

HOWEST TOEGEPASTE INFORMATICA, 2022-2023, © BRIAN BAERT

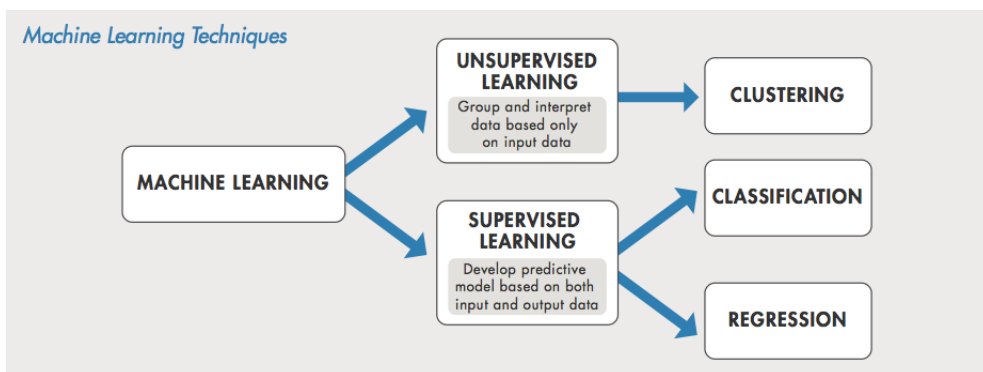
DATA ANALYTICS

HOOFDSTUK 6 - CLASSIFICEREN

howest.be

1

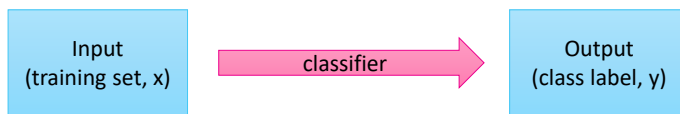
SUPERVISED VS UNSUPERVISED LEARNING



2

CLASSIFICEREN

- **Wat is classificeren?**
 - Op basis van een verzameling records (*training set*) een model zoeken voor het class attribuut als functie van de waarden uit de training set (*classifier*).
 - Het doel is voorheen onbekende records toekennen aan een bepaalde *class*.



3

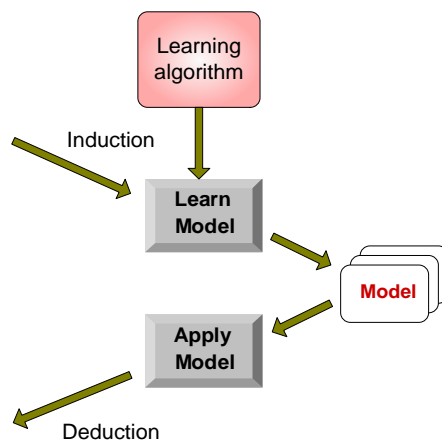
CLASSIFICEREN

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|-----|---------|---------|---------|-------|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|-----|---------|---------|---------|-------|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

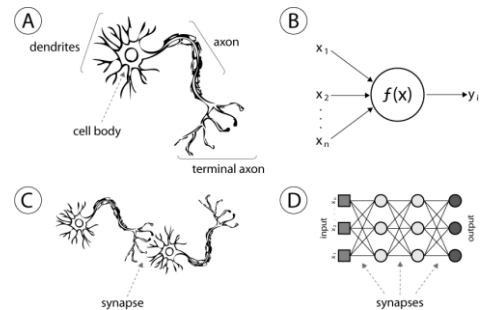
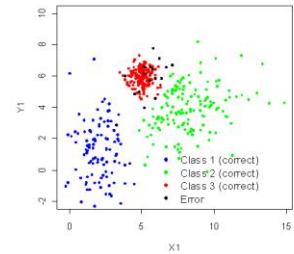
Test Set



4

CLASSIFICATIE TAKEN

- Voorbeelden
 - voorspellen van een tumorcel of die goedaardig of kwaadaardig is
 - classificeren van kredietkaarttransacties als frauduleus of legitiem
 - categoriseren van twitterberichten op emotie
 - ...
- Technieken
 - beslissingsbomen
 - regel-gebaseerde methodes
 - neurale netwerken
 - Naïef Bayes en Bayesiaanse netwerken
 - Support Vector Machines

