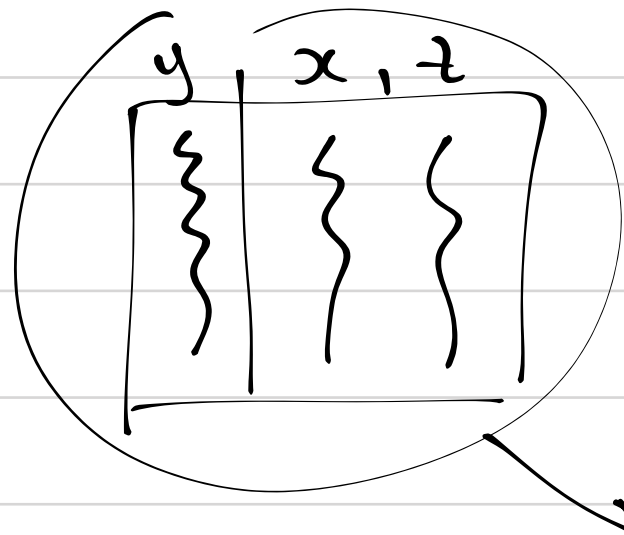
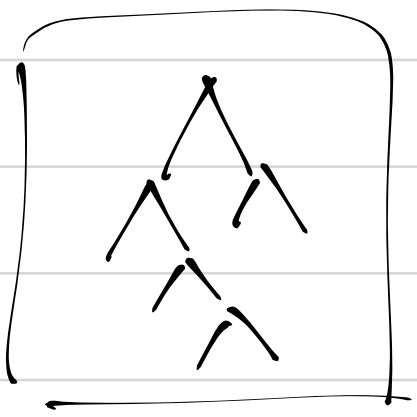
Hi :)

trees!

→ classification + regression.

