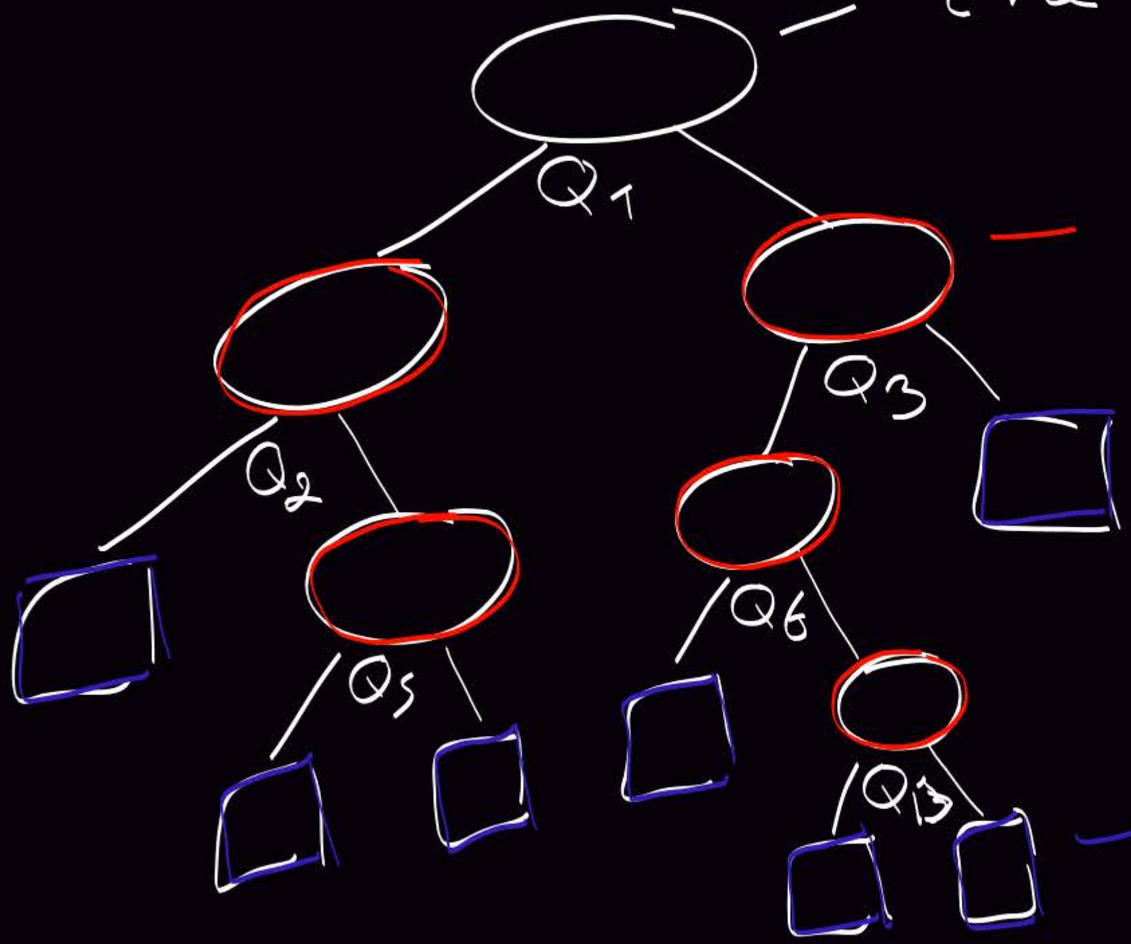
$Y_i \in \mathcal{C} = \{1, \dots, I\}$

$\Rightarrow Y$ is a qualitative variable.

In both cases, the goal is the same:

find a function that makes a link between the response variable and the explanatory ones.

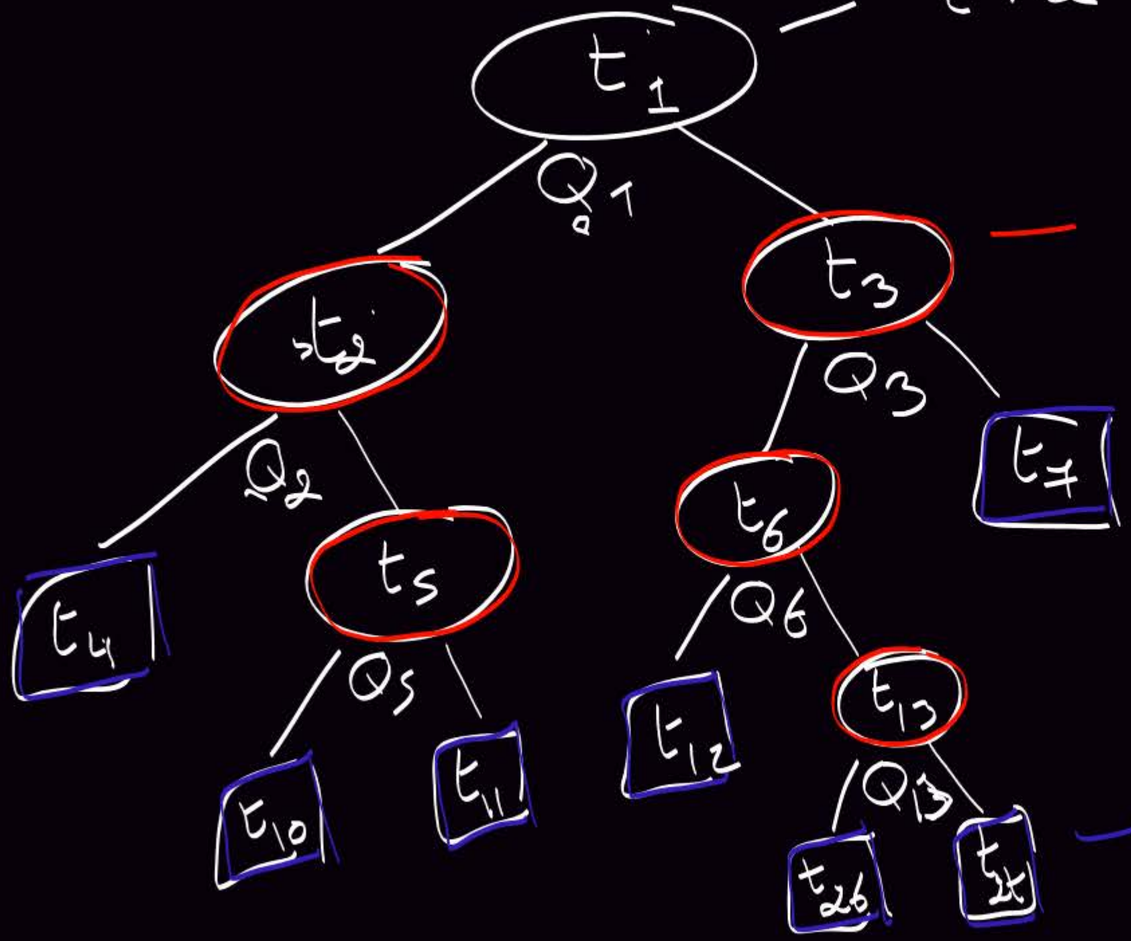
what is the Tree ?



