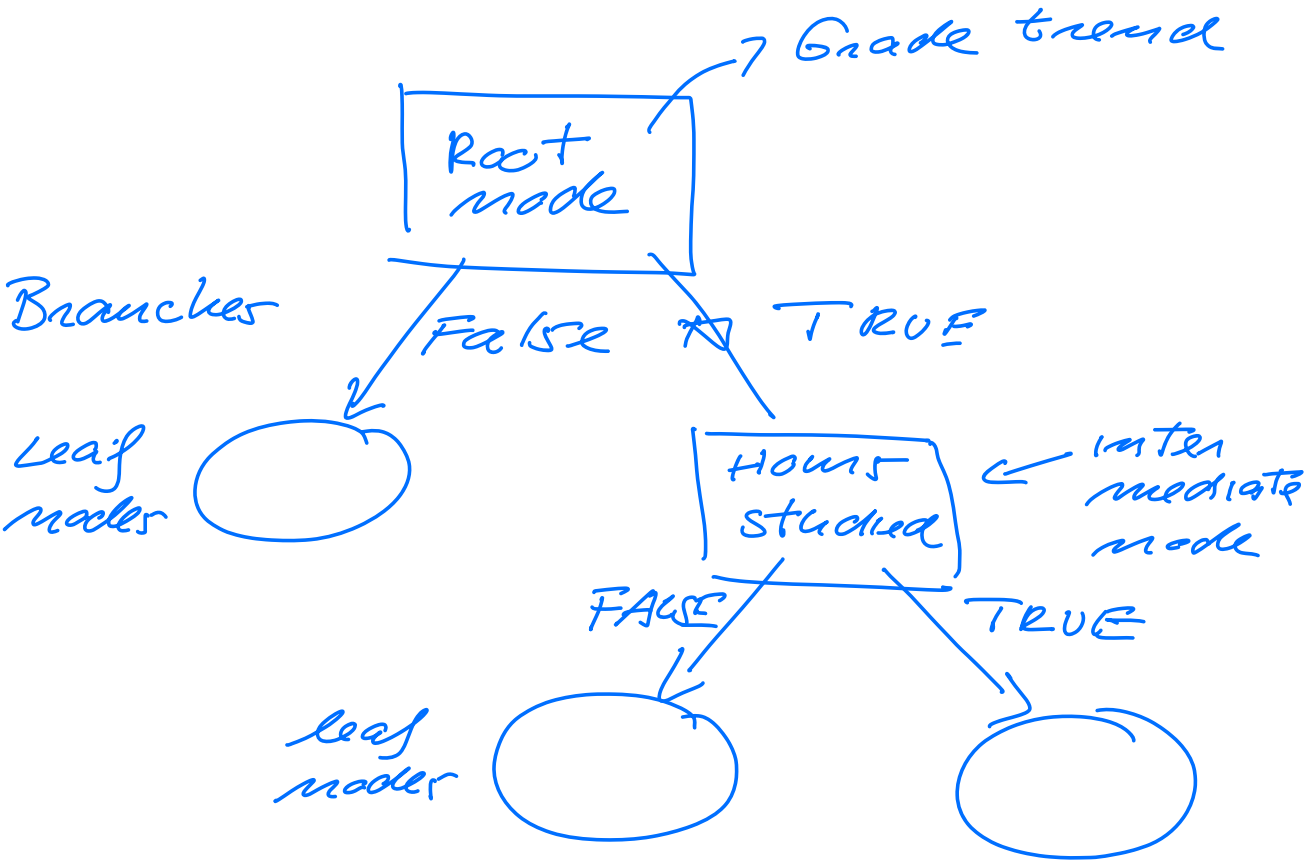
