Oplossingen oefeningen hoofdstuk 6 - 2.7

Oefening 1

Beschouw volgende training records van een binair classificatieprobleem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a. Bereken de Gini-index voor de volledige verzameling aan training records.
b. Bereken de Gini-index voor het `Customer ID` en het `Gender` attribuut.
c. Bereken de Gini-index voor het `Car Type` en `Shirt Size` attribuut gebruik makend van meervoudige splitsing.
d. Welk van de attributen verkies je?

Oplossing Oefening 1

a)

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$
$$= 1 - \left[\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right]$$
$$= 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$$

b)

Gini(customer ID) $= 1 - 1 = 0$ (*geen informatie want voor elk record verschillend*)

$$\text{Gini(Gender=Female)} = 1 - \left[\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right]$$
$$= 1 - [(0{,}16) + (0{,}36)]$$
$$= 1 - 0{,}52 = 0{,}48$$
$$\text{Gini(Gender=Male)} = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right]$$
$$= 1 - [(0{,}36) + (0{,}16)]$$
$$= 1 - 0{,}52 = 0{,}48$$

$$\textbf{Gini(Gender)} = \left(\frac{10}{20}\right).0{,}48 + \left(\frac{10}{20}\right).0{,}48 = 0{,}48$$

c)

$$\text{Gini(Car type=Family)} = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right]$$
$$= 1 - [(0{,}625) + (0{,}563)]$$
$$= 1 - 0{,}625 = 0{,}375$$

$$\text{Gini(Car type=Luxury)} = 1 - \left[\left(\frac{1}{8}\right)^2 + \left(\frac{7}{8}\right)^2\right]$$
$$= 1 - [(0{,}016) + (0{,}766)]$$
$$= 1 - 0{,}782 = 0{,}218$$

$$\text{Gini(Car type=Sports)} = 1 - \left[\left(\frac{8}{8}\right)^2 + \left(\frac{0}{8}\right)^2\right]$$
$$= 1 - 1 = 0$$

$$\textbf{Gini(Car type)} = \left(\frac{4}{20}\right).0{,}375 + \left(\frac{8}{20}\right).0{,}218 + \left(\frac{8}{20}\right).0 = 0{,}163$$

$$\text{Gini(shirt size=small)} = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$$
$$= 1 - 0{,}52 = 0{,}48$$

$$\text{Gini(shirt size=medium)} = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right]$$
$$= 1 - 0{,}51 = 0{,}49$$

$$\text{Gini(shirt size=large)} = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$$
$$= 1 - 0{,}5 = 0{,}5$$

$$\text{Gini(shirt size=extra large)} = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$$
$$= 1 - 0{,}5 = 0{,}5$$

$$\textbf{Gini(shirt size)} = \left(\frac{5}{20}\right).0{,}48 + \left(\frac{7}{20}\right).0{,}49 + \left(\frac{4}{20}\right).0{,}5 + \left(\frac{4}{20}\right).0{,}5 = 0{,}492$$

d) `Car type` heeft de laagste Gini-index, dus dit attribuut krijgt de voorkeur. Het geeft ook de hoogste informatiewinst. We kunnen dit ook m.b.v. de informatiewinst
($\Delta = \text{Gini(dataset)} - \text{Gini(attribuut)}$) weergeven.

Voor Gender: $\Delta_{Gender} = \frac{1}{2} - 0{,}48 = 0{,}02$

Voor Car type: $\Delta_{car\ type} = \frac{1}{2} - 0{,}163 = 0{,}327$

Voor shirt size: $\Delta_{shirt\ size} = \frac{1}{2} - 0{,}492 = 0{,}008$

Customer ID laten we hier achterwege, omdat het onmogelijk is om een beslissingsboom te maken waarbij voor elk record een afzonderlijke tak gemaakt wordt.

Oefening 2

Beschouw volgende training records van een binair classificatieprobleem.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

a) Wat is de entropie van de verzameling aan training records?

b) Wat is de informatiewinst van a_1 en a_2 in relatie tot de training records?

c) Voor a_3, een continu attribuut, bereken de informatiewinst voor elke mogelijke splitsing.

d) Wat is de beste splitsing (tussen a_1, a_2 en a_3) rekening houdend met de informatiewinst?

e) Wat is de beste splitsing (tussen a_1 en a_2) rekening houdend met de classificatiefout?

f) Wat is de beste splitsing (tussen a_1 en a_2) rekening houdend met de Gini-index?

