ARTIFICIAL INTELLIGENCE & CYBER SECURITY

MACHINE LEARNING
ENTROPY

Nacir Bouali

n.bouali@utwente.nl

UNIVERSITY OF TWENTE.

DECISION TREES

- To build (induce) a decision tree we need;
  - Data
  - An algorithmic method to select the best attribute at the root of the tree.
  - Repeat the process recursively.

ALGORITHMIC METHOD

- Several methods to select the best attribute/feature exist.
- We will work with information gain.
- Information gain is based on the entropy of a dataset.
- Entropy is a measure of chaos in the dataset. The less chaos the better. The more chaos, the more information needed to tell the class of a datapoint/example.

ENTROPY

- Entropy is a measure of chaos in the dataset. The less chaos the better. The more chaos, the more information needed to tell the class of a datapoint/example.
- Given by the formula below;

$$\mathrm{E(S)} = \sum_{i=1}^{c} -Pi\ Log_2(Pi)$$

ENTROPY

Class/Label

Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
1	Yes	Mild	No	Wet	Taste	Yes
2	No	Strong	No	Wet	Both	Yes
3	Yes	Mild	No	Absent	Taste	No
4	No	Mild	No	Wet	Smell	Yes
5	No	Mild	Yes	Wet	Taste	Yes
6	Yes	Lite	No	Absent	Smell	Yes
7	No	Lite	No	Dry	Smell	No
8	No	Strong	No	Absent	Taste	Yes
9	No	Absent	Yes	Wet	Both	No
10	No	Mild	No	Wet	Smell	Yes
11	No	Absent	Yes	Absent	Taste	No
12	No	Lite	No	Dry	Both	No
13	Yes	Strong	No	Wet	Both	Yes
14	No	Absent	Yes	Wet	Both	No
15	No	Mild	No	Absent	Taste	Yes
16	Yes	Lite	No	Dry	Smell	No
17	No	Absent	No	Absent	Taste	Yes
18	No	Absent	No	Wet	Smell	Yes
19	No	Lite	No	Dry	Both	No
20	No	Lite	No	Dry	Smell	No

$$\sum_{i=1}^{c} -Pi\ Log_2(Pi)$$

2

$$-P(Covid = yes)Log_2(P(Covid = yes) - P(Covid = no)Log_2(Covid = no)$$

$$-\left(\frac{11}{20}\right)Log_2\left(\frac{11}{20}\right) - \left(\frac{9}{20}\right)Log_2\left(\frac{9}{20}\right)$$

=0.9927744539878083

ENTROPY

class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$\sum_{i=1}^{c} -Pi\ Log_2(Pi)$$