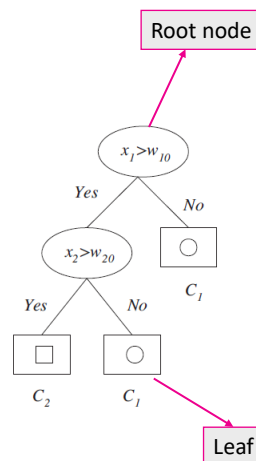
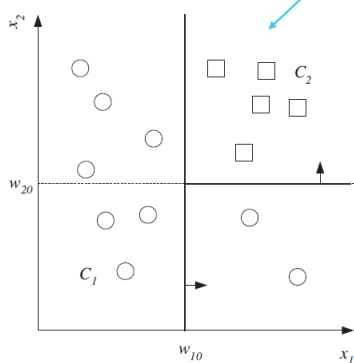


5

BESLISSINGSBOOM

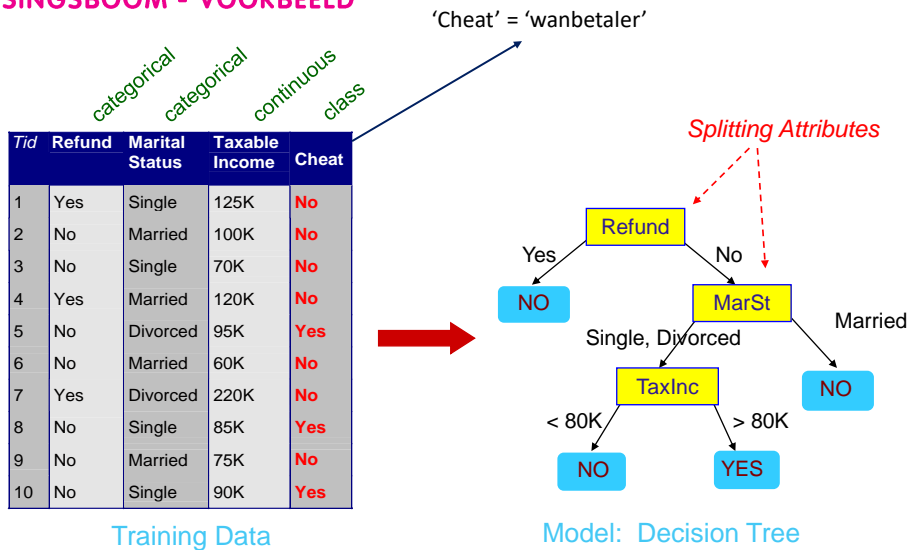
Zoeken van "Decision boundaries"