- the root
- the node
- the leaf

$Q_1, Q_2, Q_3, \ldots$ are binary questions (two issues YES and NO)

what is the Tree ?



— the root

— the node

$Q_1, Q_2, Q_3, \ldots$ are binary question (two issues: YES and NO)

— the leaf

?

1) How to define the binary questions to split the node?

2) How to decide that a node is a leaf?

3) How this tree can help me to have an idea of the link between the respons and the explanatory variables?

1) How to define the binary questions?

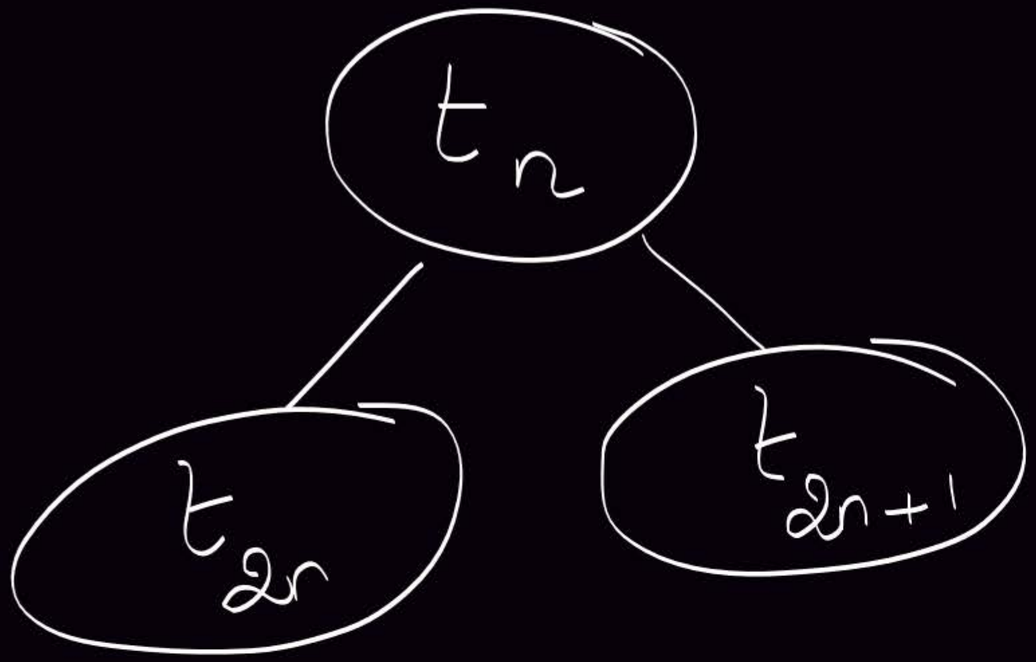
The binary questions are

$(X^i \leq a_i)$ or $(X^i > a_i)$ ← for X^i which numeric

or $(X^i \in \mathcal{C}_i)$ ← for X^i which is qualitative

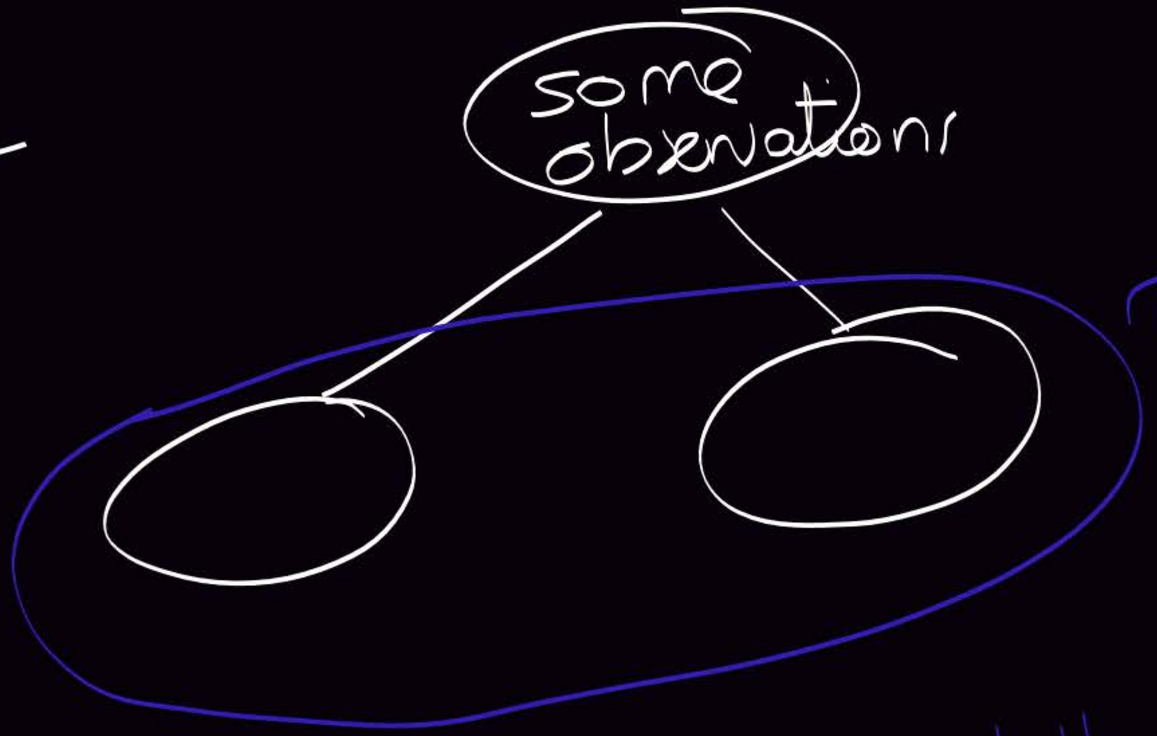
2 questions now that we know the form of the binary question:

1) How to define the variable involved in the question?
2) How to define a_i or b_i?


t_n
t_{2n}
t_{2n+1}

The answer to those two questions will depend on the framework.

Why?



