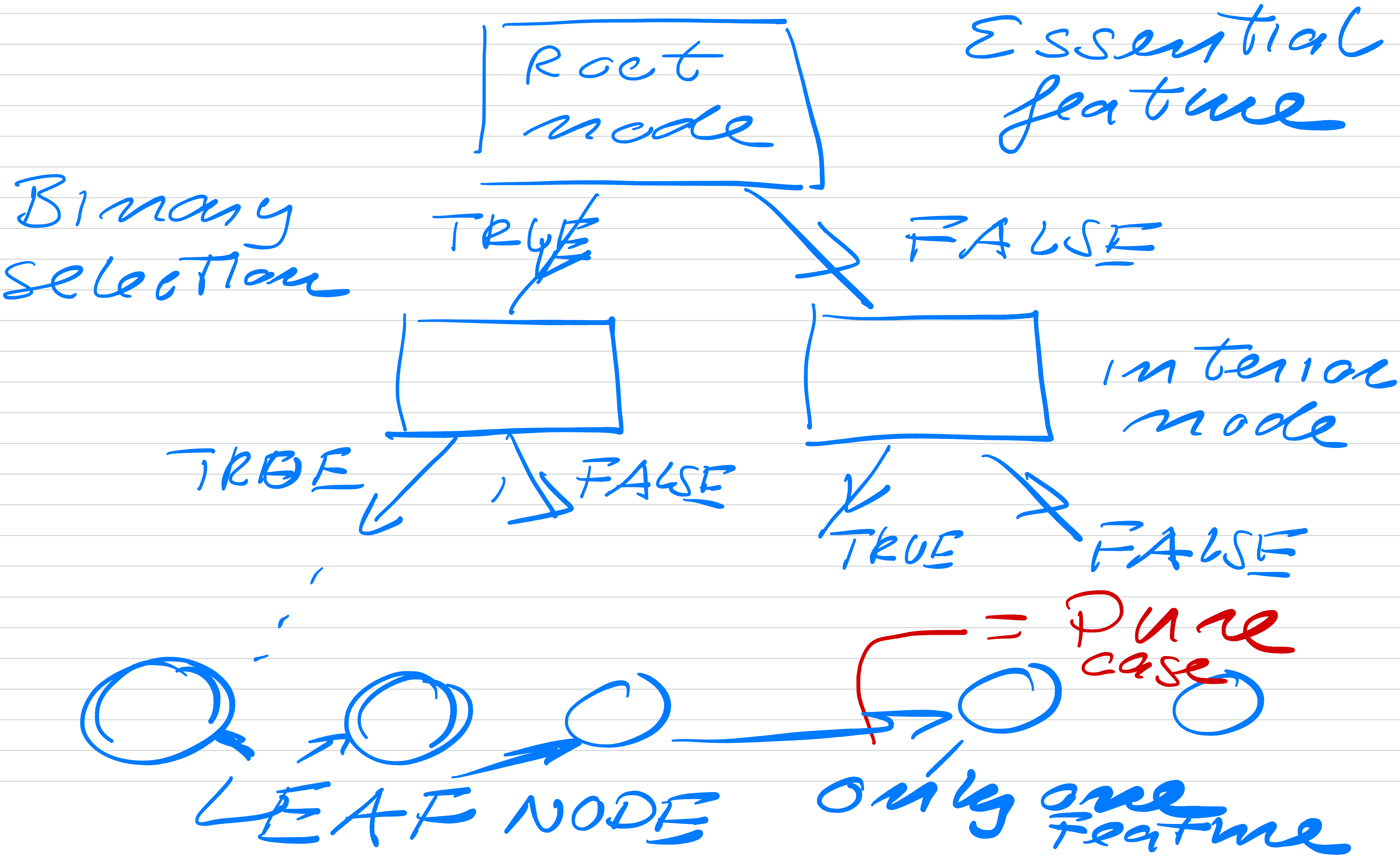
Lecture FYS-STK3155/4155, November 11, 2024

FYS-STK3155/4155
november 11

Why Decision trees?