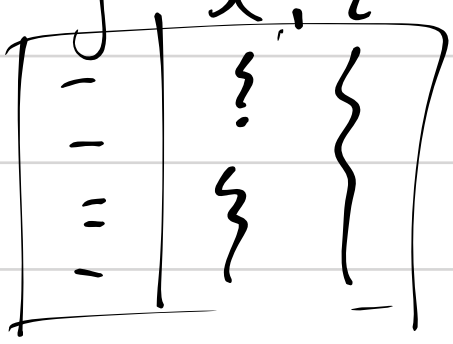
Step 1:
Build the tree



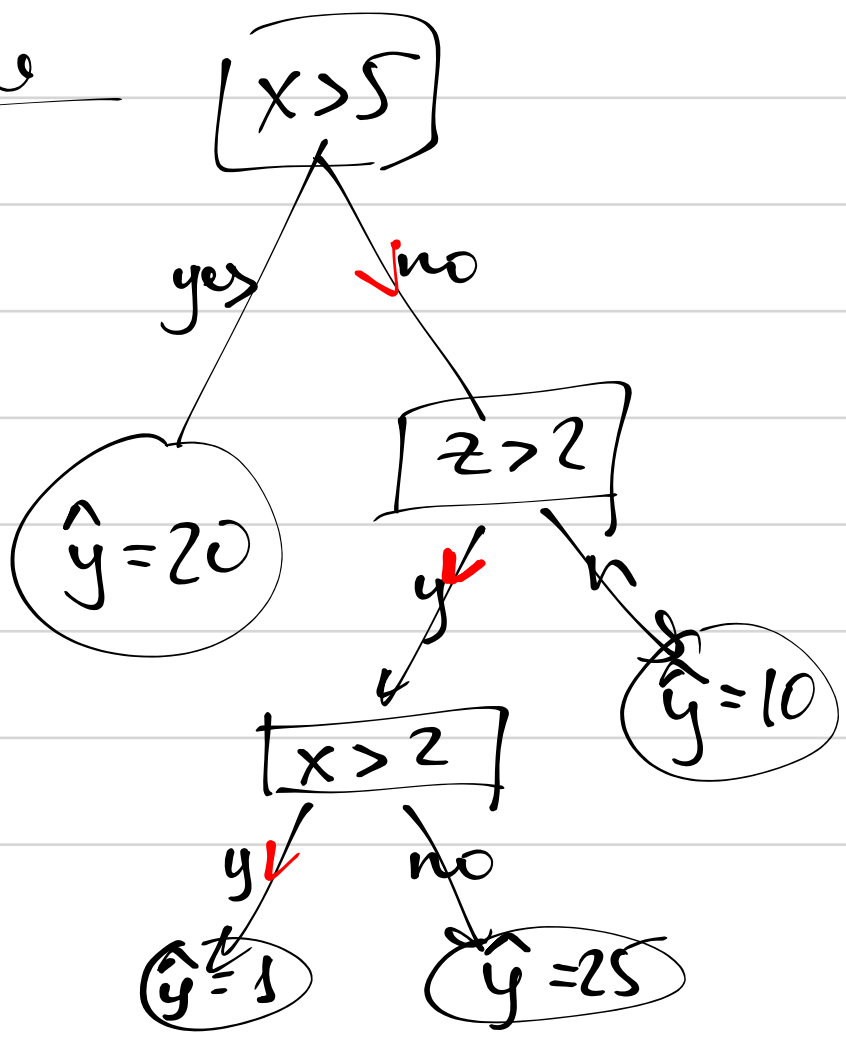


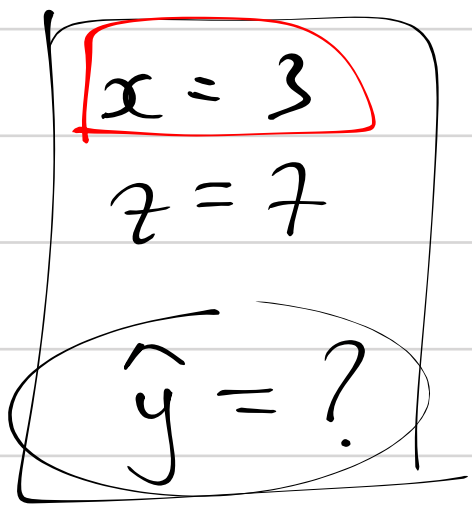
y - response variable
x, z [...] - predictors

Step 2 Use the tree to forecast y (the response variable)