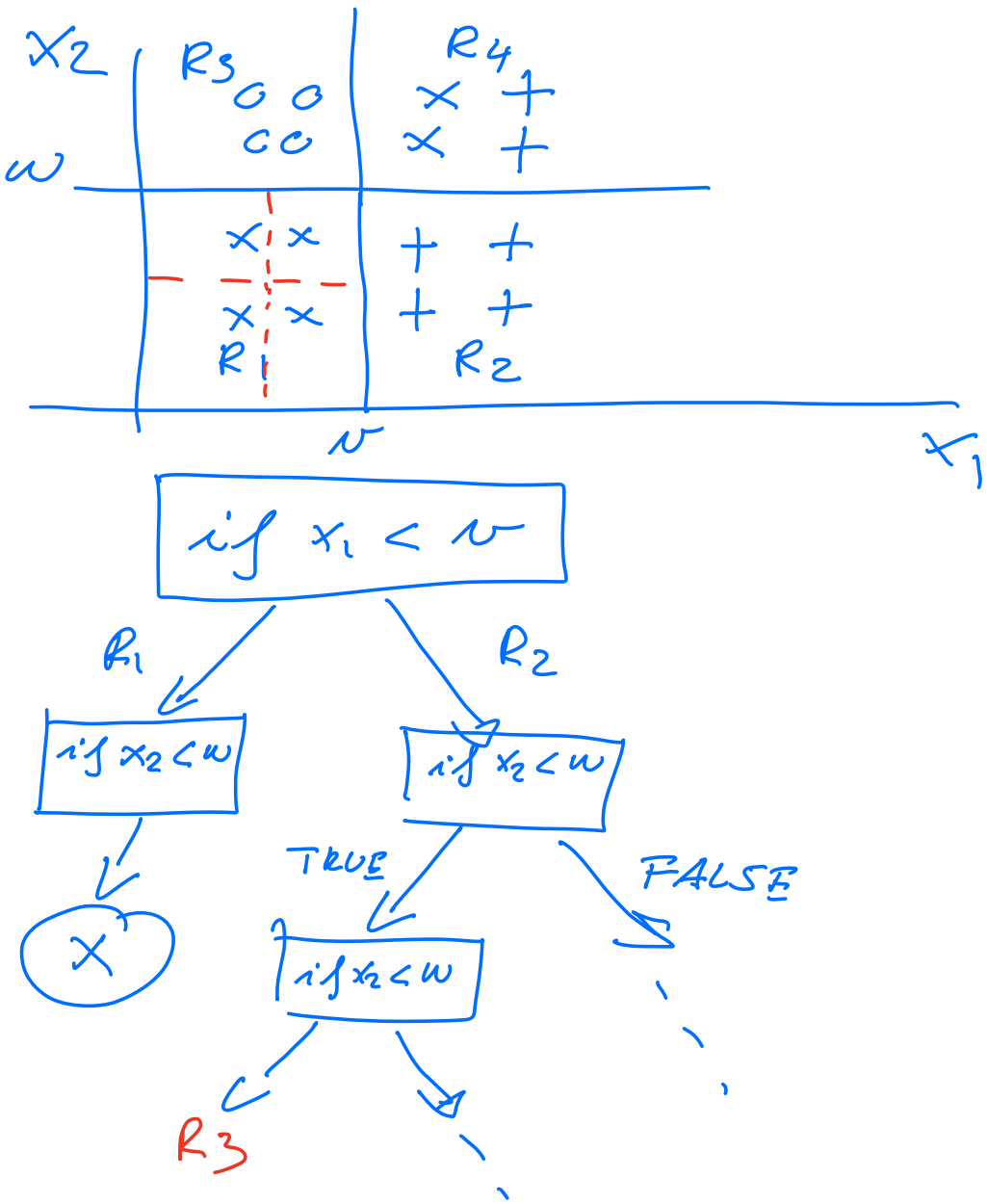
FYS-STK 4155, NOV 3, 2022