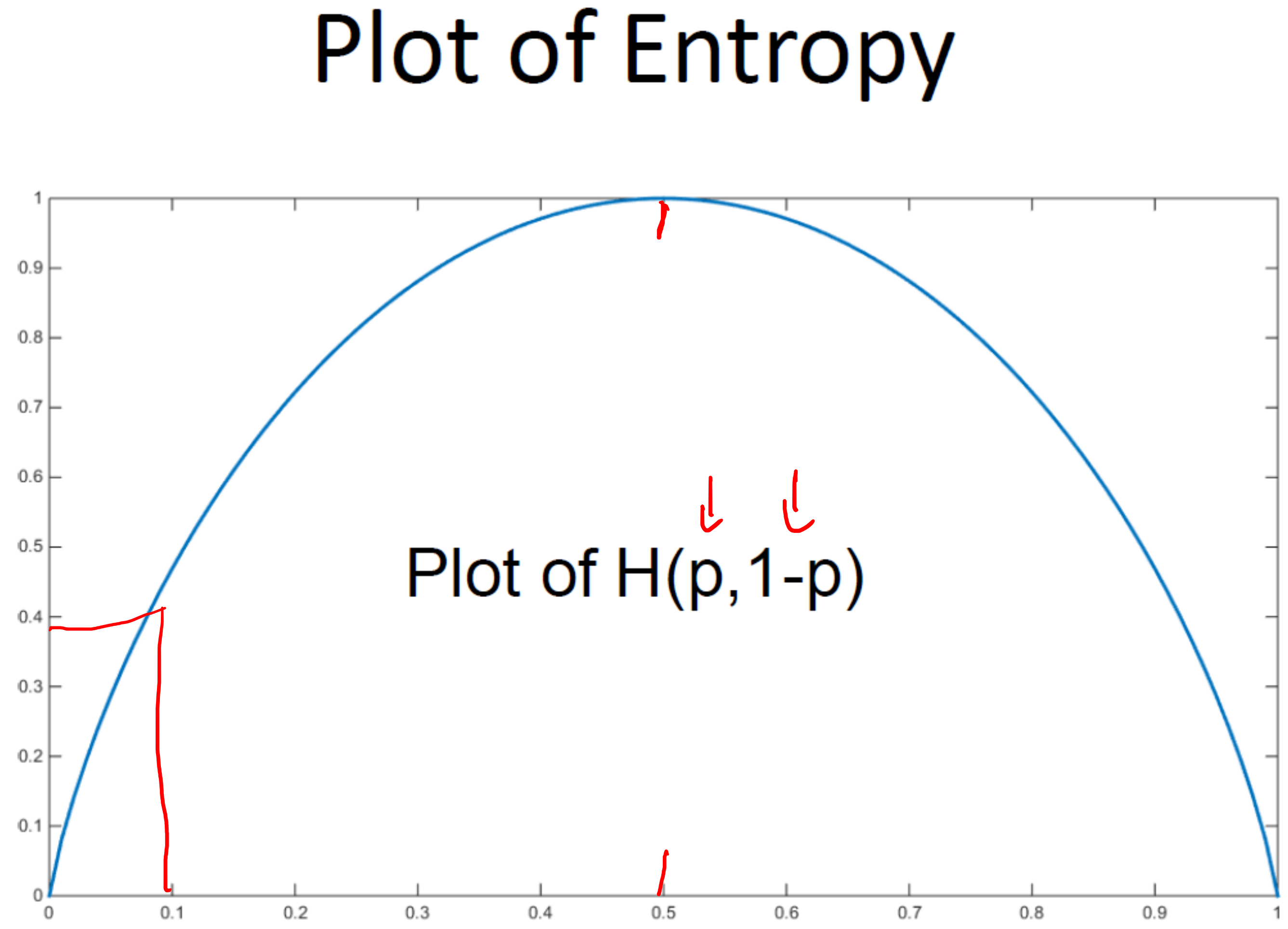
$$= -P(Cheat = yes)Log_2(P(Cheat = yes) - P(Cheat = no)Log_2(Cheat = no)$$

$\frac{3}{10}$ $\quad$ $7/10$

$$-\left(\frac{3}{10}\right)Log_2\left(\frac{3}{10}\right) - \left(\frac{7}{10}\right)Log_2\left(\frac{7}{10}\right)$$

=0.88

ENTROPY



Ex.	Fever	Headache	Fatigue	Cough	TasteSmell Loss	Covid
1	Yes	Mild	No	Wet	Taste	Yes
2	No	Strong	No	Wet	Both	Yes
3	Yes	Mild	No	Absent	Taste	No
4	No	Mild	No	Wet	Smell	Yes
5	No	Mild	Yes	Wet	Taste	Yes
6	Yes	Lite	No	Absent	Smell	Yes
7	No	Lite	No	Dry	Smell	No
8	No	Strong	No	Absent	Taste	Yes
9	No	Absent	Yes	Wet	Both	No
10	No	Mild	No	Wet	Smell	Yes
11	No	Absent	Yes	Absent	Taste	No
12	No	Lite	No	Dry	Both	No
13	Yes	Strong	No	Wet	Both	Yes
14	No	Absent	Yes	Wet	Both	No
15	No	Mild	No	Absent	Taste	Yes
16	Yes	Lite	No	Dry	Smell	No
17	No	Absent	No	Absent	Taste	Yes
18	No	Absent	No	Wet	Smell	Yes
19	No	Lite	No	Dry	Both	No
20	No	Lite	No	Dry	Smell	No
21	x	x	x	x	x	?

ENTROPY

- Entropy is a measure of chaos in the dataset. The less chaos the better. The more chaos, the more information needed to tell the class of a datapoint/example.
- Given by the formula below;

$$\mathrm{E(S)} = \sum_{i=1}^{c} -Pi\ Log_2(Pi)$$

- In case of binary classification problems, max entropy is 1.