- simple
- building block in Ensemble methods
  - Bagging
  - Random Forest
  - Boosting
  - Gradient Boosting

A simple tree



what is an impurity?
when all elements of a branch (= leaf) contain only one class, we say we have a **pure case**

How can we define a measure for when a class is pure?

Two ways to measure:

gini factor g

entropy S

$g, S \in [0, 1]$

$g = 0$, pure case

$S = 0$ — " —

Example: binary case

True has probability $= p$

$p \in [0, 1]$

False : $1-P$

$$\sum_{i=1}^{2} P_i = (1-p) + P = 1$$

Criterion for g or S

Measure of information

- additive
- continuous
- non-negative
- high for unlikely events

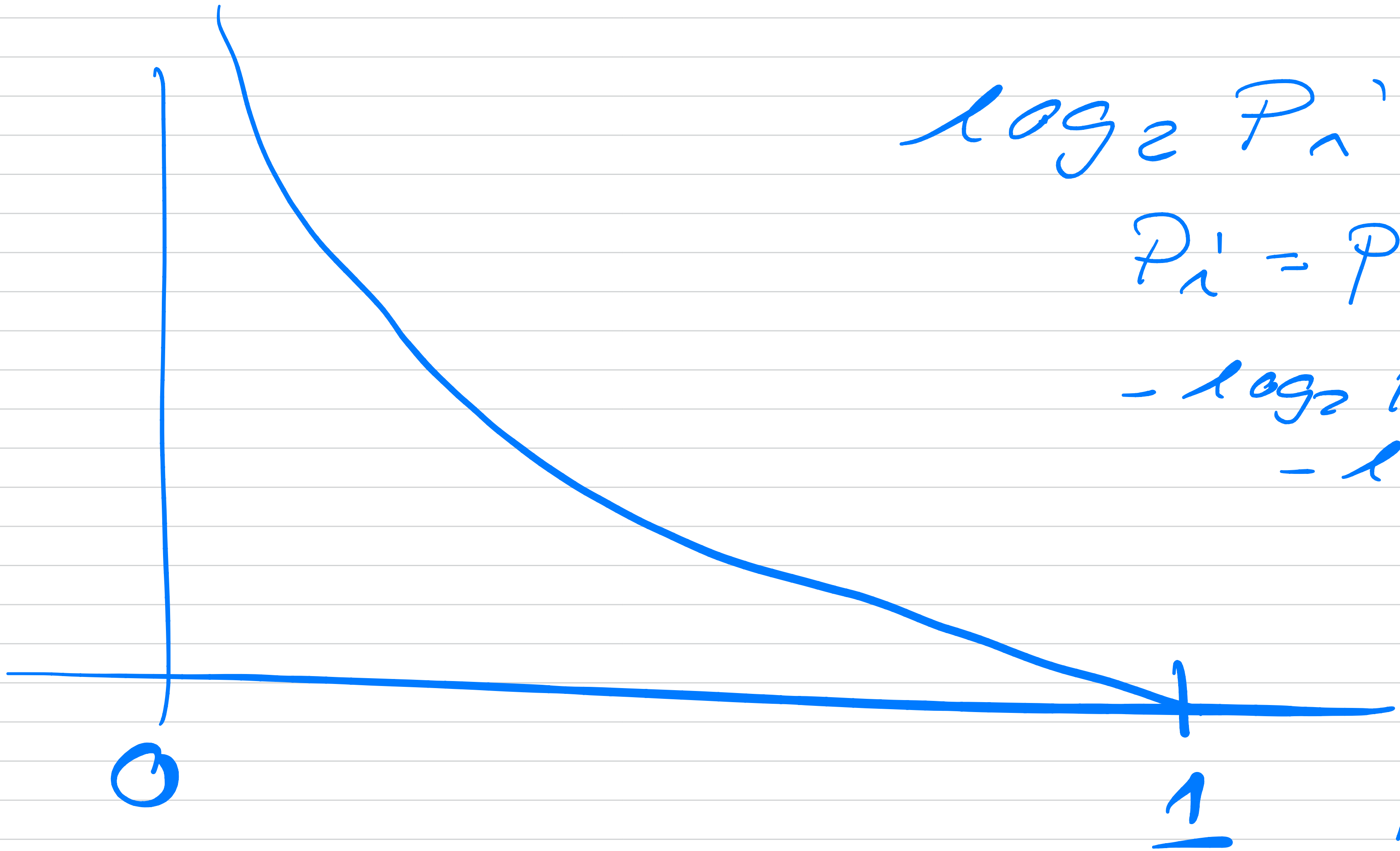
entropy :

