Example Exam Data Mining

note that this exam may be longer than normal, since it contains more questions to learn from

Algemene Opmerkingen

- Dit is geen open boek tentamen, noch mogen er aantekeningen gebruikt worden.
- Laat bij het uitvoeren van berekeningen zien hoe je aan een antwoord gekomen bent. Als je alleen een antwoord opschrijft en dat is fout, rest ons niets dan het geheel fout te rekenen.
- Een rekenmachine is toegestaan.
- De cijfers van de nagekeken tentamens zullen binnen 4 weken op de deur van de kamer 110 gepubliceerd worden.

Opgave 1. Korte vragen

Geef korte, ter zake doende antwoorden op de volgende vragen:

(a) Explain why ZeroR is a good baseline for the accuracies of classification algorithms.
The ZeroR algorithms is a simplification of OneR that simply involves no attributes in the model, and predicts by always returning the majority class (the most common class). Since this algorithm is so simple, and only involves information about the target distribution, it can serve as a baseline for classification. Any decent classification algorithm should do better than this.

(b) The purpose of discretisation is to turn a numeric attribute into a nominal one. Give a disadvantage and an advantage of this operation.
The obvious disadvantage of discretisation is that is reduces the precision of the attribute, and hence there is loss of information. The advantage is that it makes possible the use of some algorithms, often somewhat older, that can only deal with nominal data.

(c) Give two advantages of hierarchical clustering compared to the more standard method of k-means.

- There is no need to specify the number of clusters from the start. A tree of (sub-)clusters is produced, and any cut across the tree produces a set of clusters. You can therefore choose after having obtained the model, how much detail you want to see.

- Hierarchical clustering can be applied to nominal data more easily than k-means.

(d) What is the name of the standard repository of datasets that is often used for experimenting with data mining algorithms?
The UCI repository

(e) A man needs to distribute 100 balls over 5 boxes. Explain in what situation the entropy of the distribution is the highest.
When all balls are distributed uniformly over the 5 boxes, hence 20 balls per box.

(f) Explain what two criteria are being balanced in the SD quality measure Weighted Relative Accuracy.

- The unusualness of the distribution of the target.
- The size of the subgroup.

Opgave 2. Frequent Pattern Mining

Gegeven een transactionele database met de volgende itemsets over $\{A, \ldots, E\}$:

tid	Items
1	$\{A, C, B, E\}$
2	$\{D\}$
3	$\{A, B, E\}$
4	$\{A, C\}$
5	$\{A, D, C\}$
6	$\{C, B, E\}$
7	$\{D, C\}$
8	$\{A, C, B, E\}$
9	$\{B, E\}$
10	$\{D, C\}$

(a) Our database has 5 unique items, and can be used to derive itemsets and association rules. Give the maximum numbers of itemsets and association rules that can theoretically be derived.

- Theoretical maximal number of itemsets: $2^d - 1 = 2^5 - 1 = 31$. The answer 32 would have been acceptabele here also, although an empty itemset in most cases would not be very sensical.
- Theoretical maximal number of association rules: $3^d - 2^{d+1} + 1 = 243 - 64 + 1 = 180$.

(b) Explain how the terms *maximal itemset* and *closed itemset* are defined.