Oplossing oefening 2

a)

$$\text{Entropie}(t) = -\sum_{i=0}^{c-1} p(i|t).\log_2 p(i|t) = -\left[\left(\frac{4}{9}\right).\log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right).\log_2\left(\frac{5}{9}\right)\right]$$
$$= -[(-0{,}519) + (-0{,}471)] = 0{,}991$$

b)

$$\text{Entropie}(a_1 = T) = -\left[\left(\frac{3}{4}\right).\log_2\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right).\log_2\left(\frac{1}{4}\right)\right]$$
$$= -[(-0{,}311) + (-0{,}5)] = 0{,}811$$

$$\text{Entropie}(a_1 = F) = -\left[\left(\frac{1}{5}\right).\log_2\left(\frac{1}{5}\right) + \left(\frac{4}{5}\right).\log_2\left(\frac{4}{5}\right)\right]$$
$$= -[(-0{,}464) + (-0{,}258)] = 0{,}722$$

$$\text{Entropie}(a_1) = \left[\left(\frac{4}{9}\right).0{,}811 + \left(\frac{5}{9}\right).0{,}722\right] = 0{,}762$$

$$\Delta_{a_1} = 0{,}991 - (0{,}762) = 0{,}229$$

$$\text{Entropie}(a_2 = T) = -\left[\left(\frac{2}{5}\right).\log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right).\log_2\left(\frac{3}{5}\right)\right]$$
$$= -[(-0{,}529) + (-0{,}442)] = 0{,}971$$

$$\text{Entropie}(a_2 = F) = -\left[\left(\frac{2}{4}\right).\log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right).\log_2\left(\frac{2}{4}\right)\right]$$
$$= -[(-0{,}5) + (-0{,}5)] = 1$$

$$\text{Entropie}(a_2) = \left[\left(\frac{5}{9}\right).0{,}971 + \left(\frac{4}{9}\right).1\right] = 0{,}984$$

$$\Delta_{a_2} = 0{,}991 - (0{,}984) = 0{,}007$$

c)

	1		3		4		5		6		7		8			
	0,5		2		3,5		4,5		5,5		6,5		7,5		splitspunt	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
-	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0
Δ	0		0,143		0,003		0,073		0,007		0,018		0,102		0	

$$\text{entropie(parent)} = 0{,}99107$$
$$\text{entropie}(\leq 2) = -\left[\left(\frac{1}{1}\right).\log_2\left(\frac{1}{1}\right) + 0.\log_2(0)\right] = 0$$
$$\text{entropie}(> 2) = -\left[\left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right) + \left(\frac{5}{8}\right).\log_2\left(\frac{5}{8}\right)\right] = 0{,}954$$
$$\Delta_3 = 0{,}991 - \left[\left(\frac{1}{9}\right).0 + \left(\frac{8}{9}\right).0{,}954\right] = 0{,}143$$

d) volgens de informatiewinst is a_1 de beste splitsing omwille van de hogere informatiewinst.

e)
$$E_{|a_1=T} = 1 - \max\left\{\frac{3}{4};\frac{1}{4}\right\} = \frac{1}{4} = 0{,}25$$
$$E_{|a_1=F} = 1 - \max\left\{\frac{1}{5};\frac{4}{5}\right\} = \frac{1}{5} = 0{,}20$$
$$\Delta_{a_1} = \frac{4}{9} - \frac{4}{9}.\frac{1}{4} - \frac{5}{9}.\frac{1}{5} = \frac{2}{9}$$

$$E_S = 1 - \max\left\{\frac{4}{9}:\frac{5}{9}\right\} = \frac{4}{9}$$

$$E_{|a_2=T} = 1 - \max\left\{\frac{2}{5};\frac{3}{5}\right\} = \frac{2}{5} = 0{,}20$$
$$E_{|a_2=F} = 1 - \max\left\{\frac{2}{4};\frac{2}{4}\right\} = \frac{2}{4} = 0{,}50$$
$$\Delta_{a_2} = \frac{4}{9} - \frac{5}{9}.\frac{2}{5} - \frac{4}{9}.\frac{2}{4} = \frac{0}{9}$$

ook hier krijgt a_1 de voorkeur want $\Delta_{a_1} > \Delta_{a_2}$.

f)

$\text{Gini}(a_1) = 0{,}344; \text{Gini}(a_2) = 0{,}4938$ ook hier krijgt a_1 de voorkeur.