- Bevat drie soorten nodes:
 - root node
 - internal node
 - leaf of terminals



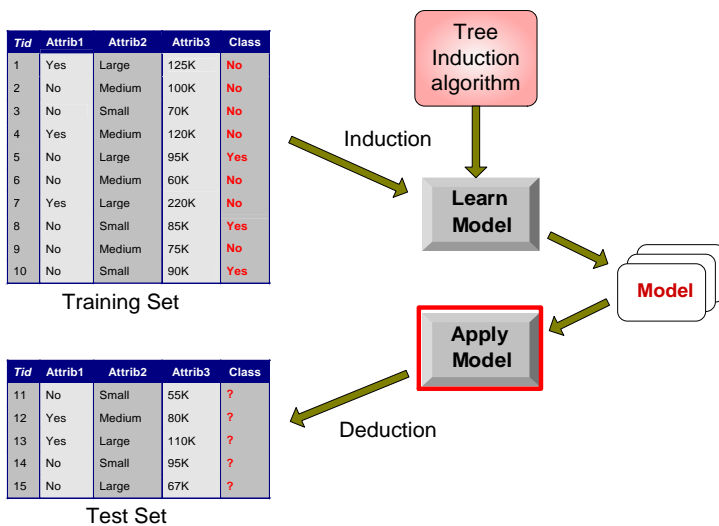
6

BESLISSINGSBOOM - VOORBEELD



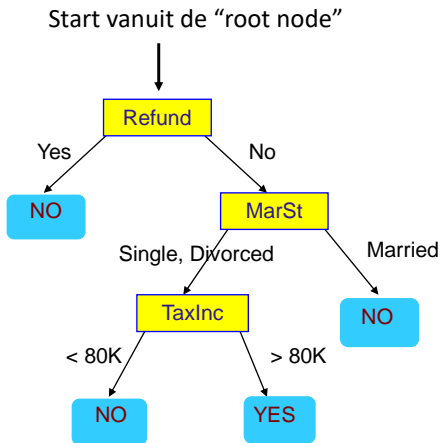
7

BESLISSINGSBOOM CLASSIFICATIETAAK



8

MODEL TOEPASSEN OP TEST SET

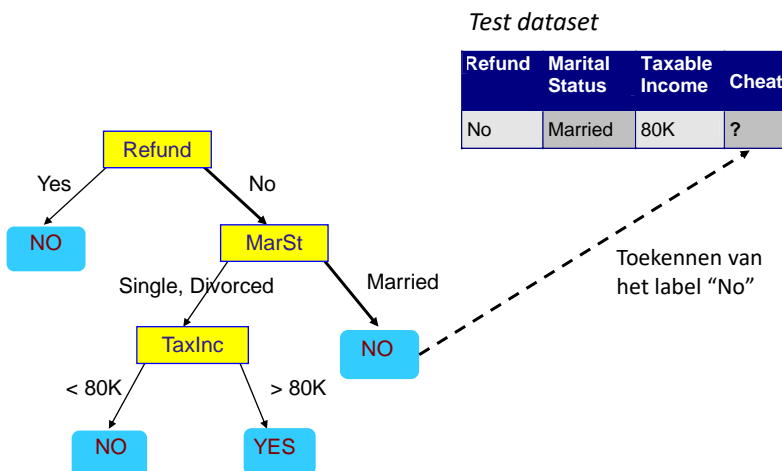


Test dataset

| Refund | Marital Status | Taxable Income | Cheat |
|--------|----------------|----------------|-------|
| No | Married | 80K | ? |

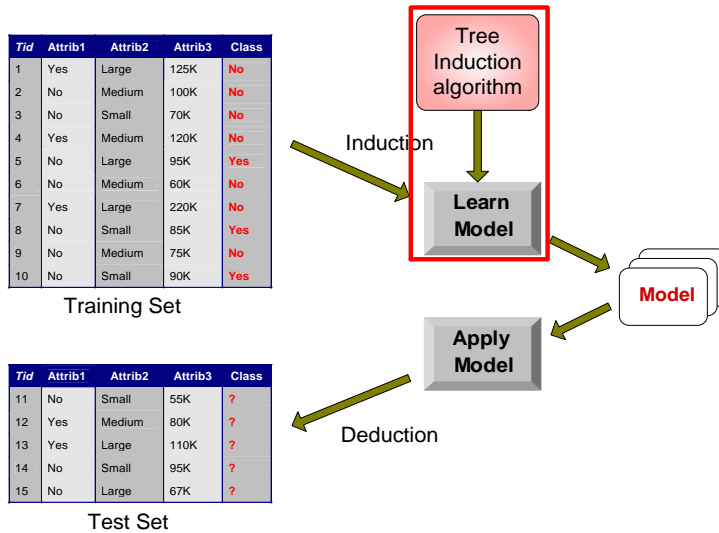
9

MODEL TOEPASSEN OP TEST DATA



10

BESLISSINGSBOOM CLASSIFICATIETAAK



11

BESLISSINGSBOOM ALGORITMES

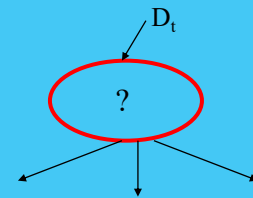
- Algoritme van Hunt
- CART
- ID3 , C4.5 , C4.8
- CHAID
- MARS
- SLIQ
- SPRINT