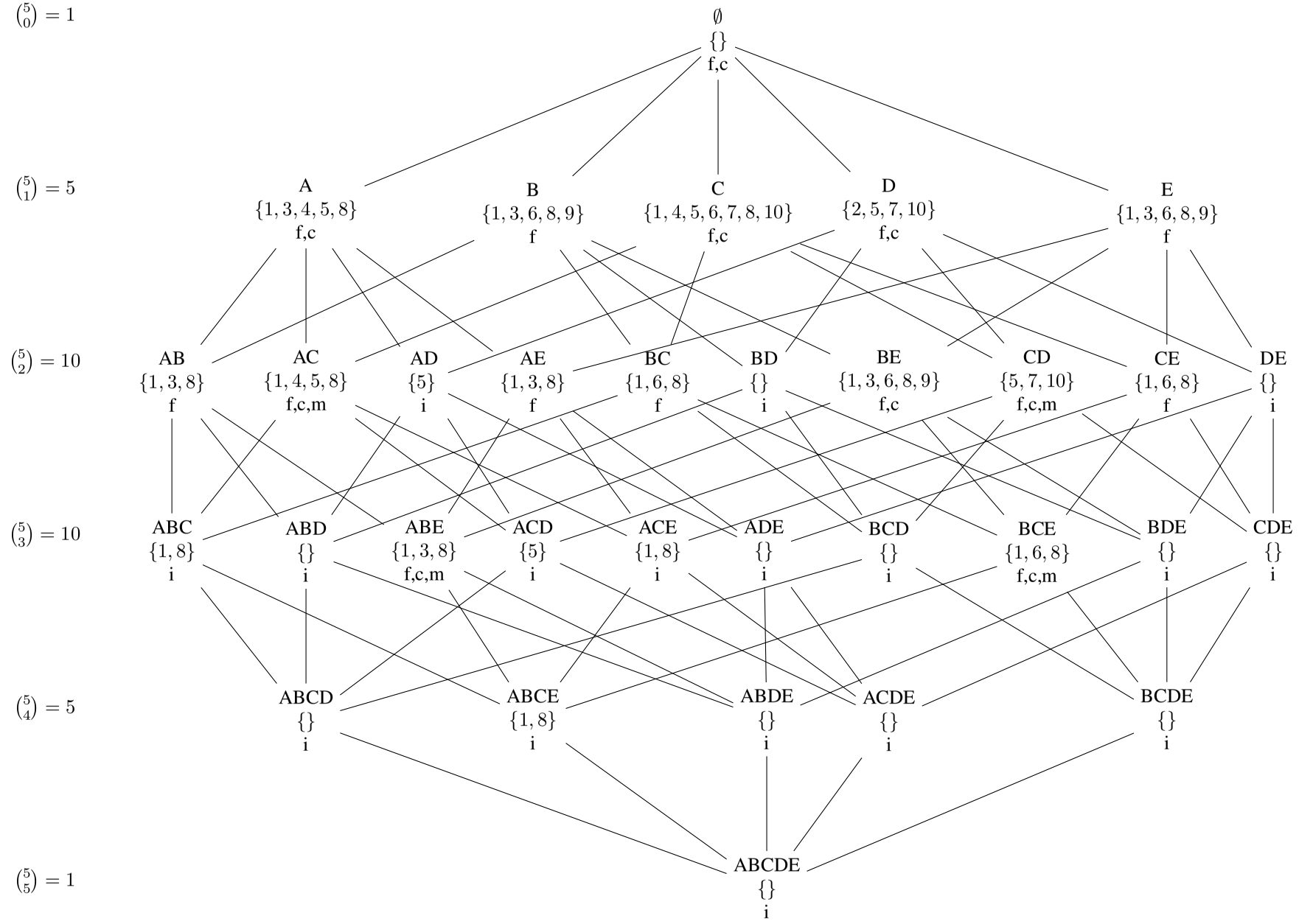
- An itemset is maximal frequent if none of its immediate supersets is frequent.

- An itemset is closed if none of its (immediate) supersets has the same support.

(c) Gegeven een minimal support $minsup = 0.3$, teken de itemset lattice en label elke knoop met minstens een van de volgende letters die van toepassing zijn: *I*= infrequent itemset, *F*=frequent itemset, *M*=maximal itemset, *C*=closed itemset.
Zie onderstaand schema voor de uitwerking. Elke knoop heeft ten minste een van de labels $\{f, i, c, m\}$, welke voorkomen in de volgende aantallen in onderstaande tabel.

label	frequentie
f	14
i	17
c	8
m	4

De labels c en m staan hier respectievelijk voor closed frequent itemset en maximal frequent itemset.
Wanneer een knoop met c of m gelabeled is, kunnen respectievelijk de f en f, c labels achterwege gelaten worden. Wanneer de infrequency border getekend is, hoeft niet elke knoop als infrequent gelabeled te worden.


$\binom{5}{0} = 1$
$\binom{5}{1} = 5$
$\binom{5}{2} = 10$
$\binom{5}{3} = 10$
$\binom{5}{4} = 5$
$\binom{5}{5} = 1$
∅ {} f,c
A {1,3,4,5,8} f,c
B {1,3,6,8,9} f
C {1,4,5,6,7,8,10} f,c
D {2,5,7,10} f,c
E {1,3,6,8,9} f
AB {1,3,8} f
AC {1,4,5,8} f,c,m
AD {5} i
AE {1,3,8} f
BC {1,6,8} f
BD {} i
BE {1,3,6,8,9} f,c
CD {5,7,10} f,c,m
CE {1,6,8} f
DE {} i
ABC {1,8} i
ABD {} i
ABE {1,3,8} f,c,m
ACD {5} i
ACE {1,8} i
ADE {} i
BCD {} i
BCE {1,6,8} f,c,m
BDE {} i
CDE {} i
ABCD {} i
ABCE {1,8} i
ABDE {} i
ACDE {} i
BCDE {} i
ABCDE {} i

Opgave 3. Distributies

(a) Stel, je moet een vragenlijst samenstellen voor een sollicitatieprocedure. Geef een voorbeeld van een ja/nee vraag met hoge entropie, en eentje van lage entropie (> 0).
"Zou je binnen een maand kunnen beginnen?"
"Heb je de juiste vooropleiding voor deze functie?"

