Decision Tree tutorial

Dataset

age	income	student	credit-rating	target buys_computer
Youth	high	yes	fair	
middleaged	medium	No	excellent	
senior	low			

Total_data = 14

: impurity measure :

- cross entropy | gini-Index

Multi-way split using cross entropy

Consider Attribute 'Age'

Youth = 3-No, 2-Yes   senior: 3-Yes, 2-No   middleaged: 4 Yes, 0-No

cross_entropy: $- \sum_{k=1}^{K} p_{mk} \log p_{mk}$ $K=2$

initially $m=1$

$k=1 \rightarrow$ class yes
$k=2 \rightarrow$ class No

$-(p_{11} \log p_{11} + p_{12} \log p_{12})$

For attribute age = Youth $p_{11} = \frac{2}{5}$, $p_{12} = \frac{3}{5}$

$$\text{cross entropy}_{Age} = +(5/14)\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right) + \frac{4}{14}\left(-\frac{4}{5}\log\frac{4}{5}\right) + \frac{5}{14}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{3}{5}\right)$$

$$= \boxed{0.6935}$$

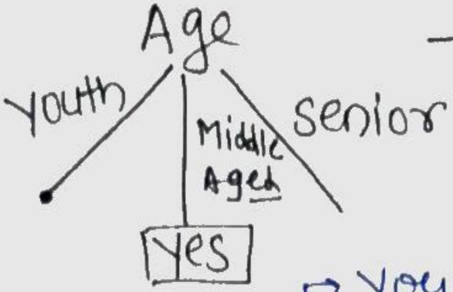
cross_entropy credit_range (D) = $\boxed{0.8922}$

cross_entropy_income : $\boxed{0.9111}$

cross_entropy_student : $\boxed{0.7885}$

less entropy →

Decision Tree

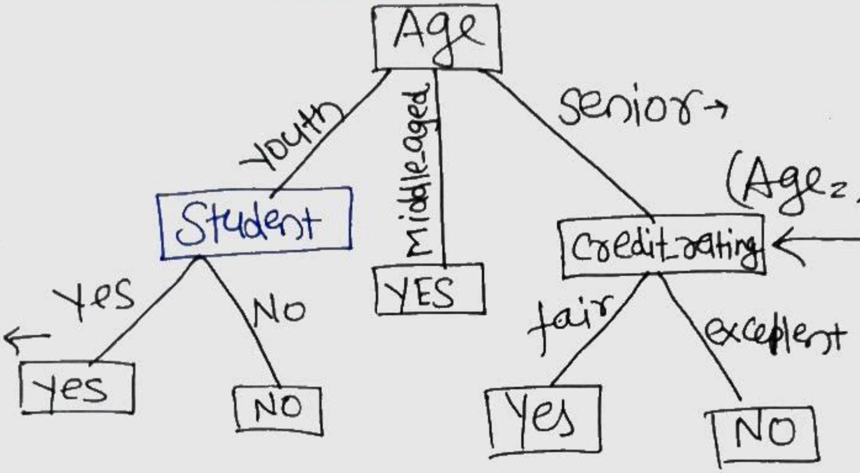


↱ You will filter dataset where <u>age=youth</u>.

$\text{cross\_entropy}_{\text{income}}(\text{Age}=\text{youth}) =$

$\text{cross\_entropy}_{\text{student}}(\text{Age}=\text{Youth}) = 0$ → <u>Minimum value of entropy</u>

$\text{cross\_entropy}_{\text{credit\_score}}(\text{Age}=\text{youth}) =$



We will stop since <u>entropy=0</u>

(Age=senior)

↓

CE_student = 0.95

CE_credit_rating=0

CE_income = 0.95

Filter data
↓
where
(Age=youth) $
<u>(Student=Yes)</u>

Decision-Tree:

Binary split using Gini index

Consider 'Age'

Totadata points = 14

(+ve) class_proportion: Age = Youth: 2/5

Age = Midd : 1

Age: Senior : 3/5

Ordering the probability.

Youth, senior, middle

$$\text{Gini-index} = \sum_{k=1}^{K} p_{mk}(1-p_{mk})$$

Possible splits

{Youth}, {senior, middle} | {youth, senior}, {middle}

⇒ we need to calculate impurity measure for both the above splits.

$$\text{Gini}_{age\in\{youth\}}(D) := 2p(1-p)$$

$$= \frac{5}{14}\left(2*\frac{2}{5}*\frac{3}{5}\right) + \frac{9}{14}\left(2*7/9*2/9\right)$$

$$= 0.6508$$

For '2' class problem $G = p(1-p) + (1-p)p$

$$= \boxed{2p(1-p)}$$

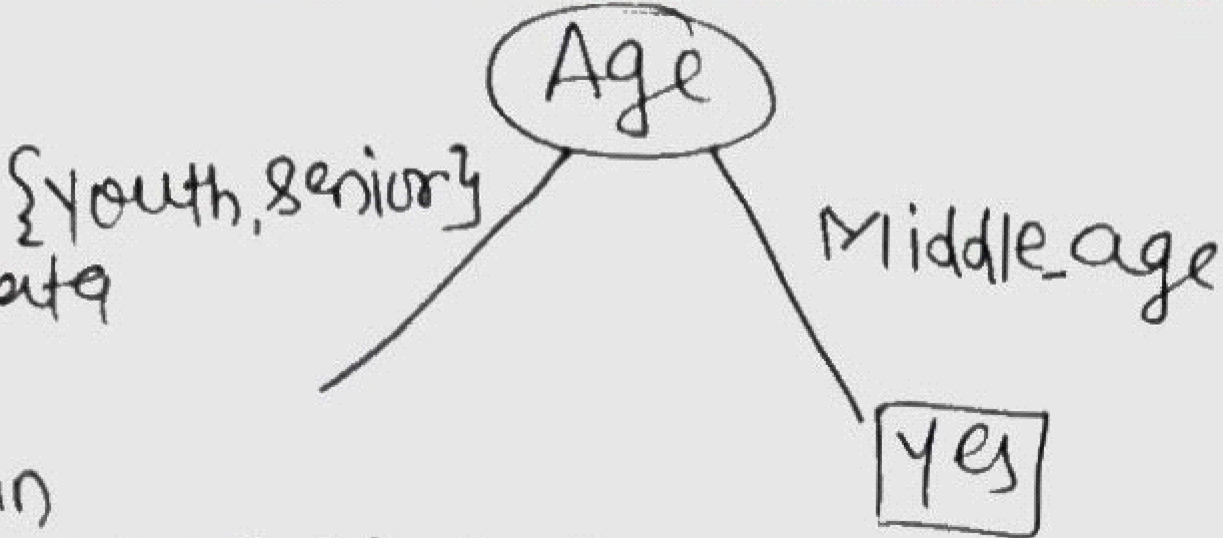
Gini-age = {Youth, senior} =

Gini-student {Yes} = $7/14\left(2*\frac{3}{7}\cdot\frac{4}{7}\right) + 7/14\left(2\cdot\cancel{2}/\cancel{7}\cdot 1/2\right)$

Gini-income → Two splits =

Gini-creditrating {~~fair~~} $= \frac{8}{14}\left(2*6/8*2/8\right) + \frac{6}{14}\left(2*\frac{3}{6}*\frac{3}{6}\right)$

Choose the one having less impure.: lesser value



filter data
↓
again
Calculate gini_index for all other attribute other than age.