partition of the previous observations

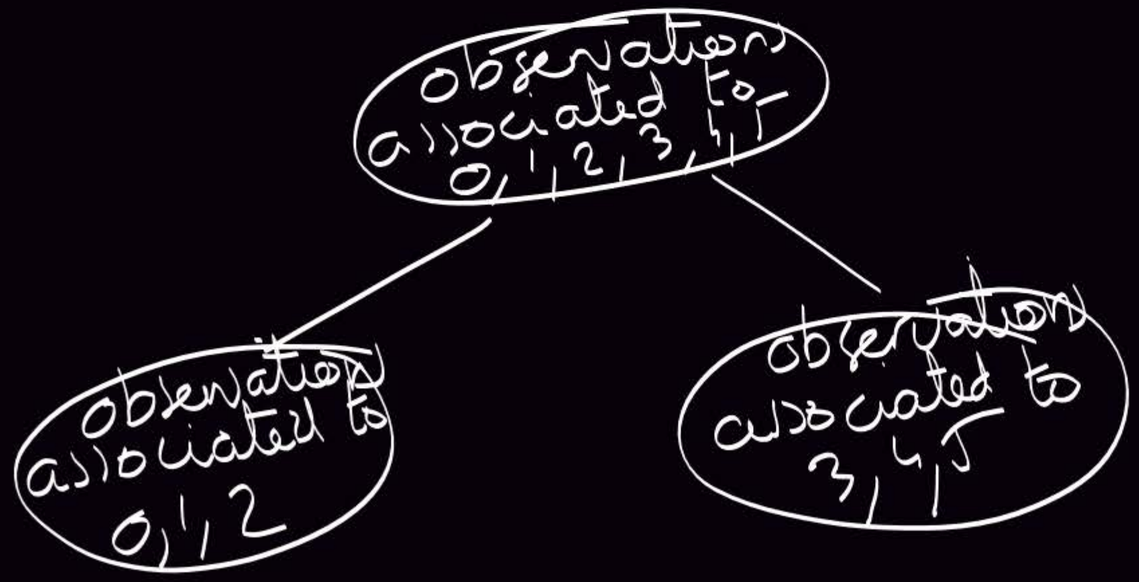
⇒ a better separation of the response

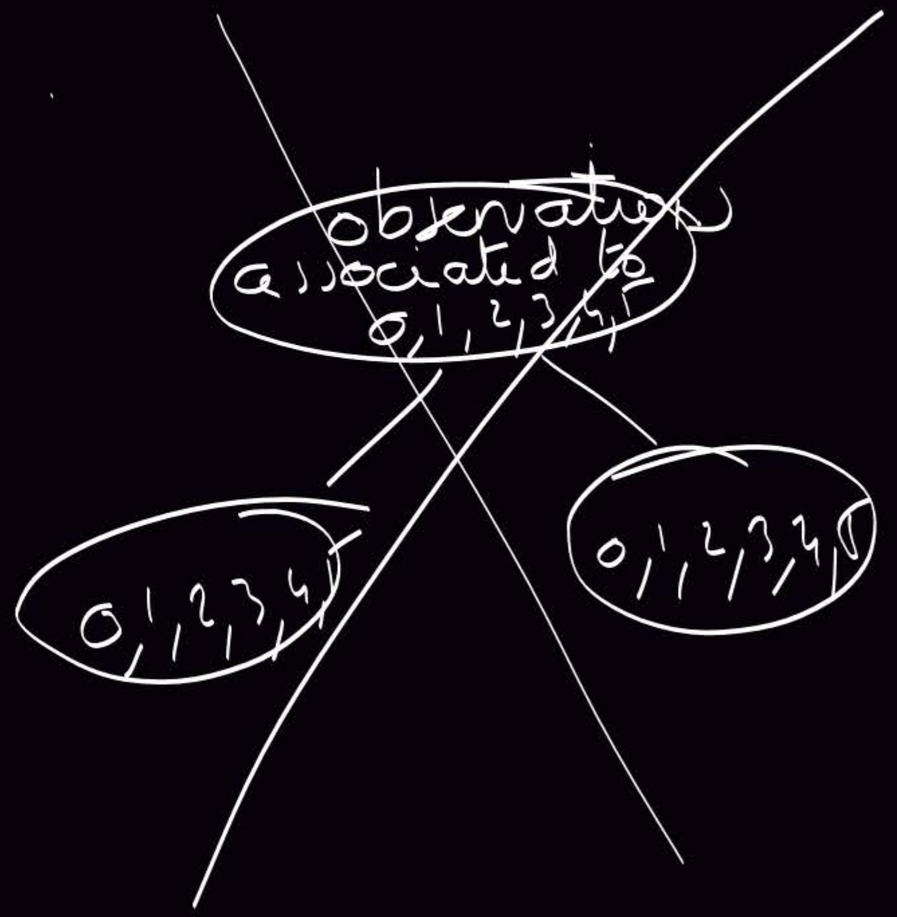
Example:

In classification:

judgment for a film.

0 1 2 3 4 5





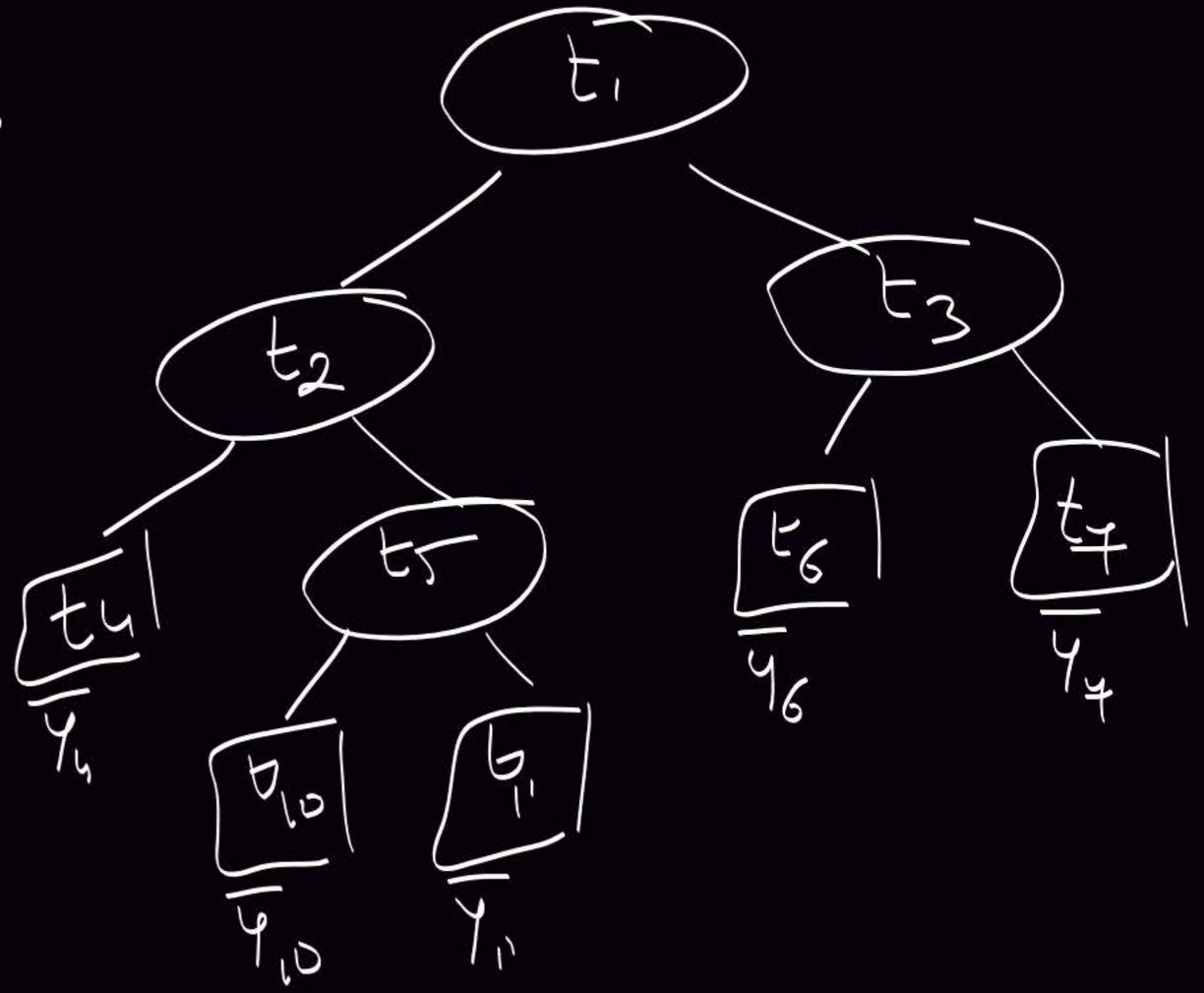
to be able to give an answer to those two points, we need to define:

- the notion of predicted value associated to a node
- a criteria to measure the purity of a node

Regression framework

$$Y = f(X) + \varepsilon$$

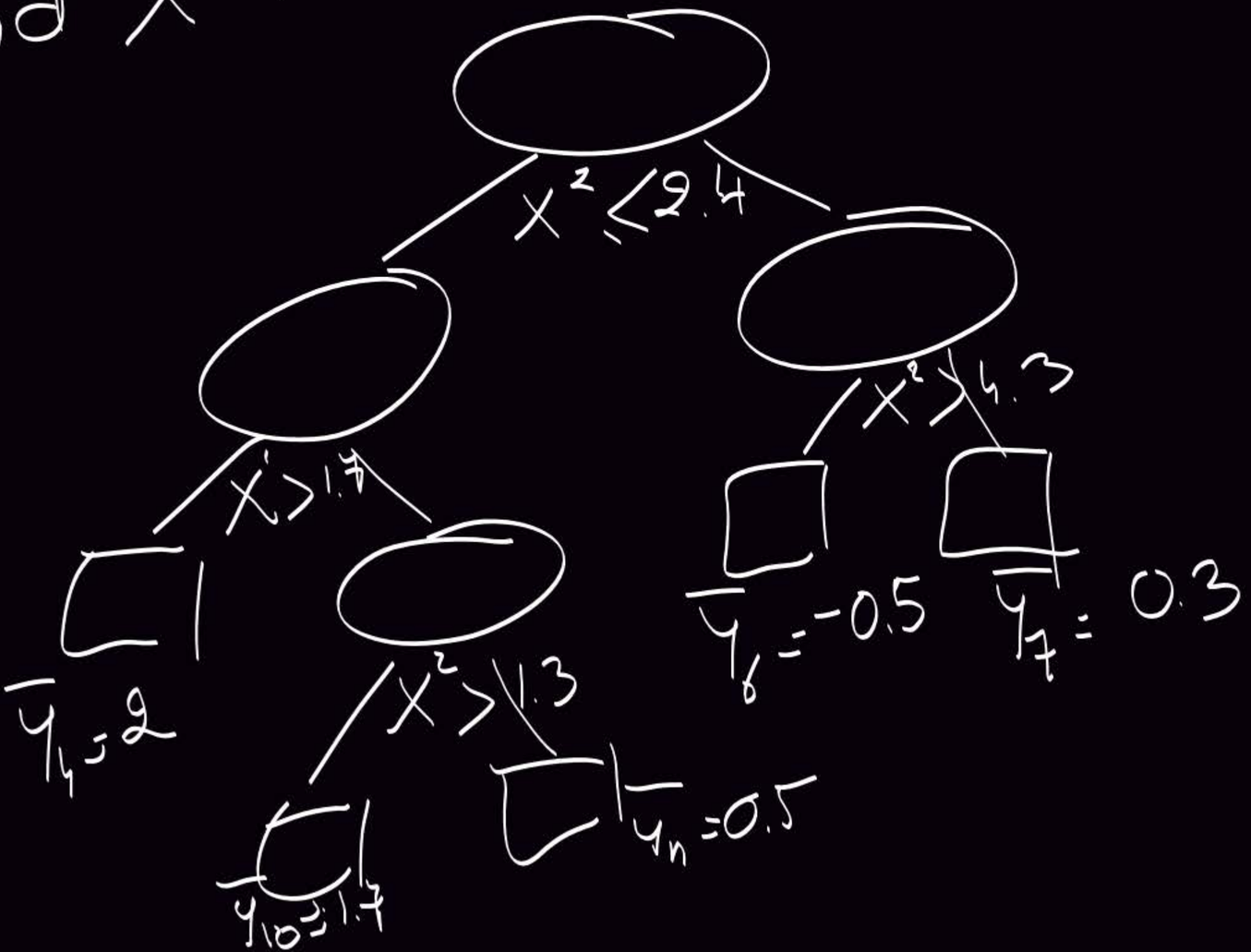
with $X = (X^1, \ldots, X^p)$



$$\hat{Y} = \begin{cases} \bar{Y}_4 & \text{if } Q_1 \text{ is YES and } Q_2 \text{ is YES} \\ \bar{Y}_{10} & \text{if } Q_1 \text{ is YES, } Q_2 \text{ is NO, } Q_5 \text{ is YES} \\ \bar{Y}_{11} & \ldots \\ \bar{Y}_6 & \ldots \\ \bar{Y}_7 & \ldots \end{cases}$$

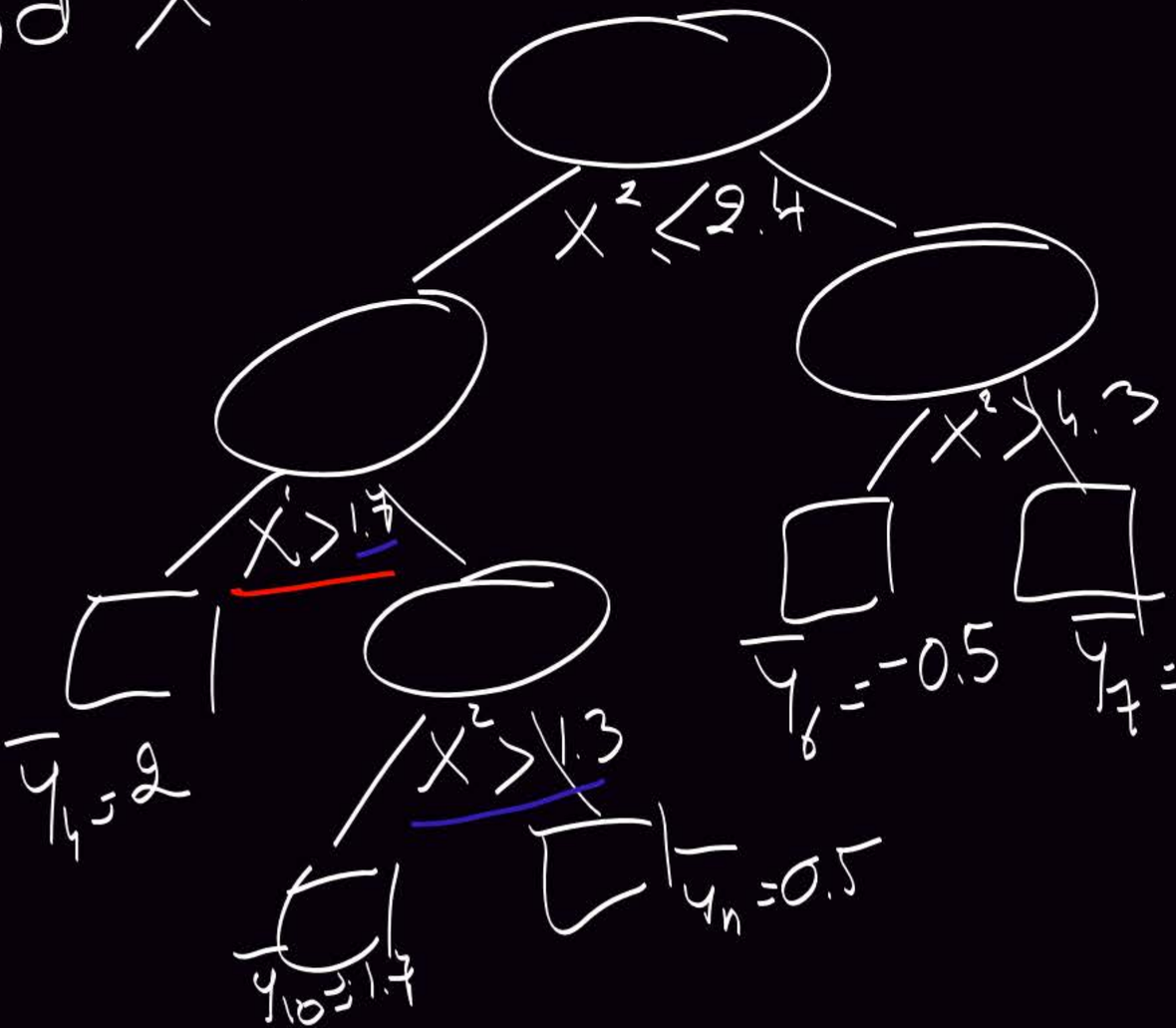
Example:

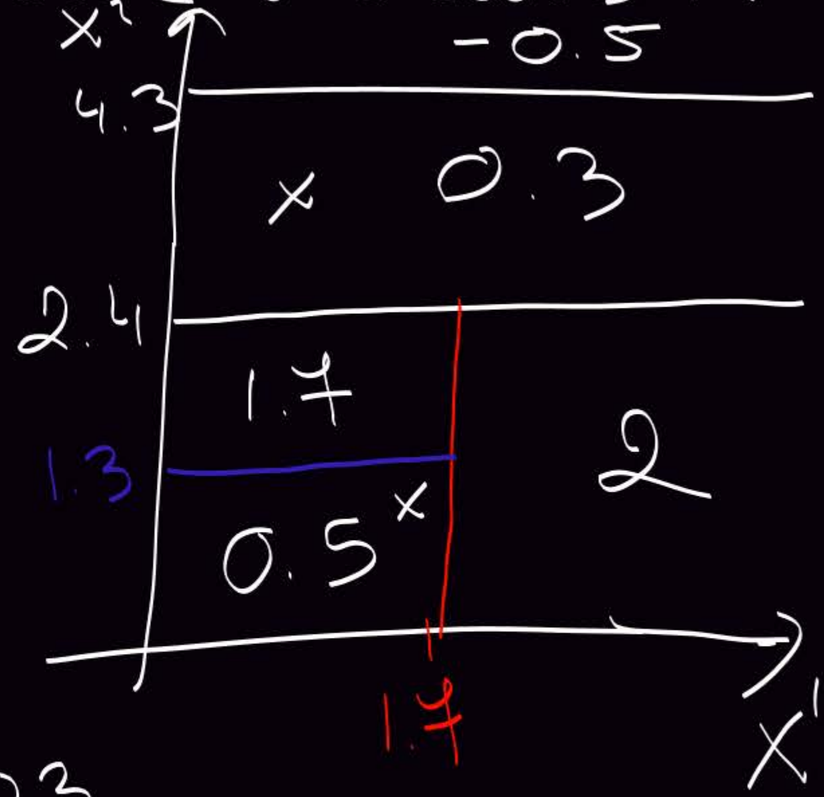
a data set with 2 quantitative variables X^1 and X^2.



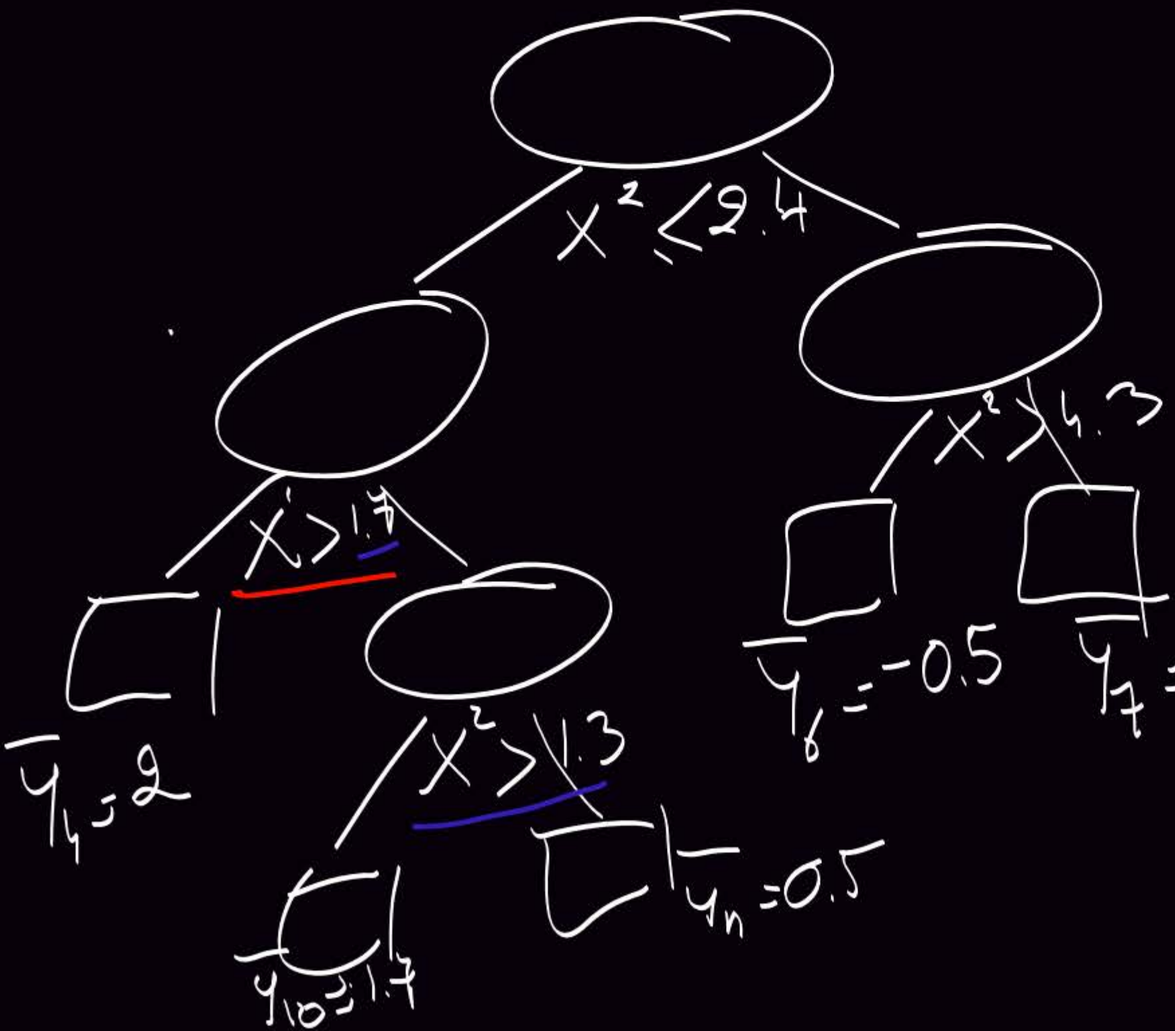
Example:

a data set with 2 quantitative variables X^1 and X^2.





$$\hat{f}(x^1, x^2) = \begin{cases} 2 & \text{if } x^2 \leq 2.4 \text{ and } x^1 > 1.7 \\ 1.7 & \text{if } 1.3 < x^2 \leq 2.4 \text{ and } x^1 \leq 1.7 \\ 0.5 & \text{if } x^1 \leq 1.7 \text{ and } x^2 \leq 1.3 \\ -0.5 & \text{if } x^2 > 4.3 \\ 0.3 & \text{if } 2.4 < x^2 \leq 4.3 \end{cases}$$



Assume that the form of the tree is determined.