Oefening 3

Stel dat je op een onbewoond eiland zit. Er groeien vele soorten paddenstoelen op het eiland, er is geen andere voedselbron. Sommige van de paddenstoelen zijn giftig, andere niet (dit is d.m.v. trial en error onderzocht door je voorgangers). Je bent de enige overlevende en hebt enkel onderstaande data om je beslissing te maken:

Example	*NotHeavy*	*Smelly*	*Spotted*	*Smooth*	*Edible*
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

Je weet dat paddenstoelen A, B en C eetbaar zijn, D tot en met H niet. Wat zou je dan durven besluiten over U, V en W? Denk eraan je leven kan er van afhangen…

a. Wat is de entropie van "`Edible`"?
b. Welk attribuut verkies je als root node van de beslissingsboom? (Je kan dit in principe zelfs visueel uit de dataset halen…)
c. Wat is de informatiewinst van het attribuut dat je in b. verkiest?
d. Bouw een beslissingsboom om paddenstoelen als giftig of niet te classificeren.
e. Wat beslis je dan over U, V en W?

Oplossing oefening 3

Voor entropie maken wij in het vervolg steeds gebruik van de letter 'H'.

a. H(Edible) = $-\sum_{i=0}^{1} P(i|t).\log_2 P(i|t) = -\frac{5}{8}.\log_2\left(\frac{5}{8}\right) - \frac{3}{8}.\log_2\left(\frac{3}{8}\right) \approx 0{,}954$

b. We gaan dit eerst mbv de nodige berekeningen oplossen.

H(NotHeavy) = $\frac{3}{8}$.H(Notheavy=0) + $\frac{5}{8}$.H(NotHeavy=1)

$$= \frac{3}{8}\left[-\frac{2}{3}.\log_2\left(\frac{2}{3}\right) - \frac{1}{3}.\log_2\left(\frac{1}{3}\right)\right] + \frac{5}{8}\left[-\frac{3}{5}.\log_2\left(\frac{3}{5}\right) - \frac{2}{5}.\log_2\left(\frac{2}{5}\right)\right]$$
$$\approx \frac{3}{8}.(0{,}390 + 0{,}528) + \frac{5}{8}.(0{,}442 + 0{,}229) \approx 0{,}344 + 0{,}607 \approx 0{,}951$$

H(Smelly) = $\frac{5}{8}$.H(Smelly=0) + $\frac{3}{8}$.H(Smelly=1)

$$= \frac{5}{8}\left[-\frac{3}{5}.\log_2\left(\frac{3}{5}\right) - \frac{2}{5}.\log_2\left(\frac{2}{5}\right)\right] + \frac{3}{8}\left[-\frac{2}{3}.\log_2\left(\frac{2}{3}\right) - \frac{1}{3}.\log_2\left(\frac{1}{3}\right)\right] \approx 0{,}951$$

(idem als de vorige)

H(Spotted) = $\frac{5}{8}$.H(Spotted=0) + $\frac{3}{8}$.H(Spotted=1)

$= \frac{5}{8}\left[-\frac{2}{5}.\log_2\left(\frac{2}{5}\right) - \frac{3}{5}.\log_2\left(\frac{3}{5}\right)\right] + \frac{3}{8}\left[-\frac{2}{3}.\log_2\left(\frac{2}{3}\right) - \frac{1}{3}.\log_2\left(\frac{1}{3}\right)\right] \approx 0{,}951$

(idem als de vorige)