12

ALGORITME VAN HUNT

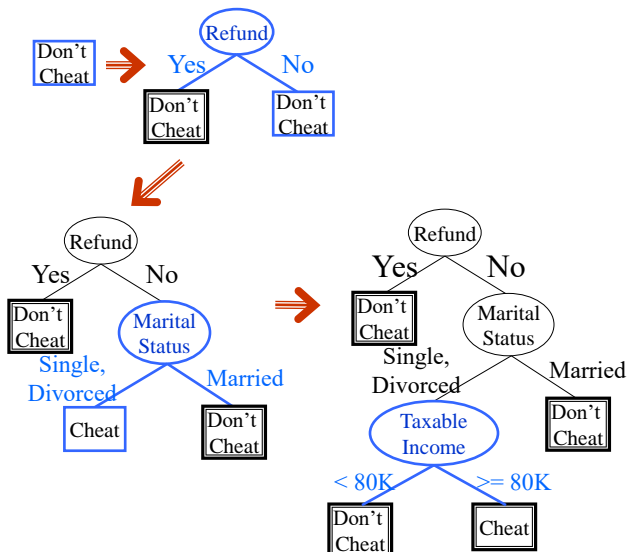
- Stel D_t de verzameling van training records die een node t bereiken en $y = \{y_1, y_2, \dots, y_c\}$ de class labels
- Algemene procedure
 - 1) Als D_t records bevat die behoren tot eenzelfde klasse y_t , dan is t een bladnode (leaf node) met als label y_t .
 - 2) Als D_t records bevat die behoren tot meer dan één klasse, gebruik dan een attribuuttestvoorwaarde om de data te splitsen in meerdere, kleinere deelverzamelingen. We herhalen deze procedure recursief.

| 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 |



13

BESLISSINGSBOOM - VOORBEELD



| 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 |

14

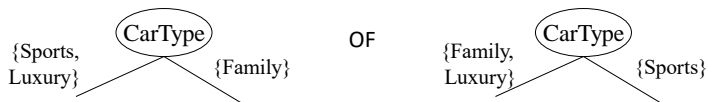
BOOMINDUCTIE

- 'Greedy' strategie
Splits de records op basis van een attribuuttest die een bepaald criterium optimaliseert.
- *Probleem 1*: Hoe splitsen we de records?
 - Specificeren van attribuuttestvoorwaarde
 - Bepalen van beste splitsing
- *Probleem 2*: Hoe bepalen wanneer we stoppen met splitsen?
 - Als de node-onzuiverheid kleiner is dan een vooropgestelde threshold (grenswaarde)

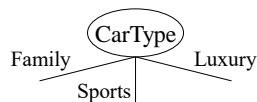
15

BEPALEN VAN ATTRIBUUTTESTVOORWAARDE

- Afhankelijk van *attribuuttype*
 - Nominaal → Multiway of Binary split
 - Ordinaal → Multiway of Binary split
 - Continu → vergelijkende test
- **Binary split** (of binaire split)



- **Multiway split** (meervoudige splitsing)



Delen van waarden in twee deelverzamelingen, nodig om een optimale partitionering te vinden.

Zoveel partities als verschillende waarden

16

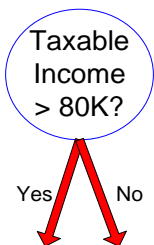
BEPALEN VAN ATTRIBUUTTESTVOORWAARDE

- Continue attributen
verschillende manieren van afhandeling
 - **Discretiseren**
 - Statisch (discretiseren bij de start)
 - Dynamisch (bereik kan gevonden worden door b.v. percentielen, gelijke intervallen)
 - **Binaire beslissing ($A < v$ of $A \geq v$)**
 - Bepaal alle mogelijke splitsingen en zoek de beste
 - Computerintensief

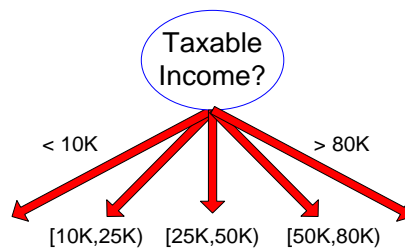
17

BEPALEN VAN ATTRIBUUTTESTVOORWAARDE

- Continue attributen



(i) Binary split



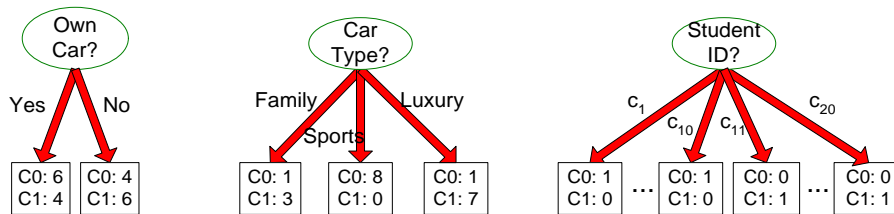
(ii) Multi-way split

$$v_i \leq A < v_{i+1}, \text{ voor } i = 1, \dots, k$$

18

HOE BEPALEN WE DE BESTE SPLITSING?