How to define the response (predicted) associated to each leaf?

let $R(\hat{f}) = \frac{1}{n} \sum_{i=1}^{\hat{}} (\hat{f}(x_i) - f(x_i))^2$

we want to minimize $R(\hat{f})$

Since $R(\hat{f})$ is unknown, we replace $R(\hat{f})$ by $R_n(\hat{f})$ defined by:

$$R_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2.$$

Prop:

Let T denotes a tree.

Let $\widetilde{T}$ denotes the leaves of T.

We have $\hat{f}(x) = \sum_{t \in \widetilde{T}} \bar{y}_t \times 1_{x \in t}$

To minimize $R_n(\hat{f})$, we need to have

$$\overline{y_t} = \frac{1}{n_t} \times \sum_{i \in \{1,\dots,n\} \text{ such that } X_i \in t} y_i$$

with n_t : number of observations from $\mathcal{L}$ that are in t

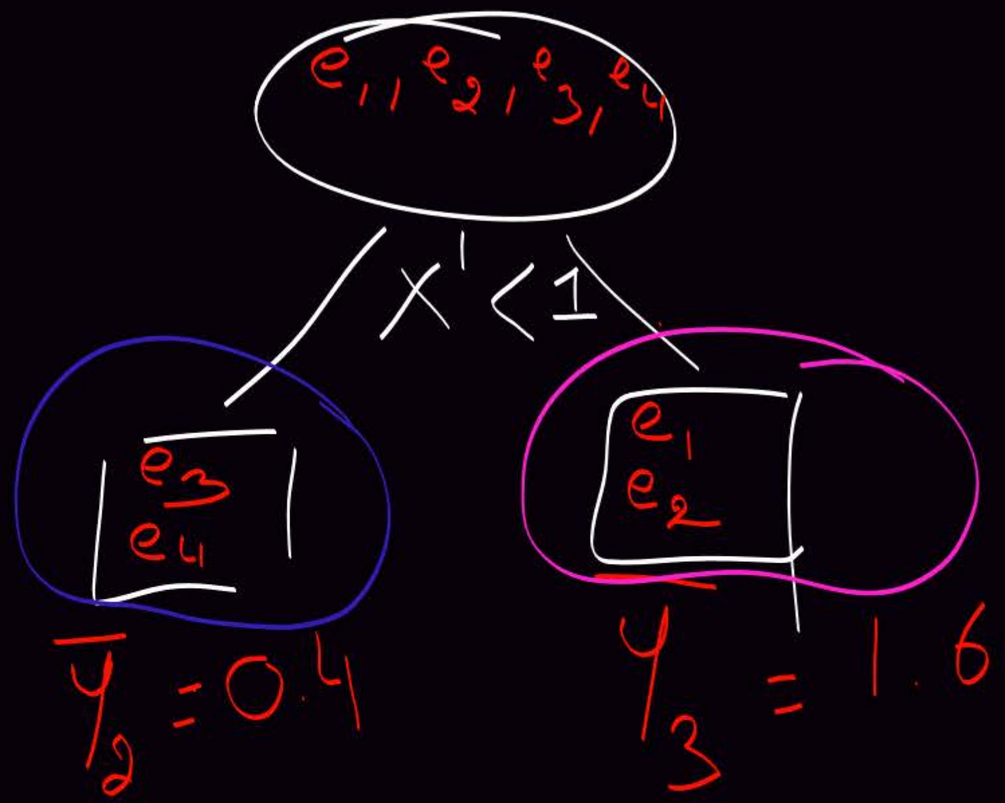
RB:

$$\mathbb{1}_{x \in A} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Example:

dataset

	x^1	x^2	y
e_1	1.4	0.2	1.2
e_2	1.5	-0.2	2
e_3	0.8	1.2	0.5
e_4	0.4	2.1	0.3



Proof:

$$R_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

$$= \frac{1}{n} \sum_{t \in T} \left(\sum_{\substack{i \in \{1,\dots,n\} \\ x_i \in t}} (y_i - \hat{f}(x_i))^2 \right)$$

To minimize $R_n(\hat{f})$, we just have to minimize each term of the sum (because each term is positive and does not involve observations used in the other terms).

$\forall t \in \tilde{T}$, we have to minimize $\sum_{\substack{i \in \{1,\dots,n\} \\ x_i \in t}} (y_i - \hat{p}(x_i))^2$

But

$$\sum_{\substack{i \in \{1,\dots,n\} \\ x_i \in t}} \left(y_i - \hat{f}(x_i)\right)^2 = \sum_{\substack{i \in \{1,\dots,n\} \\ x_i \in t}} \left(y_i - \overline{y}_t\right)^2$$

$$= \sum_{x_i \in t} \left(y_i^2 - 2 y_i \overline{y}_t + \overline{y}_t^2\right)$$

$$= n_t \overline{y}_t^2 - 2\overline{y}_t \sum_{x_i \in t} y_i + \sum_{x_i \in t} y_i^2$$

$$= n_t \left(\bar{y}_t^2 - 2\bar{y}_t \left(\frac{1}{n_t} \sum_{x_i \in t} y_i \right) + \frac{1}{n} \sum_{x_i \in t} y_i^2 \right)$$

$$= n_t \left(\underbrace{\left(\bar{y}_t - \frac{1}{n_t} \sum_{x_i \in t} y_i \right)^2}_{\text{This should be equal to 0 to solve the } p^b \text{ of minimization}} + \frac{1}{n_t} \sum_{x_i \in t} y_i^2 - \frac{1}{n_t^2} \left(\sum y_i \right)^2 \right)$$

$$g(\overline{\gamma_t}) = f(\overline{\gamma_t}) + h(\text{does not depend on } \overline{\gamma_t})$$

To construct the tree:

the criteria is proportional to the variance.

let t denotes a node of T:

$$R(t) = \frac{1}{n} \sum_{x_i \in t} \left(\bar{y}_t - y_i \right)^2$$

Let $t_{L,s}$ and $t_{R,s}$ two children of the node t, children generated by a binary question s.