$$S_i = - P_i \log_2 P_i$$

$$\log_2 P_i$$

$$P_i = p(x_1)\, p(x_2)$$

$$-\log_2 p(x_1)$$
$$-\log_2 p(x_2)$$



$$S = -\sum_{i=1}^{2} p_i \log_2 p_i$$

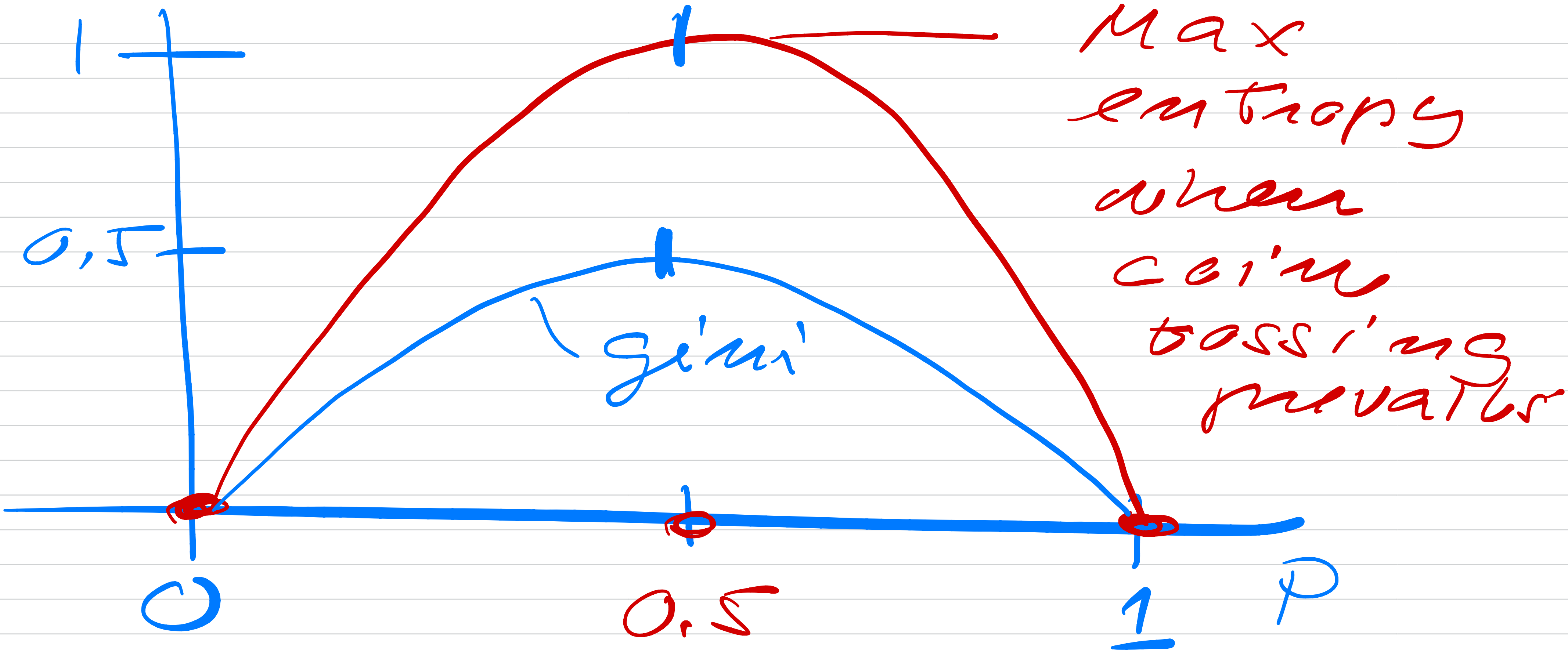
$$(0 \log_2 0 = 0)$$

$$-\underbrace{p}_{\text{True}} \log_2 p - \underbrace{(1-p)}_{\text{False}} \log_2 (1-p)$$

$S = 0$ if $p = 0$ FALSE (pure)

$S = 0$ if $p = 1$ TRUE (pure)

$p = 0{,}5 \Rightarrow S = 1$


1
0,5
0
0,5
1
p
gini
Max entropy when coin tossing prevails

p = 0.5, max entropy is the probability of an imperfect classification. Both cases are equally probable.

Gini factor

$$g = \sum_{i=1}^{M} p_i (1 - p_i)$$

$$= \sum_i p_i - \sum_i p_i^2 =$$

$= 1$

$$1 - \sum_{i=1}^{M} p_i^2 \qquad g = [0, 1]$$

For our simple example

$$g = 1 - \sum_{i=1}^{2} p_i^2$$

$$= 1 - p^2 - (1-p)^2 = 2p(1-p)$$

$p = 0 \Rightarrow g = 0$

$p = 1 \Rightarrow g = 0$

$p = 0.5 \Rightarrow g = 0.5$

Classification Example

- Features/Attributes
  - Grade trend (above/below)
  - # of hours studied (high/low)
  - # of hours slept (high/low)

- output
  grade above or below average

Trend	Study	Sleep	Grade
Above	Low	H	A
Below	High	L	B
A	L	H	A
A	H	H	A
B	L	H	B
A	L	L	B
B	H	H	B
B	L	H	B
A	L	L	B
A	H	H	A

Gini factor in order to make a tree

$P(\text{Trend} = A) = 6/10$

$P(\text{Trend} = B) = 4/10$

if (Trend = above & grade = below)

$P = 2/6 = 1/3$

if (Trend = above & grade Above)

$P = 4/6$

$$\text{gini} = 1 - \left(\left(\frac{4}{6}\right)^2 + \left(2/6\right)^2\right) = 0.45$$

if (Trend = below & grade above)

$$P = 0$$

if (Trend = below & grade below)

$$P = 4/4 = 1$$

gini index $1 - (1^2 + 0^2) = 0$

weighted sum for Trend:

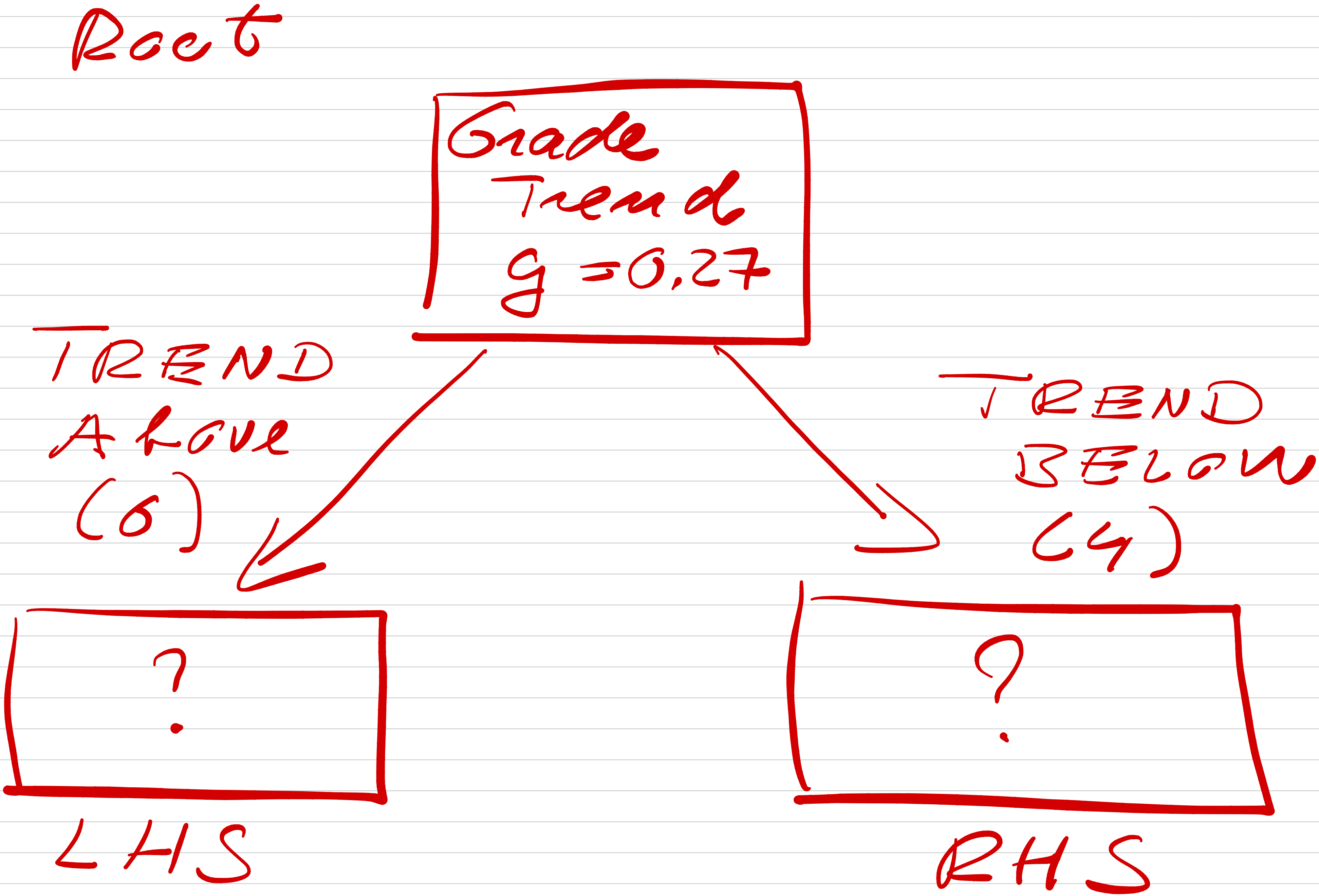
$$6/10 \times 0.45 + 4/10 \cdot 0 = \boxed{0.27}$$

gini factor for hours slept

$g = \boxed{0.47}$ (weighted)

weighted factor for hours studied

$g = \boxed{0.34}$


Root
Grade Trend
g = 0,27
TREND ABOVE (6)
TREND BELOW (4)
?
?
LHS
RHS

LHS

T	sleep	study	G
A	L	H	A
A	L	H	A
A	H	H	A
A	L	L	B
A	L	L	B
A	H	H	A

$P(\text{sleep} = H) = 2/6$

$P(\text{SLEEP} = L) = 4/6$

if (H & grade above)

$P = 1$

if (H & grade below)