(b) Zet de volgende drie attributen op volgorde van oplopende entropie (in het ideale geval): een numeriek attribuut, een binair attribuut, en een nominaal attribuut van 8 waarden.

- binair (laagste entropie)
- nominaal
- numeriek (hoogste entropie)

(c) Geef twee nadelen van een histogram (over een numeriek attribuut), en geef een alternatieve techniek om de distributie te bepalen.

- De grenzen van een bin kunnen te ver van elkaar liggen, waardoor er veel items in deze bin vallen, die niet meer van elkaar te onderscheiden zijn
- De grenzen van een bin kunnen te dicht bij elkaar liggen, waardoor een bin overdreven leeg is (terwijl zijn buurman te vol, bijvoorbeeld).

Een alternatieve methode is Kernel Density Estimation.

Opgave 4. Clustering

Hieronder staan gegevens van 4 (fictieve) studenten van de data mining cursus. We noteerden respectievelijk het aantal bijgewoonde lessen, hun score (op 10), het aantal dagen examenvoorbereiding, en of ze al dan niet voor het examen kwamen opdagen:

$$s_1 : (3, 1, 0, 1)$$

$$s_2 : (10, 9, 2, 1)$$

$$s_3 : (7, 9, 2, 1)$$

$$s_4 : (6, 1, 0, 1)$$

We willen deze studenten clusteren, wat we kunnen doen door middel van de k-means methode. Kies $k = 2$, en als initiële random cluster centroids:

$$c_1 : (3, 7, 1, 0) \quad c_2 : (10, 3, 1, 0)$$

Pas de k-means methode toe op D tot een maximum van 3 iteraties. Noteer voor elke iteratie welke clusters gevormd worden en wat de cluster centroids zijn. Convergeert de methode? Zo ja, motiveer.

Gebruik de standaard, Euclidische afstandmaat voor het berekenen van afstanden:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Iteratie 1

Student 1
$d(s1, c1) = \sqrt{38}$
$d(s1, c2) = \sqrt{55}$
Cluster 1

Student 2
$d(s2, c1) = \sqrt{55}$
$d(s2, c2) = \sqrt{38}$
Cluster 2

Student 3
$d(s3, c1) = \sqrt{22}$
$d(s3, c2) = \sqrt{47}$
Cluster 1

Student 4
$d(s4, c1) = \sqrt{47}$
$d(s4, c2) = \sqrt{22}$
Cluster 2

New centroids
$c1 = (5, 5, 1, 1)$
$c2 = (8, 5, 1, 1)$

Iteratie 2

Student 1
$d(s1, c1) = \sqrt{21}$
$d(s1, c2) = \sqrt{42}$
Cluster 1

Student 2
$d(s2, c1) = \sqrt{42}$
$d(s2, c2) = \sqrt{21}$
Cluster 2

Student 3
$d(s3, c1) = \sqrt{21}$
$d(s3, c2) = \sqrt{18}$
Cluster 2

Student 4
$d(s4, c1) = \sqrt{18}$
$d(s4, c2) = \sqrt{21}$
Cluster 1

New centroids
$c1 = (4.5, 1, 0, 1)$
$c2 = (8.5, 9, 2, 1)$

Iteratie 3

Indeling en centroids van beide clusters blijft gelijk.
Dus de methode is geconvergeerd.

Opgave 5. Maximally Informative k-Itemsets

Gegeven de volgende dataset van binaire attributen:

A	B	C	D
1	0	0	1
1	0	0	1
1	1	1	0
1	1	1	0
1	1	0	0
0	1	0	0
0	0	1	1
0	0	1	1

(a) Geef de entropie van elk van de 4 attributen (over de hele dataset). Gebruik eventueel de onderstaande tabel met benaderde waarden voor de entropie $H(p)$.

p	$H(p)$
0	0
1/8	0.54
2/8	0.81
3/8	0.95
4/8	1
5/8	0.95
6/8	0.81
7/8	0.54
1	0

Merk op dat de entropie $H(p)$ in de tabel gebruik maakt van het logaritme met basis 2, ook wel $\log_2$ of lg.