Example





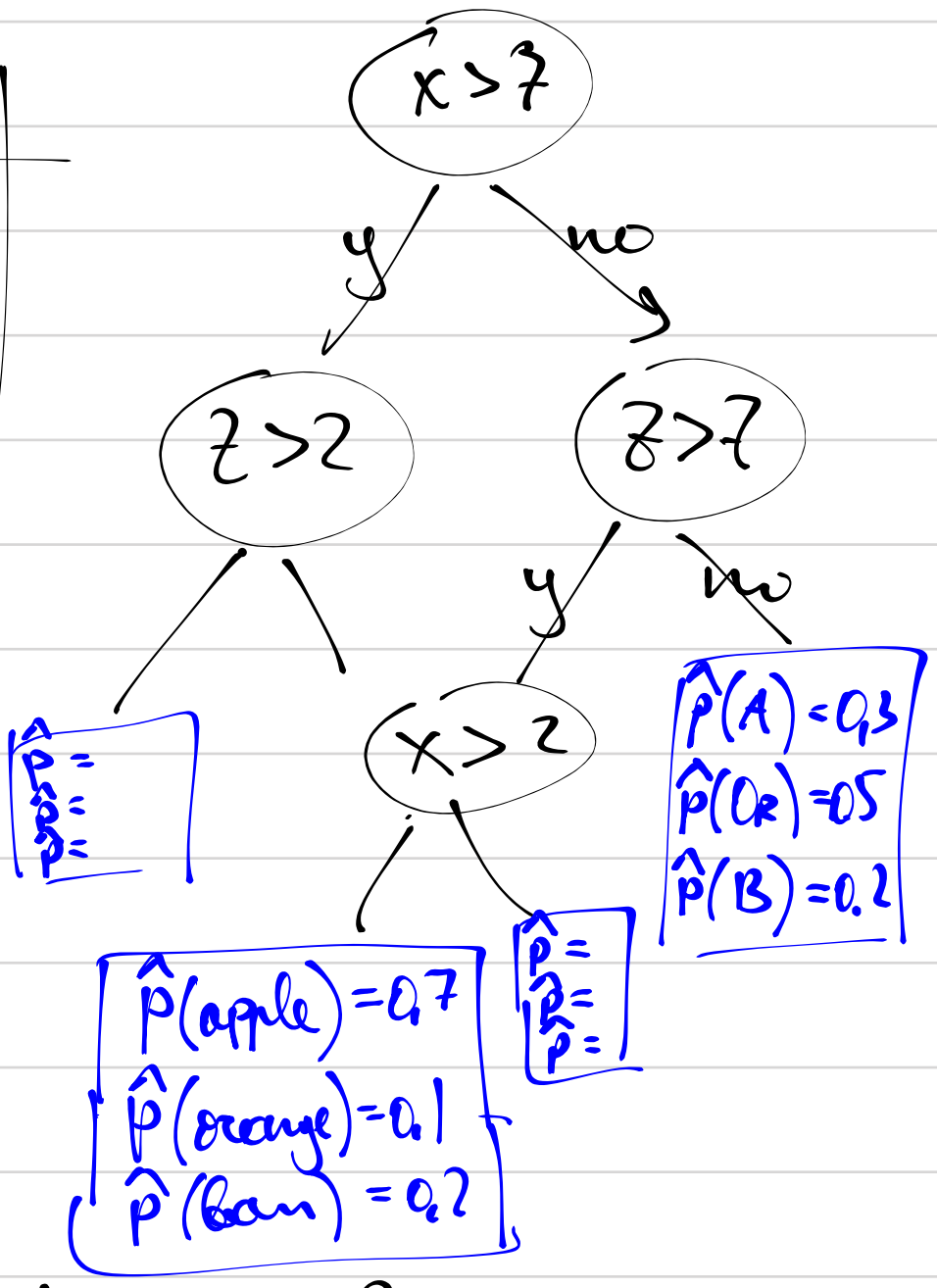
ŷ = 1

Step 1. How to construct a tree?

Classification task. response variable y takes finitely many values (non-numerical values)

response: predictors.

y	x	z
apple	20	5
orange	10	1
ban	20	3
ban	10	4
apple	-5	5



! there are many variants of the procedure!

Q1. How the node is split in two nodes? Which function is optimized?

Q2. When we should stop the splitting process?

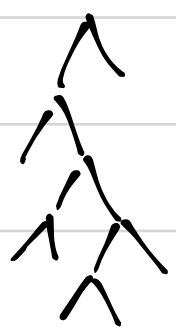
Q3. How the tree processes missing values of predictors?

y	x	z
orange	20	–

Q4. Should we cut the tree after construction? And how?

1. grow a tree

2. cut the tree.



Q1. How to split a node in two nodes?

two criteria: entropy, Gini impurity index

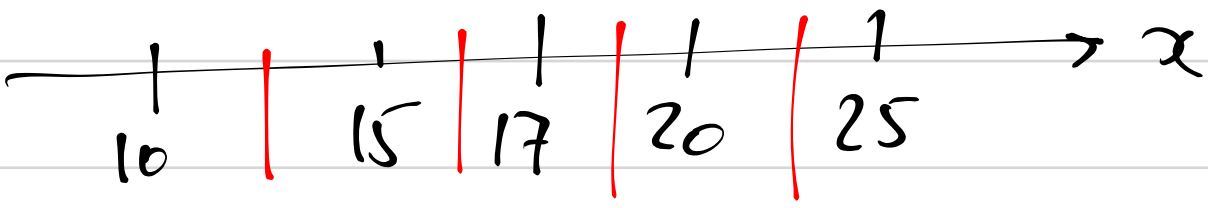
ask a question → predictor > south.