$P = 0$

⋮

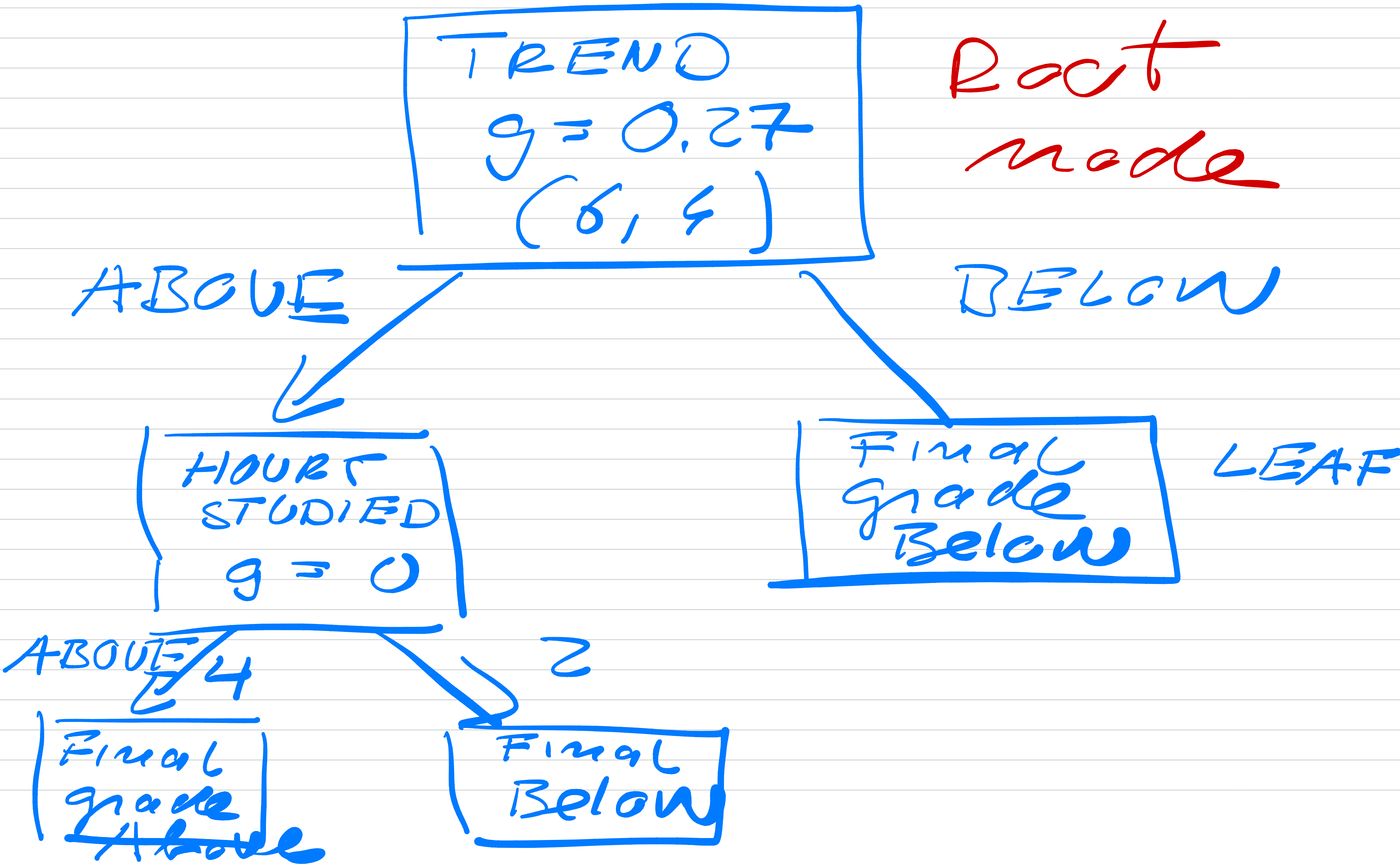
weighted gini for hours slept $= 0.33$

weighted gini for hours