The set of all possible binary question is S

$$S = \bigcup_{i \in \{1,\dots,p\}} \left\{ \begin{array}{l} (X^i \leq a_i,\ a_i \in \{s_i^{(1)},\dots,s_i^{(n-1)}\}) \\ \text{or } (X^i > a_i,\ a_i \in \{s_i^{(1)},\dots,s_i^{(n-1)}\}) \\ \text{or } (X^i \in \mathcal{G}^i,\ \mathcal{G}^i \subset \{1,\dots,I_i\}) \end{array} \right\}$$

where $\{1,\dots,I_i\}$: the set of modalities for X^i when X^i is a qualitative variable

$x_1^i, \ldots, x_n^i$ are the observations of the variable X^i.

let $x_{(1)}^i \leq x_{(2)}^i \leq \ldots \leq x_{(n)}^i$ the same observations written in ascending order.

$$s_i^{(k)} = \frac{1}{2}\left(x_{(k)}^i + x_{(k+1)}^i\right)$$

Rk: Let X^i be a quantitative variable.

At the maximum, we have $2*(n-1)$ possible divisions involving X^i

Let X^i be a qualitative variable.

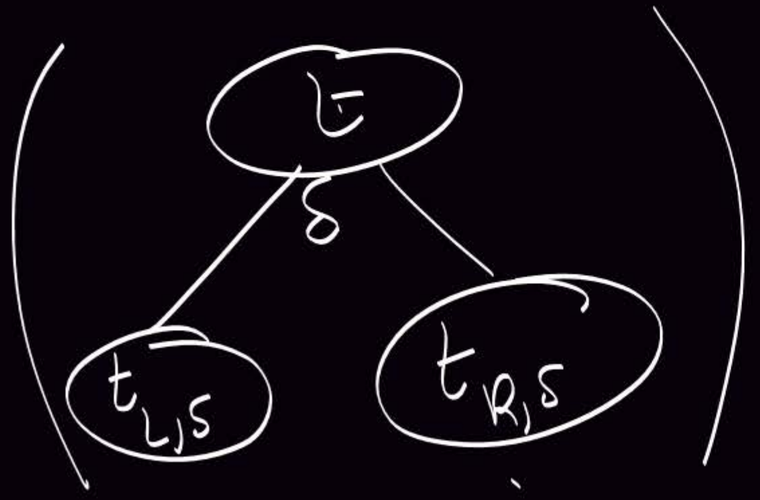
We have $2^{I_i} - 2$ possible divisions

(I_i being the number of modalities of X^i)

Let t a node that we want to split.

For each possible division s of t

we compute $\Delta R(t,s) = R(t) - (R(t_{L,s}) + R(t_{R,s}))$



The final division $\delta^*(t)$, it means the division that appears in the tree to split the node t, is defined by:

$$\delta^*(t) = \underset{\delta \in S}{\operatorname{argmax}} \; \Delta R(t, \delta)$$

$\Leftrightarrow S^*(t)$ is a division of S that maximizes $\Delta IR(t, S)$

How to say that a node is a leaf?

Stopping rule: a node is a leaf if:

- the node contains just one observation

or

- all the observations of the node have the same value for Y.

At the end, we have a maximal tree!

Rq: the error, computed on the training sampling, for a maximal tree is 0!

Example:

dataset

	x^1	x^2	y
e_1	1.4	0.2	1.2
e_2	1.5	-0.2	2
e_3	0.8	1.2	0.5
e_4	0.4	2.4	0.3

$$S = \{ (x^1 \leq 0.6), (x^1 \leq 1.1), (x^1 \leq 1.45), (x^1 > 0.8), (x^1 > 1.1), (x^1 > 1.45), (x^2 \leq 0), (x^2 \leq 0.7), (x^2 \leq 1.8), (x^2 > 0), (x^2 > 0.7), (x^2 > 1.8) \}$$

$\Delta R(t_1, s_1)$

$\Delta R(t_1, s_2)$

$\Delta R(t_1, s_3)$

$\vdots$

$\Delta R(t_1, s_{12})$

maximal value the best division is $(x' \leq 1.45)$

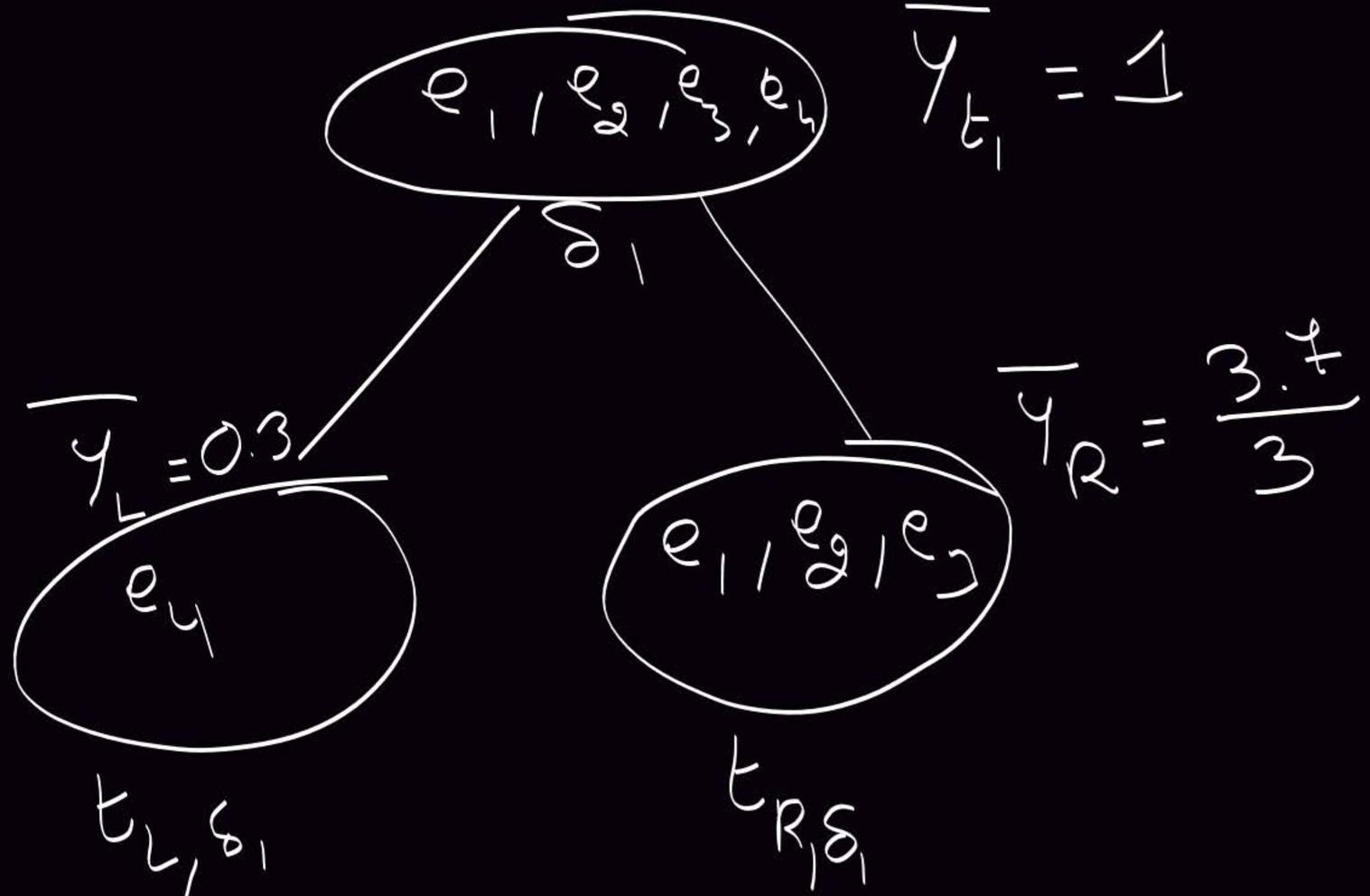
$S = \{s_1 : (x' \leq 0.6),\ s_2 : (x' \leq 1.1),\ s_3 : (x' \leq 1.45),\ s_4 : (x' > 0.6),\ s_5 : (x' > 1.1),\ s_6 : (x' > 1.45),\ s_7 : (x^2 \leq 0),\ s_8 : (x^2 \leq 0.7),\ s_9 : (x^2 \leq 1.8),\ s_{10} : (x^2 > 0),\ s_{11} : (x^2 > 0.7),\ s_{12} : (x^2 > 1.8)\}$