$x>7$ $z>3$ $x>2.5$

$z>1.7$ . . . .

y	x	z
A	20	1
A	10	2
B	15	3
B	17	4
B	20	100
B	25	200

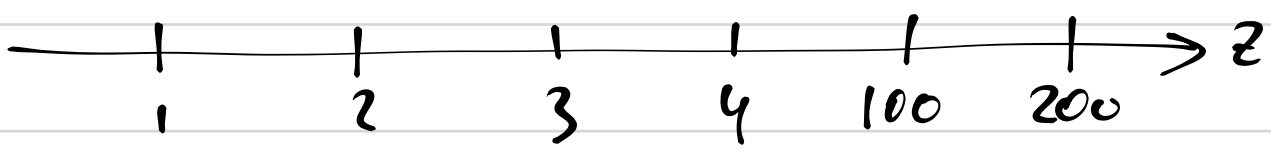
$$H(y) = -\left[\frac{2}{6}\cdot\ln\frac{2}{6}+\frac{4}{6}\cdot\ln\frac{4}{6}\right]$$



Possible questions are

$x>12.5$ $x>16$

$x>18.5$ $x>22.5$



$z>1.5$ $z>2.5$

$z>3.5$ $z>52$

$z>150$

Calculate all possible conditional entropies $H(y|\varphi)$

y	x	z
A	20	1
A	10	2
B	15	3
B	17	4
B	20	100
B	25	200

$Q = \{z > 52\}$

$H(y|Q) = ?$

$z > 52$

yes / no

y	x	z
B	20	100
B	25	200

y	x	z
A	20	1
A	10	2
B	15	3
B	17	4

$H(Y|Q=yes) = -1 \cdot \ln 1 = 0$

$H(Y|Q=NO) = -\left[\frac{2}{4}\cdot \ln\frac{2}{4} + \frac{2}{4}\cdot \ln\frac{2}{4}\right] = -\ln\frac{1}{2} \approx 0.69.$

$\underline{H(Y|Q)} = \frac{2}{6}\cdot 0 + \frac{4}{6}\cdot \ln 2 = \frac{2}{3}\cdot \ln 2 \approx 0.46.$

Q, Possible questions	H(Y|Q)
$Q = \{z > 52\}$	0.46
$\{x > 12.5\}$	⋮
$\{z > 1.5\}$	*
⋮	

9 possible questions

← lowest cond-l entropy is chosen.

$$\min_Q H(Y|Q) \iff \max_Q \left[H(Y) - H(Y|Q)\right]$$

Second criterion: Jini impurity index

y
A
A
B
B
B
B

$J(Y)$ = the probability that two players [different] will choose values of Y if they choose indep- by one observation each

$$J(Y) = P(\text{1 ch A}, \text{2 ch B}) + P(\text{1 ch B}, \text{2 ch A}) =$$

$$= \frac{2}{6}\cdot\frac{4}{6} + \frac{4}{6}\cdot\frac{2}{6} = 2\cdot\frac{2}{6}\cdot\frac{4}{6} =$$

$$= 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$$

Q possible quest-s	$J(Y\|Q)$
$\{z > 52\}$	$\frac{1}{3}$
⋮	⋮

the question with lowest $J(Y|Q)$ is chosen.

$$\min_Q J(Y|Q)$$
$$\max_Q [J(Y) - J(Y|Q)]$$

$z > 52$

yes

y	x	z
B	20	100
B	25	200

$J(Y|Q=\text{yes}) = 0$

no

y	x	z
A	20	1
A	10	2
B	15	3
B	17	4

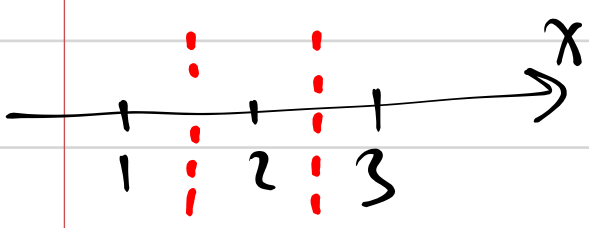
$J(Y|Q=\text{no}) = \frac{2}{4}\cdot\frac{2}{4} + \frac{2}{4}\cdot\frac{2}{4} = \frac{1}{2}$

$$J(Y|Q) = \frac{n_y}{n}\cdot J(Y|Q=\text{yes}) + \frac{n_{no}}{n}\cdot J(Y|Q=\text{no}) =$$
$$= \frac{2}{6}\cdot 0 + \frac{4}{6}\cdot\frac{1}{2} = \frac{1}{3}$$