studied $g = 0$

RHS

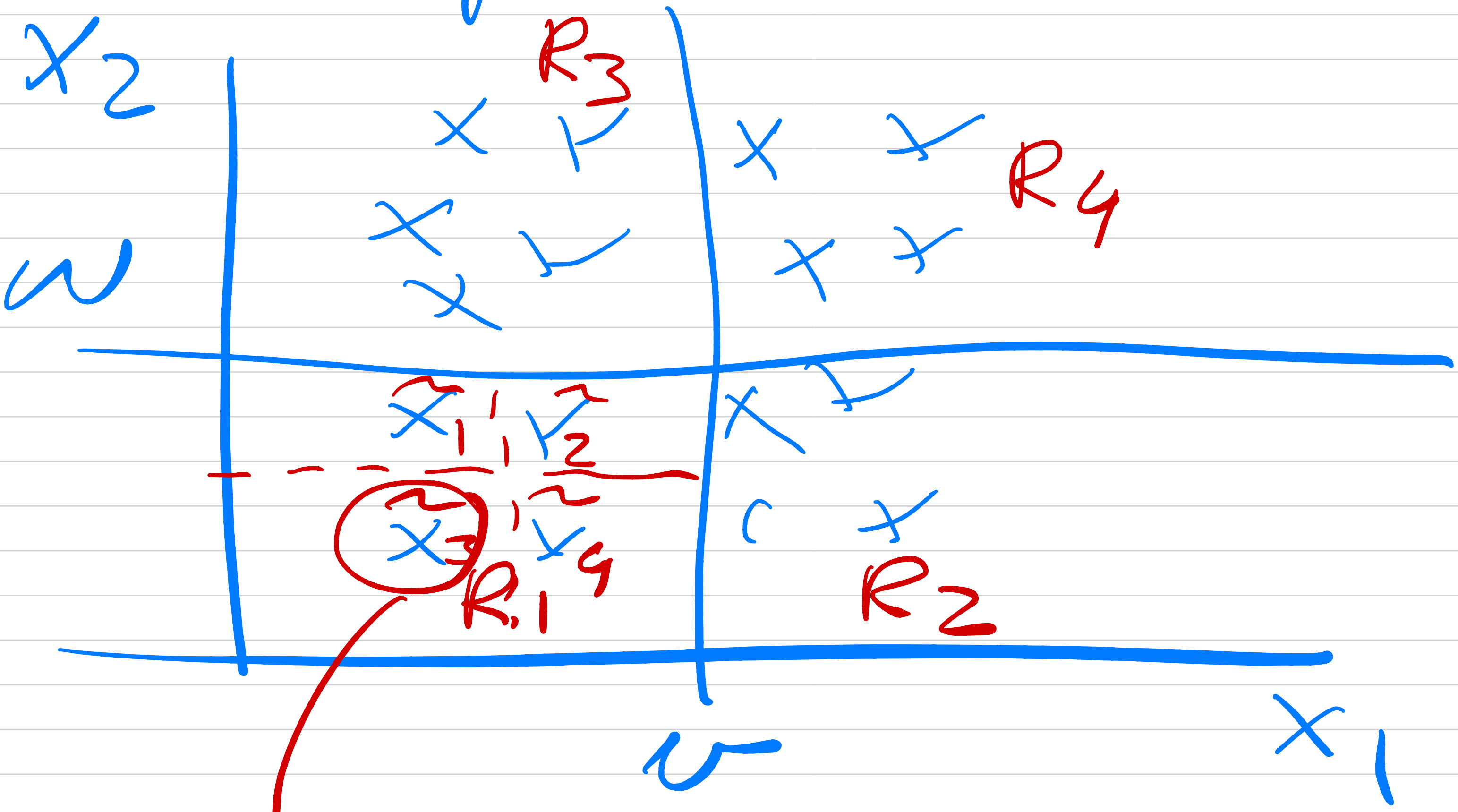
T	Sleep	Study	Grade
B	H	L	B
B	L	H	B
B	H	H	B
B	L	H	B