Decision tree classification

- Attributes/features
  - Grade trend (Above or below)
  - number of hours studied
  - number of hours slept
- output
  - Grade above or below average



Regression case



Classification algo for a tree

Nodes with zero or as less as possible impurity as

functions of the number of features/attributes/etc

- gini factor
- entropy

} Split trees and define nodes.

Example: two classes, with probability

- p
- $1-p$

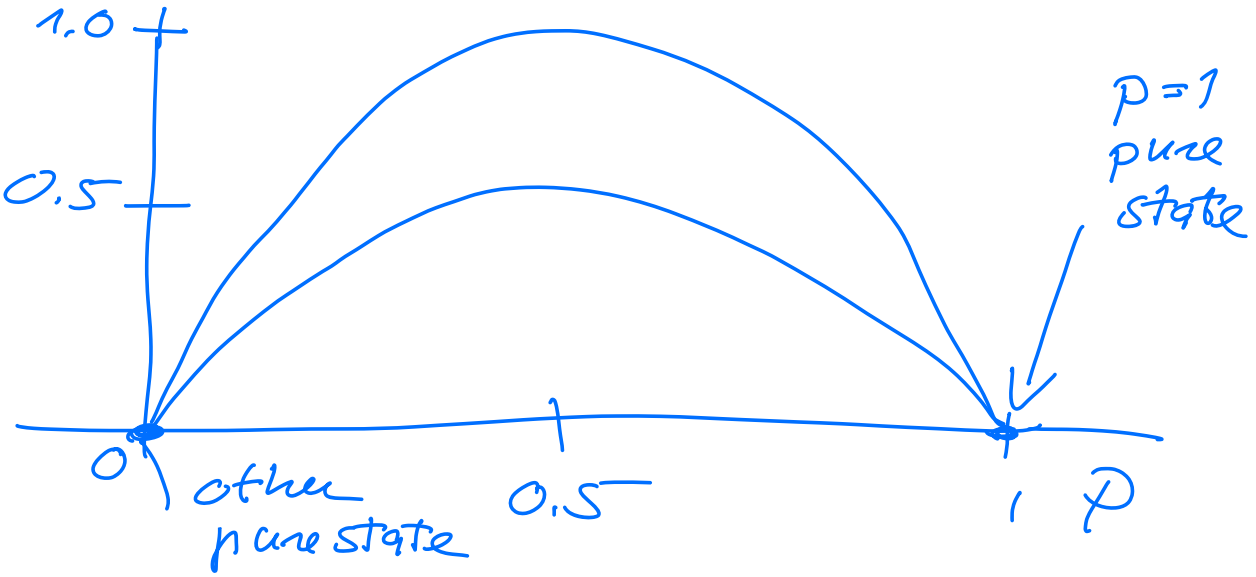
Entropy: $S = -\sum_{i=1}^{\text{classes}} p_i \log_2 p_i$

$= -p\log_2 p - (1-p)\log_2(1-p)$

Gini: $g = 1 - \sum_{i=1}^{\text{classes}} p_i^2$

$= 2p(1-p)$

$0 \log_2 0 = 0$!



0 = all elements belong to one specific attributes/ class

1 = all elements are randomly distributed across various classes

A gini factor of 0.5 denotes distributed classes.

CART with Gini factor.
(Binary splitting)

Attributes/classes

output

Grade Trend	hours Slept	hours studied	Grade
Above (1)	Low	high	Above
Below (0)	high	Low	Below
Above 1	Low	high	Above
Above 1	high	high	Above
Below 0	Low	high	Below
Above 1	Low	Low	Below
Below 0	high	high	Below
Below 0	Low	Low	Below
Above 1	Low	high	Above
Above 1	high		

Gini factor for the various classes,

P (Trend = above) = 6/10

P (Trend = below) = 4/10

Gini for Trend.

if (pasttrend = above & grade = Above)

P = 4/6

if (trend = above & grade = below)

P = 2/6 = 1/3

gini = $1 - ((4/6)^2 + (2/6)^2) = 0.45$

if (trend = below & grade above)

P = 0

if (trend = below & grade below)

P = 4/4 = 1

Gini index

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

Weighted sum for trend

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

Gini for hours slept

p (hours slept above)