For regression problem

Residual sum of squares → min.
Q

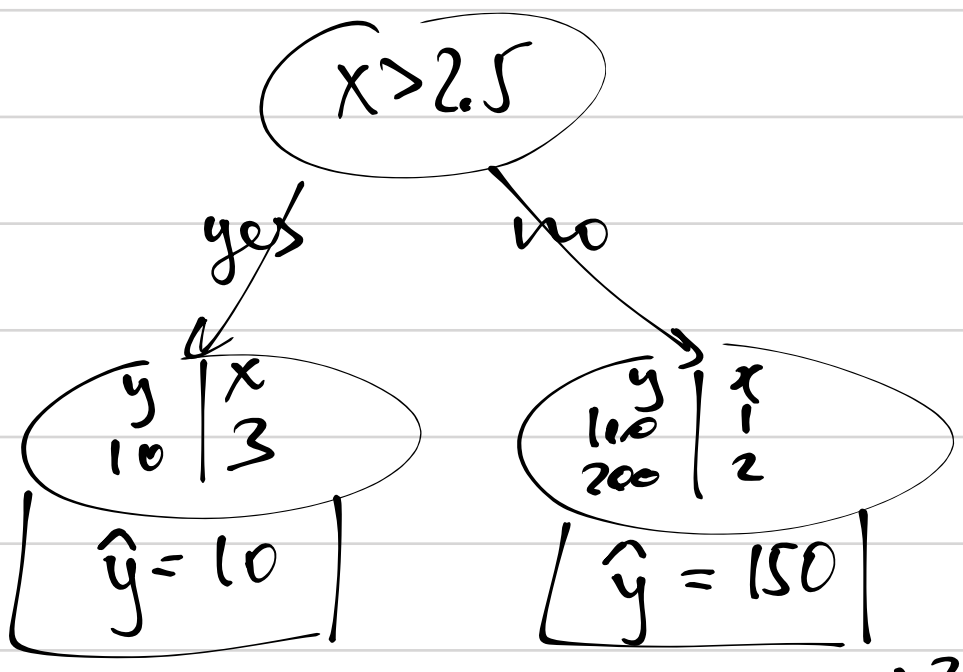
y	x
100	1
200	2
10	3



two possible questions.

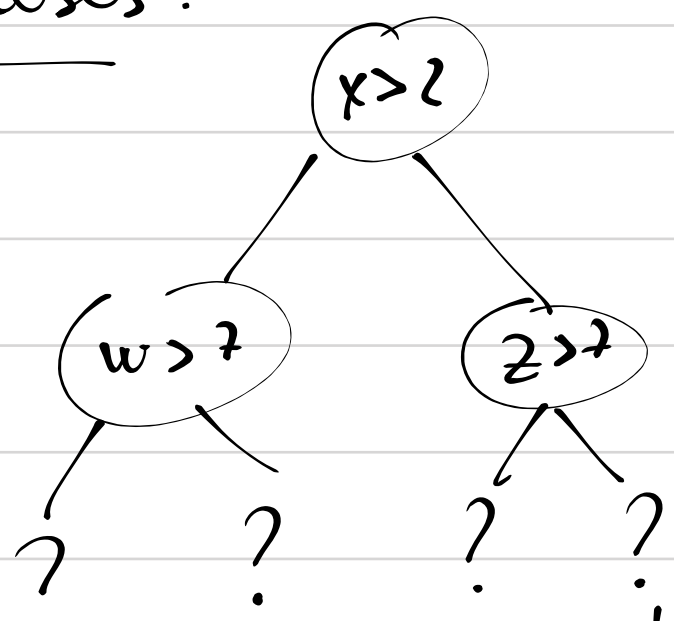
Q	ResSS
$x > 1.5$	
$x > 2.5$	?