Voor splitsing: 10 records van klasse **C0**,
10 record van klasse **C1**



Welke voorwaarde is de beste?

19

HOE BEPALEN WE DE BESTE SPLITSING?

- Greedy benadering**
Nodes met een homogene klassenverdeling zijn vereist
- Waarom kiezen voor een bepaalde feature?
→ Nood aan een maat voor “**node onzuiverheid**” (impurity)

C0: 5
C1: 5

**Niet-homogeen,
Hoge onzuiverheidsgraad**

C0: 9
C1: 1

**Homogeen,
Lage onzuiverheidsgraad**

?Greedy benadering?:

- 1) Bij elke node, kies de “beste feature” om te splitsen.
- 2) Splits
- 3) Herhaal

20

MATEN VOOR NODE ONZUIVERHEID



Entropie

$p(i|t)$ = de fractie van records die behoren tot de klasse 'i' in een node 't'

$$\text{Entropie}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

Gini

'c' = aantal klassen

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

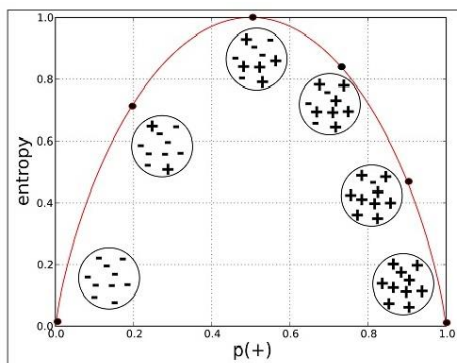
Classificatiefout

$$\text{classificatiefout}(t) = 1 - \max_i [p(i|t)]$$

21

ENTROPIE MAAT VOOR ONZUIVERHEID

- De entropie is de verwachte hoeveelheid informatie
- Doorgaans symbolisch genoteerd met $H(t)$
- Voorbeeld: *Entropie bij een binair classificatieprobleem*

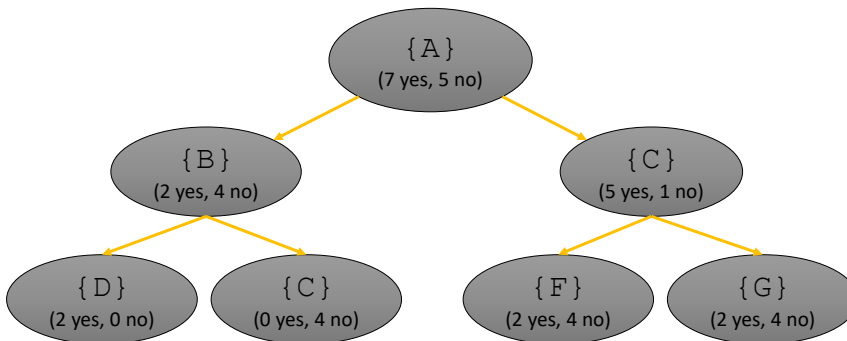


22

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Veronderstel een dataset, 12 records, met 4 features (attributen), {A, B, C, D, F, G} en 1 output (uitvoer) {yes | no}
- De beslissingsboom ziet er als volgt uit

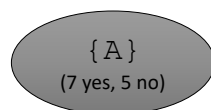


23

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Entropie ligt tussen 0 en 1
 - ❖ 0 = zuiver
 - ❖ 1 = meest onzuiver



- Berekening voor de volledige dataset (S)

$$H(S) = -\frac{7}{12} \cdot \log_2 \left(\frac{7}{12} \right) - \frac{5}{12} \cdot \log_2 \left(\frac{5}{12} \right)$$