Example:

dataset

	x^1	x^2	y
e_1	1.4	0.2	1.2
e_2	1.5	-0.2	2
e_3	0.8	1.2	0.5
e_4	0.4	2.1	0.3

$\delta_1 = (x^1 \leq 0.6)$



Example:

data set

	x^1	x^2	y
e_1	1.4	0.2	1.2
e_2	1.5	-0.2	2
e_3	0.8	1.2	0.5
e_4	0.4	2.1	0.3

$$\delta_1 = (x^1 \leq 0.6) \qquad \overline{y}_{t_1} = 1$$

$$(e_4)\ \frac{y_L}{y_R} = \frac{0.3}{\frac{3.7}{3}}$$

$$R(t_1) = \frac{1}{4}\Big((1-1.2)^2 + (1-2)^2 + (1-0.5)^2 + (1-0.3)^2\Big)$$

$$R(t_L) = \frac{1}{4}\Big((0.3-0.3)^2\Big) = 0$$

$$R(t_R) = \frac{1}{4}\Big(\Big(1.2-\frac{3.7}{3}\Big)^2 + \Big(2-\frac{3.7}{3}\Big)^2 + \Big(0.5-\frac{3.7}{3}\Big)^2\Big)$$

What happends in the classification framework?

Let $\{1, \dots, I\}$ denotes the set of modalities of Y.

$\forall i \in \{1, \dots, I\}$, $n_i = \#\{j \in \{1, \dots, n\} \text{ s.t. } Y_j = i\}$

$\pi_i = \frac{n_i}{n}$: prior probability of class i

Let t be a node of a tree T

$$n_t = \#\{i \in \{1, \dots, n\} \text{ s.t. } x_i \in t\}$$

$$i \in \{1, \dots, I\}, \quad n_{i,t} = \#\{j \in \{1, \dots, n\} \text{ s.t. } x_j \in t \text{ and } y_j = i\}$$

we estimate :

- the probability that an observation is in t and of class i by :

$$p(t,i) = \pi_i \times \frac{n_{i,t}}{n_i}$$

- the probability that an observation is in t by :

$$p(t) = \sum_{i=1}^{I} p(t,i)$$

- the posterior probability of class i in node t by:

$$p(i|t) = \frac{p(i,t)}{p(t)}$$

definition 1:

We say that a function h is an heterogeneity function on $\{(p_1, \dots, p_I) \in \mathbb{R}^I$ s.t $\forall i \in \{1, \dots, I\}, p_i \geq 0$ and $\sum p_i = 1\}$ if:

- h is symetric w.r.t $(p_1, \dots, p_I)$
- h is maximal at the point $\left(\frac{1}{I}, \dots, \frac{1}{I}\right)$
- h is minimal at the points $(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$

definition 2

Let t be a node.

Let h be an heterogeneity function.

Let define the imp function by:

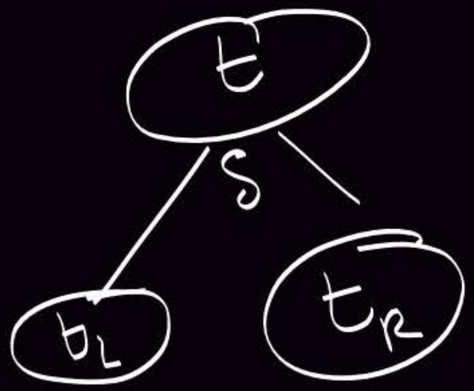
$$imp(t) = h\left(p(1|t), p(2|t), \ldots, p(I|t)\right)$$

proposition :

If the heterogeneity function h is concave, then we have

$$\Delta imp(t, s) = imp(t) - p_R\, imp(t_R) - p_L\, imp(t_L) \geq 0$$

where



and

$$p_R = p(t_R)/p(t)$$

The way to construct the maximal tree in the classification framework is similar to the regression framework except two things:

- the criterion $\Delta R(t, s)$ is replaced by $\Delta imp(t, s)$

- $\overline{y}_t$ are replaced by $j(t)$ that are defined by

$$j(t) = \underset{i \in \{1,\dots,I\}}{\operatorname{argmax}}\ p(i|t)$$

Rk: When in the computer, we do not define some prior probabilities, the computer use

the prior probabilities Π_i.

In this case, $j(t)$ is also the class which is the most represented in t.

If you have others prior probabilities, be careful this other definition is not always true!

RR : classification

Learning error : $\frac{1}{n}\sum_{i=1}^{\hat{n}} 1\!\!1_{y_i \neq \hat{y}}$

```
1/nrow(X) * sum(y != predict(T, type = 'class'))
```

Exercise

Create a matrix X where:

- the first column contains 1000 observations from $\mathcal{N}(2, 0.1)$
- the second one is associated to $\mathcal{E}(0.2)$
- the third one is associated to a $\mathcal{B}(5, 0.7)$
- the variable 4 ——— $\mathcal{N}(-2, 0.05)$

- The variable 5 is associated to $\mathcal{B}(4, 0.2)$
- 6 ——— to $\mathcal{N}(-1, 0.1)$
- 7 ——— $\mathcal{N}(1, 0.1)$
- 8 ——— $\mathcal{B}(0.5)$

$$Y = \begin{cases} 2 - 5 * V5 + 4 * V2 + \varepsilon & \text{if } V8 = 0 \\ -3 + 2 * V1 - 7 * V3 + \varepsilon & \text{if } V8 = 1 \end{cases}$$

with $\varepsilon \sim \mathcal{N}(0, 1)$

Q1 : run the CART algorithm
to explain Y thanks to all
the variables.
look at the variable importance
object.

Q2 : create Y1 which is defined by :
$Y1 = 1$ if $Y < Q_1$ (1st quartile of Y)
$Y1 = 2$ if $Q_1 \leq Y < Q_2$

$y1 = 3$ if $Q_2 \leq y < Q_3$

$y1 = 4$ if $y \geq Q_3$

Run the CART algorithm to explain Y1 thanks to V1 until V8.

$\underline{Q_3}$: Repeat Q_1 and Q_2 but you transform V3 and V5 into qualitative variables.

Q4: Repeat just Q1 with the initial variables, but create 5 bootstrap samples (with replacement)