5000



$$\text{ResSS} = \sum_i (y_i - \hat{y}_i)^2 = (10-10)^2 + (100-150)^2 + (200-150)^2 = 2 \cdot 50^2 = 5000$$

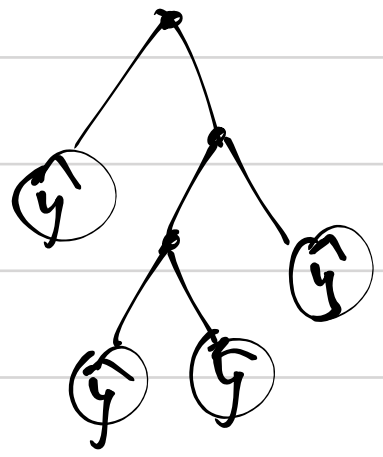
Q16. Where we should split if there are many cases?



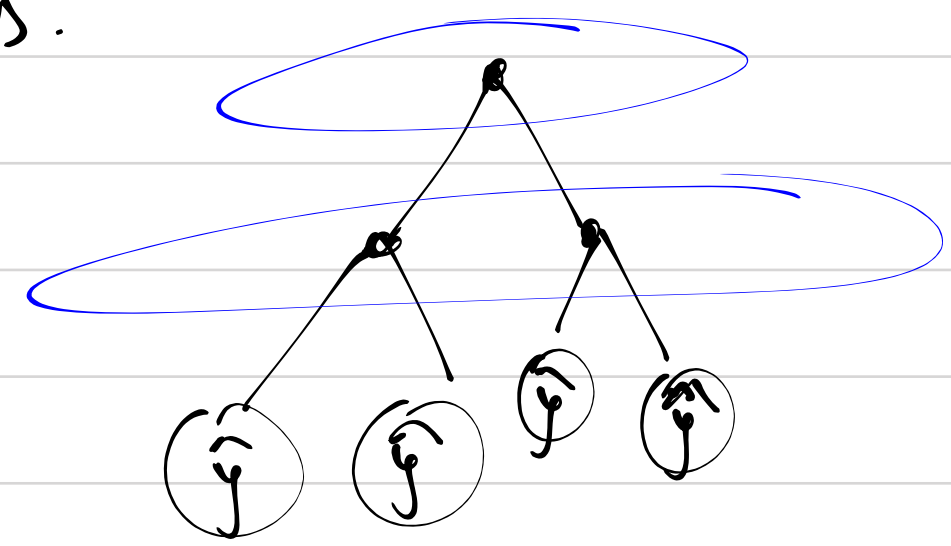
Consider all possible split places and choose the place where the entropy drop is max!

Q2. Where should we stop ?

Possible answers : → Stop after (3) splits.



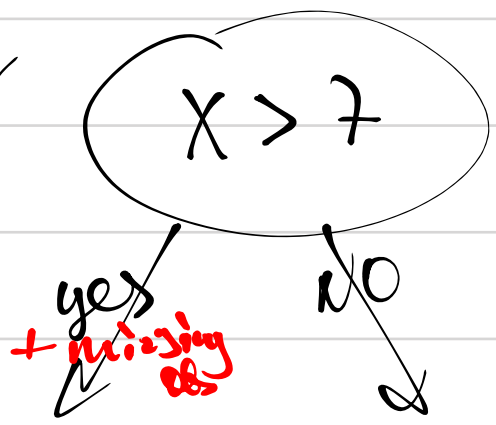
→ Stop when the tree has 2 levels of questions.



→ Split the node if there more than (5) observations inside.