= 4/10

p (hours slept less) = 6/10

if (slept above & grade above)

P = 2/4 = 1/2

if (slept above & grade below)

$P = 1/2$

Gini index $= 1 - ((\frac{1}{2})^2 + (1/2)^2) =$
0.5

if (slept less & grade above)

$P = 2/6 = 1/3$

if (slept less & grade below)

$P = 4/6 = 2/3$

Gini index $= 0.45$

weighted gini index:

$4/10 \cdot 0.5 + 6/10 \cdot 0.45 = 0.47$

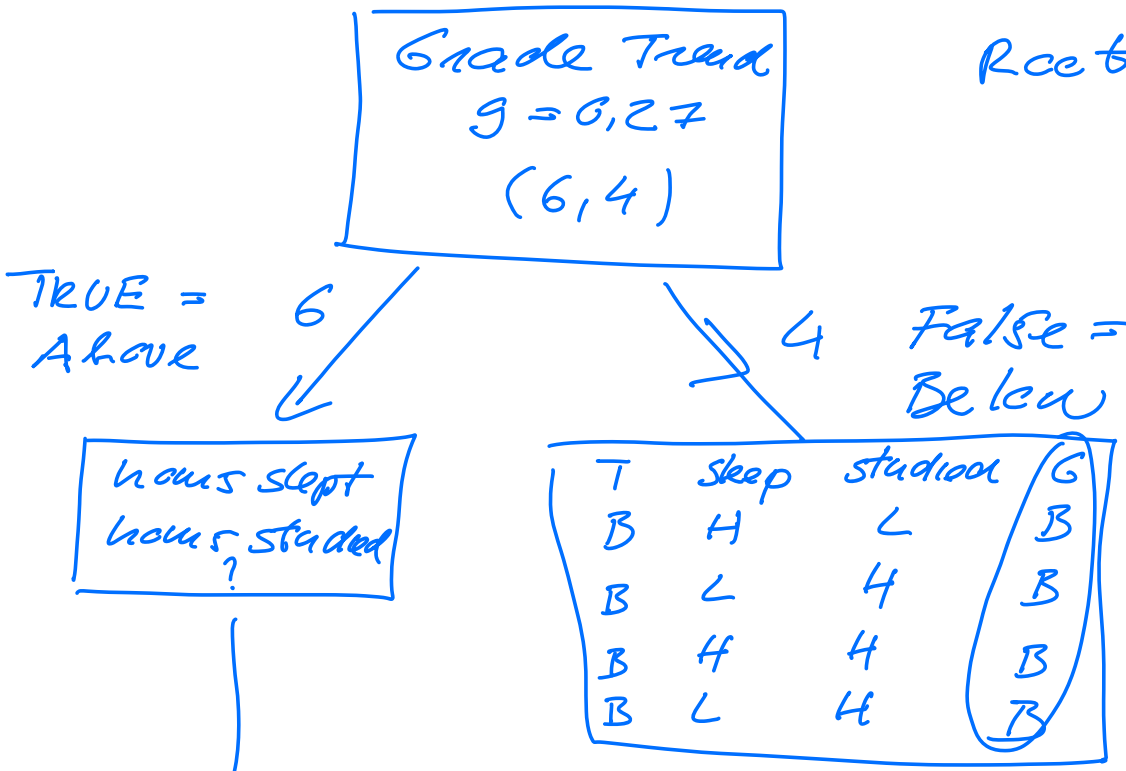
---

weighted gini for hours studied $= 0.34$

Feature	g
Trend	0.27