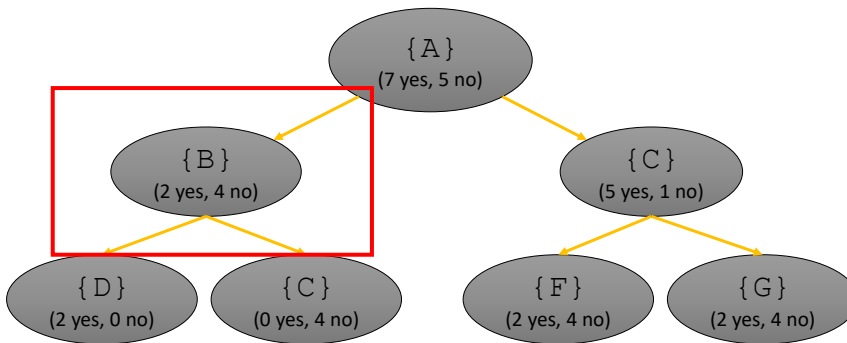
= 0,97 bits

24

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Veronderstel een dataset, 12 records, met 6 features (attributen), $\{A, B, C, D, F, G\}$ en 1 output (uitvoer) $\{yes | no\}$
- De beslissingsboom ziet er als volgt uit

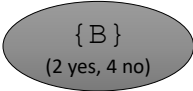


25

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Veronderstel een dataset, 12 records, met 6 features (attributen), $\{A, B, C, D, F, G\}$ en 1 output (uitvoer) $\{yes | no\}$



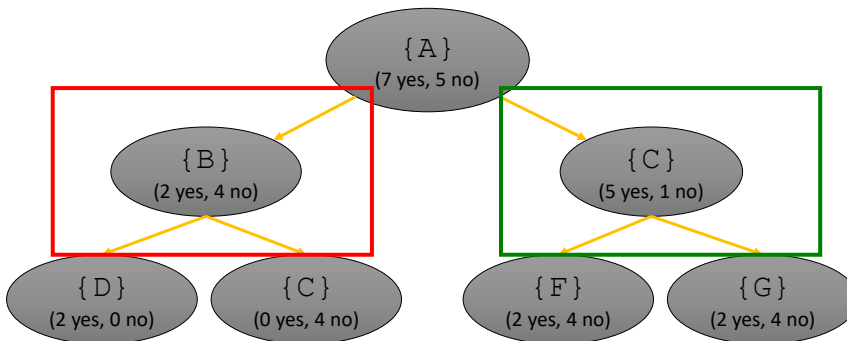
$$\begin{aligned}
 H(B) &= -\frac{2}{2+4} \cdot \log_2 \left(\frac{2}{2+4} \right) - \frac{4}{4+2} \cdot \log_2 \left(\frac{4}{4+2} \right) \\
 &= -\frac{2}{6} \cdot \log_2 \left(\frac{2}{6} \right) - \frac{4}{6} \cdot \log_2 \left(\frac{4}{6} \right) \\
 &= 0,92 \text{ bits}
 \end{aligned}$$

26

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Veronderstel een dataset, 12 records, met 6 features (attributen), {A, B, C, D, F, G} en 1 output (uitvoer) {yes | no}
- De beslissingsboom ziet er als volgt uit

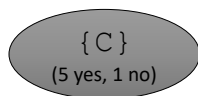


27

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Veronderstel een dataset, 12 records, met 6 features (attributen), {A, B, C, D, F, G} en 1 output (uitvoer) {yes | no}



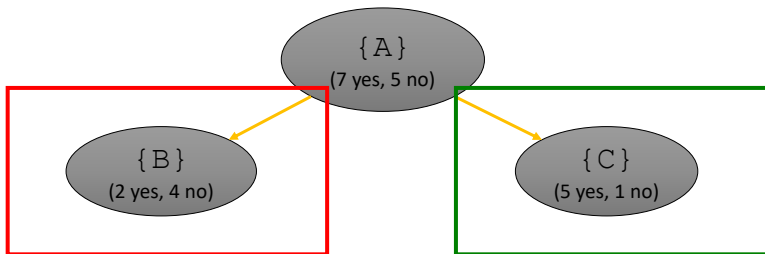
$$\begin{aligned}
 H(C) &= -\frac{5}{5+1} \cdot \log_2 \left(\frac{5}{5+1} \right) - \frac{1}{5+1} \cdot \log_2 \left(\frac{1}{5+1} \right) \\
 &= -\frac{5}{6} \cdot \log_2 \left(\frac{5}{6} \right) - \frac{1}{6} \cdot \log_2 \left(\frac{1}{6} \right) \\
 &= 0,65 \text{ bits}
 \end{aligned}$$