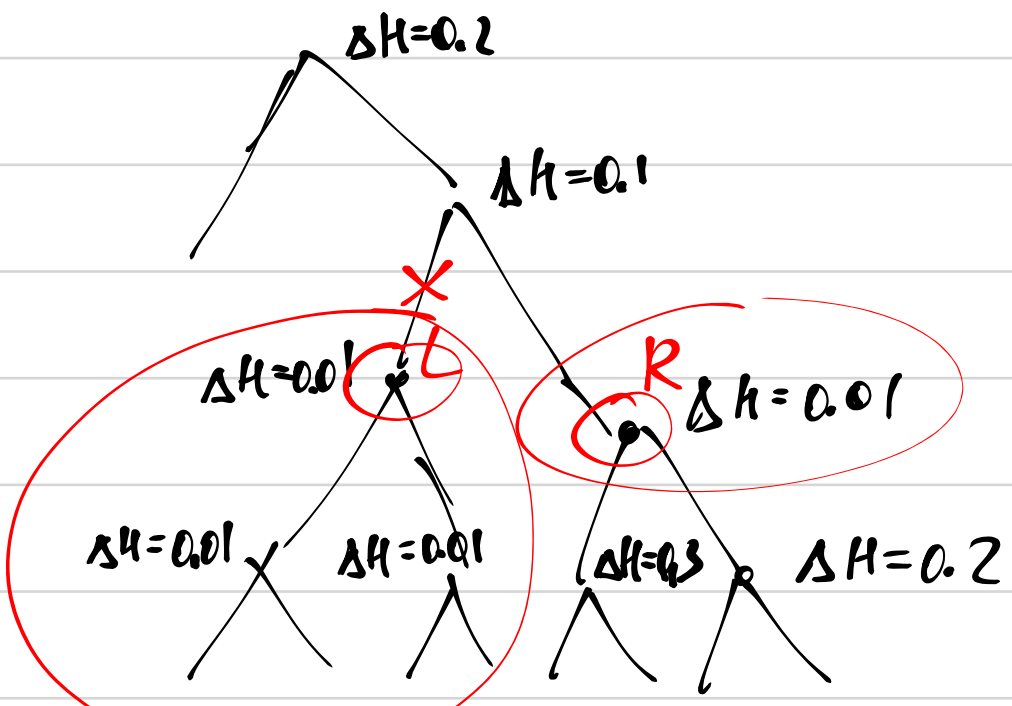
Q3. What to do with missing values in predictors?

A_1 : Choose the best side (Yes / No)
A_2 : choose (Y/No) side randomly.
A_3 : Remove missing obs-s.



Q4. Should we cut the tree? If yes then how?

Stage one.

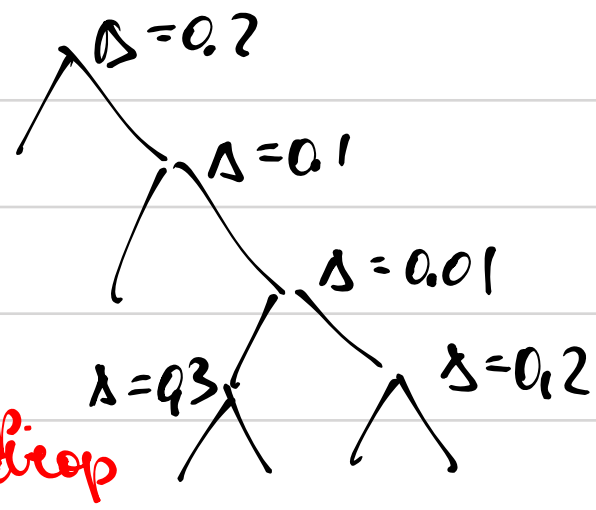


Stage two.

Cut "bad" branches.

price of quest = 0.05

cut the bad part of the tree



classif: entropy drop (max) / gini imp. index drop (max)

regression: RessS drop (max)

- Q1A : How to split a node in two?
- Q1B : Which node should we split?
- Q2 : When should we stop?
- Q3 : How to process missing values?
- Q4. : How to cut the tree?

→ split the node if it contains more than 5 obs.

→ see answ to Q1A / Q1B

→ compare the entropy drop with a branch price.