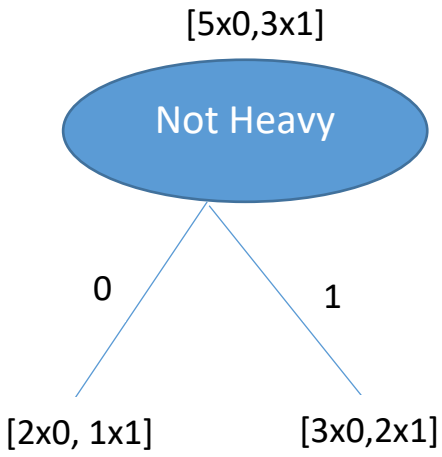
H(Smooth) = $\frac{4}{8}$.H(Smooth=0) + $\frac{4}{8}$.H(Smooth=1)

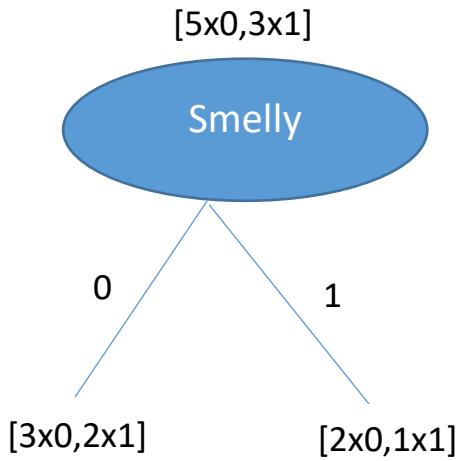
$= \frac{4}{8}\left[-\frac{2}{4}.\log_2\left(\frac{2}{4}\right) - \frac{2}{4}.\log_2\left(\frac{2}{4}\right)\right] + \frac{4}{8}\left[-\frac{3}{4}.\log_2\left(\frac{3}{4}\right) - \frac{1}{4}.\log_2\left(\frac{1}{4}\right)\right]$

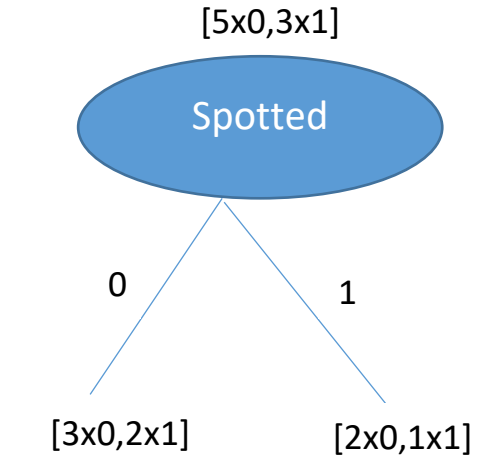
$= \frac{1}{2}.1 + \frac{1}{2}.\left(0{,}311 + \frac{1}{2}\right) \approx 0{,}906$

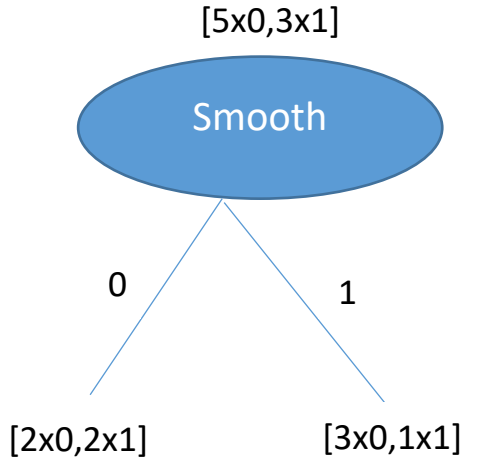
Aangezien de onzuiverheidsgraad het kleinst is bij het Smooth-attribuut, zullen we dus splitsen op Smooth.

Eigenlijk kan je dit ook meer visueel oplossen: in onderstaande figuren tellen we onder elk attribuut waarop we splitsen hoeveel keer Edible=0 en hoeveel keer Edible=1









Uit deze 4 figuren kan je gemakkelijk afleiden dat een opsplitsing op basis van Smooth aanleiding geeft tot de kleinste onzuiverheidsgraad (want de 1-tak geeft aanleiding tot de minst gelijke verdeling). Deze draagt dus onze voorkeur voor een eerste splitsing.

c. We berekenen nu de informatiewinst bij splitsing op de attributen:

Δ(NotHeavy) = H(Edible) – H(NotHeavy) ≈ 0,954 – 0,951 ≈ 0,003

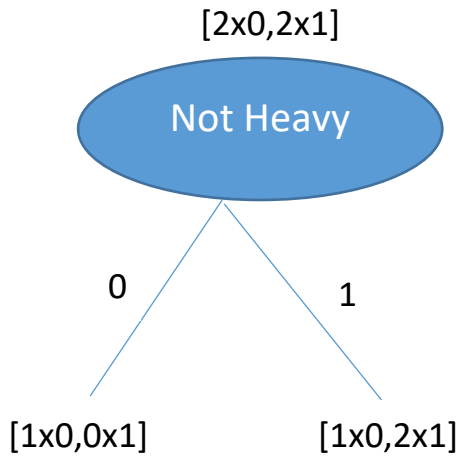
Analoog is Δ(Smelly) = Δ(Smooth) ≈ 0,003

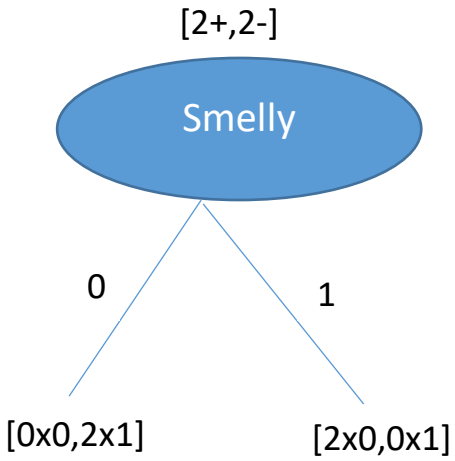
Δ(Smooth) = H(Edible) – H(Smooth) ≈ 0,954 – 0,906 ≈ 0,048

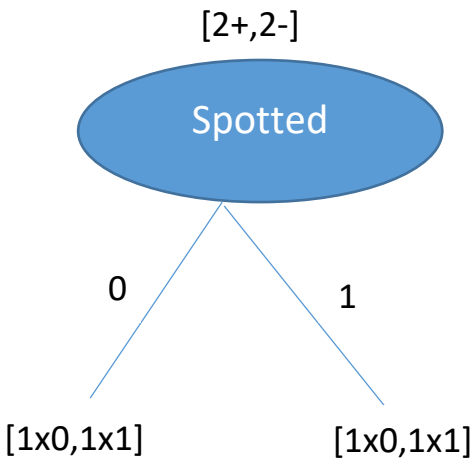
Splitsing op Smooth levert dus de hoogste informatiewinst op en is dus het attribuut waarop je best eerst splitst.

d. Na een eerste splitsing op Smooth, bepalen we de volgende splitsing louter visueel.

We bekijken eerst op welk attribuut we zullen splitsen als bij de eerste splitsing de tak **Smooth=0** gekozen werd.