hours slept	0.47
hours studied	0.34

Gini factor which is smallest

defines the root node

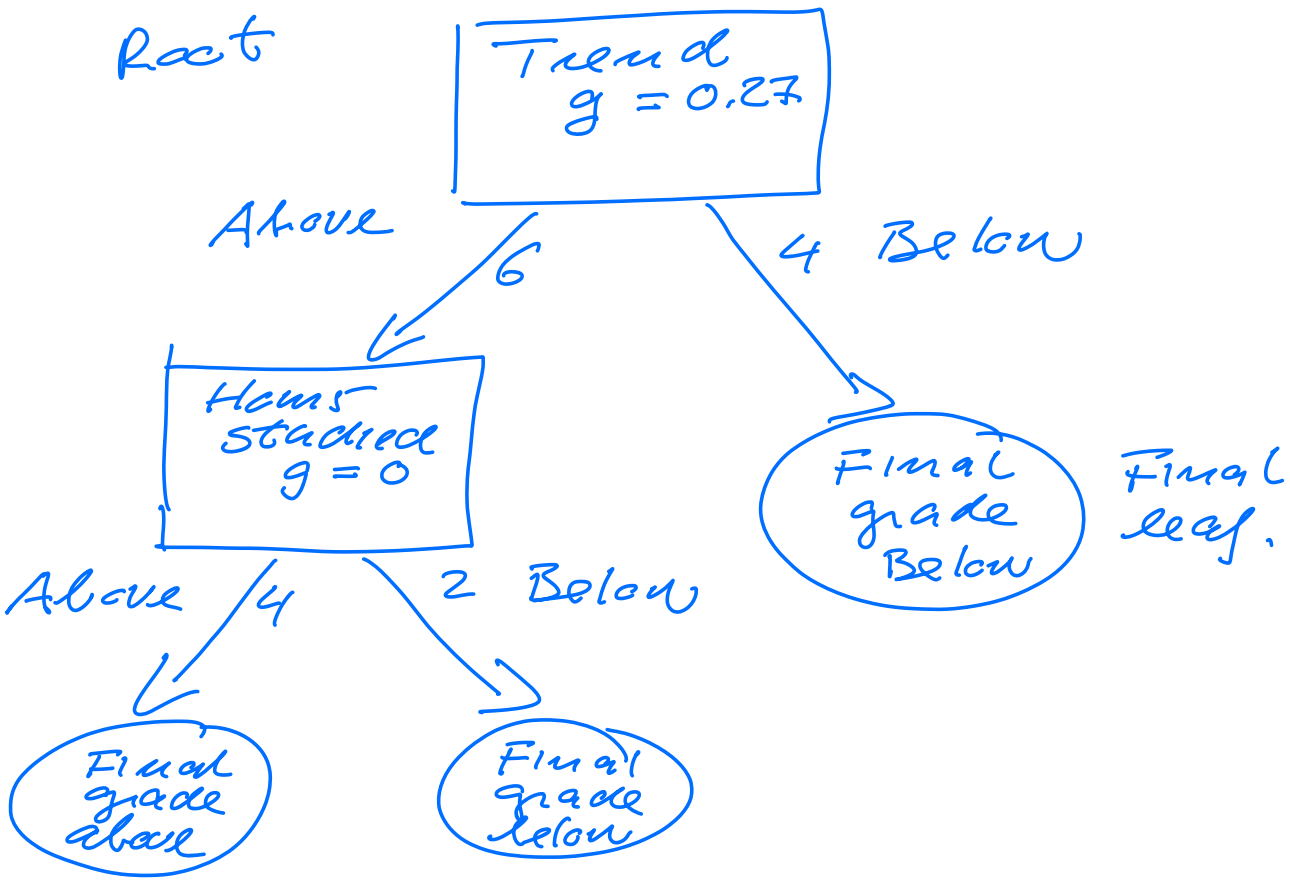


T	Sleep	Studied	G
B	H	L	B
B	L	H	B
B	H	H	B
B	L	H	B

T	Sleep	Studied	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

weighted Gini for hours studied
= 0

weighted Gini for hours slept
g = 0,33


Root
Trend
g = 0.27
Above
6
4 Below
Hours studied
g = 0
Final grade Below
Final leaf.
Above 4
2 Below
Final grade above
Final grade below