TREE

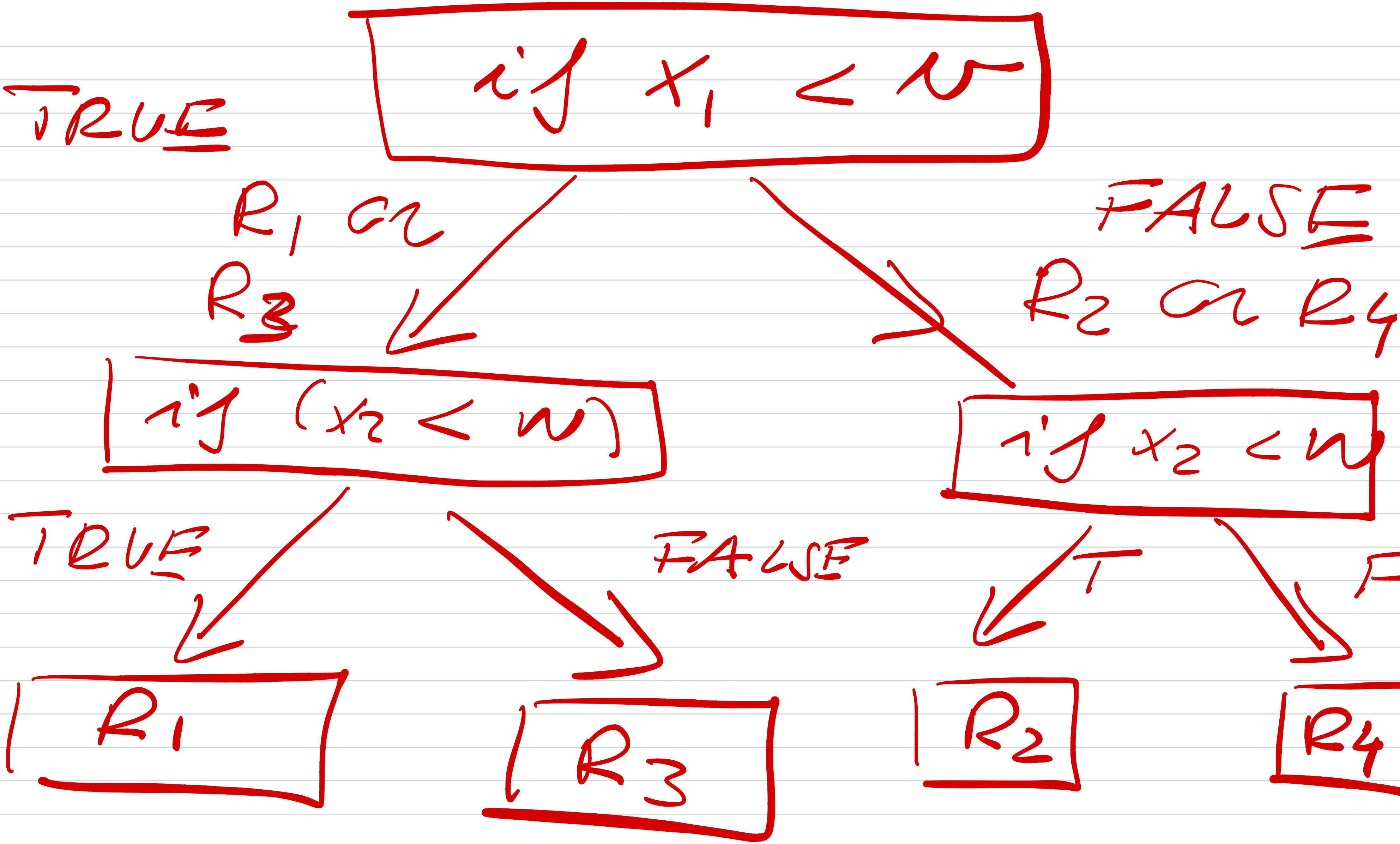


TREND
g = 0.27
(6, 4)
Root node
ABOVE
BELOW
Final grade Below
LEAF
HOURS STUDIED
g = 0
ABOVE 4
2
Final grade Above
Final Below

Regression Tree



$$\tilde{x}_{R_1} = \frac{1}{4}(\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4)$$



Regression prediction =
average value in a Region R_i