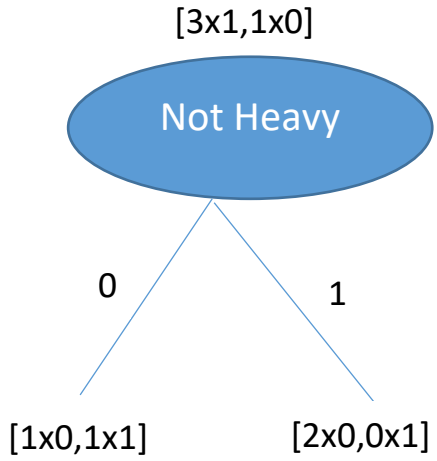


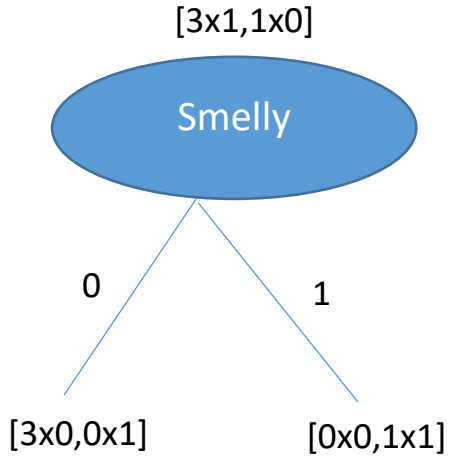


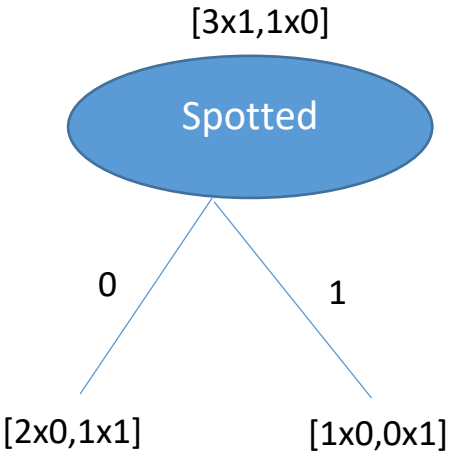


Uit deze 3 figuren kan je gemakkelijk afleiden dat je bij Smooth=0 best een opsplitsing op basis van Smelly maakt, want beide takken geven aanleiding tot leafs: als Smelly=0, dan is Edible steeds gelijk aan 1 (want Edible=0 komt 0x voor) en als Smelly=1, dan is Edible steeds gelijk aan 0 (want Edible=1 komt 0x voor).

We bekijken vervolgens op welk attribuut we zullen splitsen als bij de eerste splitsing de tak **Smooth=1** gekozen werd.