28

INFORMATIEWINST VIA ENTROPIE

Voorbeeld

- Veronderstel een dataset, 12 records, met 4 features (attributen), {A, B, C, D, E, G} en 1 output (uitvoer) {yes | no}



$$\begin{aligned}\Delta(S, A) &= 0,97 - \frac{|S_B|}{|S|} \cdot (0,92) - \frac{|S_C|}{|S|} \cdot (0,65) \\ &= 0,97 - \frac{6}{12} \cdot (0,92) - \frac{6}{12} \cdot (0,65) = \mathbf{0,185}\end{aligned}$$

👉 ALGEMEEN 👉:

Wij kiezen te splitsen op het attribuut (feature) met de hoogste informatiewinst (gain, Δ).

29

GINI MAAT VOOR ONZUIVERHEID

- Gini-index voor een gegeven node t

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

$p(i|t)$ is de relatieve frequentie van klasse i bij node t

- Maximum ($1-1/n_c$) als records gelijk verdeeld zijn over alle klassen, minst bruikbare informatie
- Minimum (0,0) als alle records tot één klasse behoren, grootste informatiewinst.

| | |
|------------|---|
| C1 | 0 |
| C2 | 6 |
| Gini=0.000 | |

| | |
|------------|---|
| C1 | 1 |
| C2 | 5 |
| Gini=0.278 | |

| | |
|------------|---|
| C1 | 2 |
| C2 | 4 |
| Gini=0.444 | |

| | |
|------------|---|
| C1 | 3 |
| C2 | 3 |
| Gini=0.500 | |

Gini is een maat voor de kans op een verkeerde classificatie van een nieuw record.

30

GINI MAAT VOOR ONZUIVERHEID - VOORBEELD

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [P(i|t)]^2$$

| | |
|----|----------|
| C1 | 0 |
| C2 | 6 |

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$\text{Gini} = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$$

| | |
|----|----------|
| C1 | 1 |
| C2 | 5 |

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$$

| | |
|----|----------|
| C1 | 2 |
| C2 | 4 |

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$\text{Gini} = 1 - (2/6)^2 - (4/6)^2 = 0.444$$

31

SPLITSEN VAN BINAIRE ATTRIBUTEN – GINI COËFFICIËNT

Voorbeeld

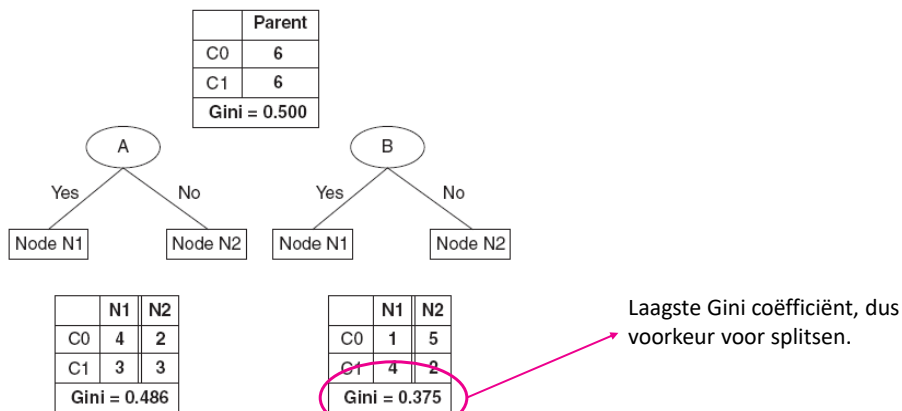


Figure 4.14. Splitting binary attributes.

32

OEFENINGENREEKS

- 6.2.7
1-3
4 (thuis)