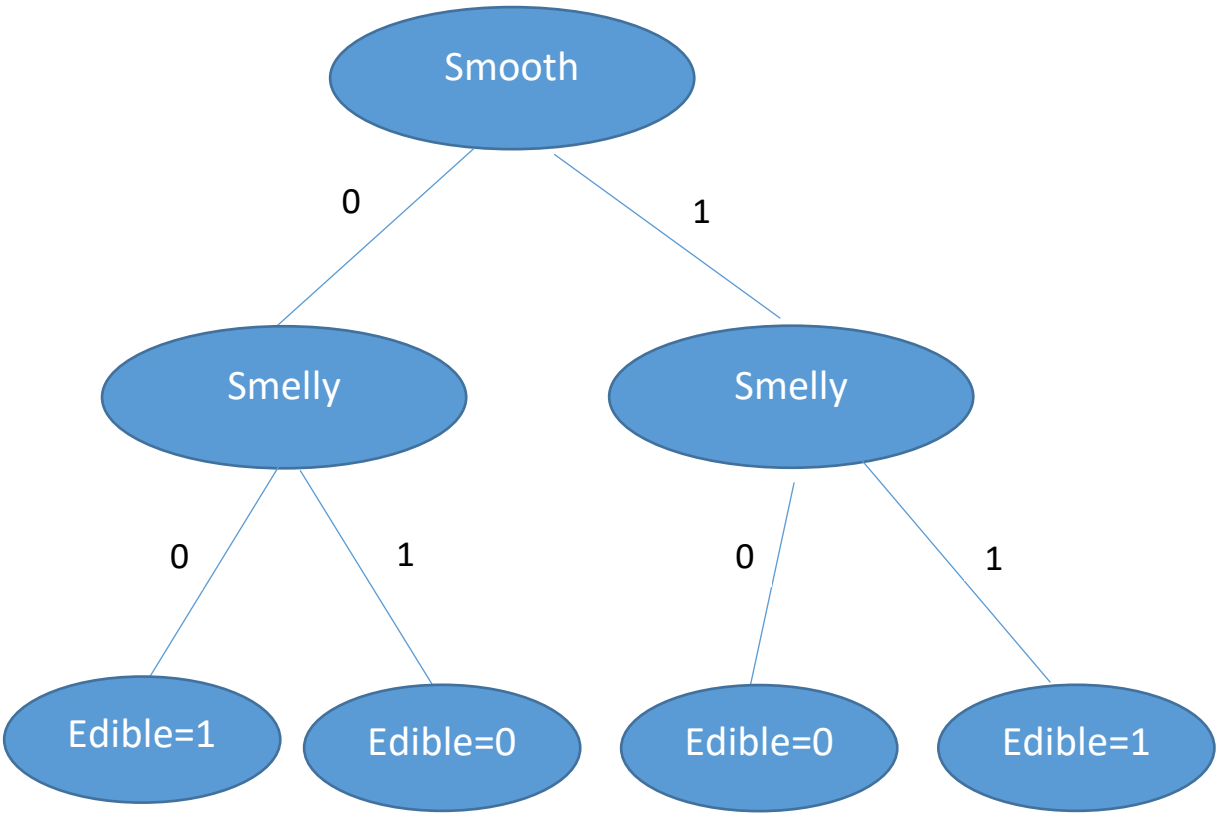






Uit deze 3 figuren kan je gemakkelijk afleiden dat je ook bij Smooth=1 best een opsplitsing op basis van Smelly maakt, want beide takken geven aanleiding tot leafs: als Smelly=0, dan is Edible steeds gelijk aan 0 (want Edible=1 komt 0x voor) en als Smelly=1, dan is Edible steeds gelijk aan 1 (want Edible=0 komt 0x voor).

De uiteindelijke beslissingsboom wordt dan:



e. We kunnen nu bovenstaande beslissingsboom gebruiken om te bepalen of de 3 paddestoelen eetbaar zijn of niet:

Voor record U: Smooth=1 en Smelly=1 => Edible = 1

Voor record V: Smooth=1 en Smelly =1 => Edible = 1

Voor record W: Smooth=0 en Smelly =1 => Edible = 0

Oefening 4 (thuis)

Van een 800-tal exoplaneten werd nagegaan of zij bewoonbaar (`habitable`) zijn of niet. Men onderzoekt dit op basis van een aantal features zoals "`Size`" en "`Orbit`". Elke rij geeft ook weer of de planeet "`Habitable`" is of niet. Zo zijn er bijvoorbeeld in de tabel 20 grote "Big" planeten die dicht "`Near`" rond hun ster draaien die "`habitable`" zijn.

Size	Orbit	Habitable	Count
Big	Near	Yes	20
Big	Far	Yes	170
Small	Near	Yes	139
Small	Far	Yes	45
Big	Near	No	130
Big	Far	No	30
Small	Near	No	11
Small	Far	No	255

Bouw een beslissingsboom op basis van deze data, maak hiervoor gebruik van de classificatiefout als informatiewinstcriterium.

Oplossing oefening 4

We gebruiken de letter 'E' als afkorting van de classificatiefout (error rate).

Wij starten met de classificatiefout te bepalen van "habitable"

$$E_{|\text{habitable}} = 1 - \max\left\{\frac{374}{800}; \frac{426}{800}\right\} = 1 - \frac{426}{800} = 0{,}467$$

Vervolgens de classificatiefout en informatiewinst voor elk van de twee attributen

Size

$$E_{|\text{Size=Big}} = 1 - \max\left\{\frac{190}{350}; \frac{160}{350}\right\} = 1 - \frac{190}{350} = 0{,}457$$

$$E_{|\text{Size=Small}} = 1 - \max\left\{\frac{184}{450}; \frac{266}{450}\right\} = 1 - \frac{266}{450} = 0{,}409$$

De informatiewinst voor "Size" is dan:

$$\begin{aligned}\Delta_{\text{Size}} &= E_{|\text{Habitable}} - \frac{350}{800}.E_{|\text{Size=Big}} - \frac{450}{800}.E_{|\text{Size=Small}} \\ &= 0{,}467 - \frac{350}{800}.0{,}457 - \frac{450}{800}.0{,}409 \\ &= 0{,}037\end{aligned}$$

Orbit

$$E_{|\text{Orbit=near}} = 1 - \max\left\{\frac{159}{300}; \frac{141}{300}\right\} = 1 - \frac{159}{300} = 0{,}47$$

$$E_{|\text{Orbit=far}} = 1 - \max\left\{\frac{215}{500}; \frac{285}{500}\right\} = 1 - \frac{285}{500} = 0{,}43$$

De informatiewinst voor "Orbit" is dan:

$$\begin{aligned}\Delta_{\text{Orbit}} &= E_{|\text{Habitable}} - \frac{300}{800}.E_{|\text{Orbit=near}} - \frac{500}{800}.E_{|\text{Orbit=near}} \\ &= 0{,}467 - \frac{300}{800}.0{,}47 - \frac{500}{800}.0{,}43 \\ &= 0{,}022\end{aligned}$$

Wij verkiezen "Size" als eerste splitsing, daarna "orbit".