- $H(A) = 0.95$
- $H(B) = 1$
- $H(C) = 1$
- $H(D) = 1$

(b) Bewijs dat $\{C, D\}$ een miki is voor $k = 2$ en geef de joint entropy. Is het de enige miki van 2 attributen?

$\{C, D\}$ is een k-itemset met $k = 2$, dus dit levert maximaal $2^k = 2^2 = 4$ combinaties. Elk van deze combinaties komt even vaak voor: twee keer. De joint entropy van $\{C, D\}$ is dus 2 bits, want er zijn twee items. Daarmee is deze itemset dus een miki, want een hogere joint entropy is niet te bereiken.
$\{B, C\}$ is ook een miki, het deelt de database, net als $\{C, D\}$, op in 4 gelijke stukken.

(c) Geef een bovengrens van $H(\{B, C, D\})$ op basis van de hierboven berekende joint entropy van $\{C, D\}$. Bepaal daarna de exacte waarde van $H(\{B, C, D\})$.
De bovengrens van $H(\{B, C, D\})$ wordt gevonden met behulp van de relatie:

$$\begin{aligned} H(\{B, C, D\}) &\leq H(\{C, D\}) + H(\{B\}) \\ &= 2 + 1 \\ &= 3. \end{aligned}$$

In dit specifieke geval is de bovengrens dus gelijk aan het theoretisch maximum:

$$lg(2^k) = lg(2^3) = lg(8) = 3.$$

B is de inverse van D, en daardoor zal B geen informatie toevoegen aan $\{C, D\}$.
De exacte waarde voor $H(\{B, C, D\})$ is dus:

$$H(\{B, C, D\}) = H(\{C, D\}) + 0 = 2 + 0 = 2.$$

(d) Stel, de tabel wordt gesorteerd op A, B. Wat is de invloed hiervan op de entropie van $\{C, D\}$?
Sorteren heeft geen invloed op de (joint) entropie. Entropie is gevoelig voor verandering in verdeling of frequenties, deze veranderen niet wanneer de database gesorteerd wordt.

(e) Geef de joint entropie $H(\{A, B, C, D\})$.
De volgende combinaties komen voor:
1,0,0,1 (2 keer)
1,1,1,0 (2 keer)
1,1,0,0 (1 keer)
0,1,0,0 (1 keer)
0,0,1,1 (2 keer)
De joint entropy is dus

$$\begin{aligned} -2/8lg(2/8) - 2/8lg(2/8) - 1/8lg(1/8) - 1/8lg(1/8) - 2/8lg(2/8) = \\ 0.5 + 0.5 + 0.375 + 0.375 + 0.5 = 2.25 \end{aligned}$$

Opgave 6. Regressie

(a) Zowel Regression Trees als Model Trees kunnen gebruikt worden om regressie te doen. Leg uit wat het grote verschil tussen deze methoden is.
Bij Model Trees worden er in de bladeren nog lineaire modellen gebruikt, terwijl bij Regression Trees er alleen constante waarden in de bladeren staan.

(b) Welke waarden kan R^2 aannemen, en hoe moeten deze waarden geïnterpreteerd worden?
De waarde ligt tussen de 0 en 1, waarbij 0 betekent dat het model niets bijdraagt, terwijl bij de waarde 1 sprake is van een perfecte fit.

(c) In welke gevallen zou je eerder kiezen voor bomen (regressie of model) dan voor lineaire modellen? Geef twee redenen.

- Als er sprake is van duidelijk niet-lineaire relaties of data waarbij de target duidelijk afhangt van een drempelwaarde voor (één van) de voorspellende attributen.
- Als er nominale attributen beschikbaar zijn.

(d) Binnen subgroup discovery kun je ook een regressie setting kiezen, maar daar is de maat R^2 nog niet ingeburgerd. Geef een andere maat die geschikt is voor regressie bij SD.
z-score

(e) Het algoritme $M5'$ bouwt bomen die uitsluitend uit binaire splitsingen bestaan. Leg uit hoe dit werkt in het geval van een nominaal splitsingsattribuut.
Voordat het modelleren begint, worden de verschillende waarden van het nominale attribuut gesorteerd op basis van de gemiddelde waarde van de target binnen elke nominale waarde. Voor k nominal waarden zijn er $k-1$ splitsingen in twee groepen te bepalen. Elk van deze splitsingen wordt nu aan de dataset toegevoegd door middel van een (binaire) indicator-variabele.

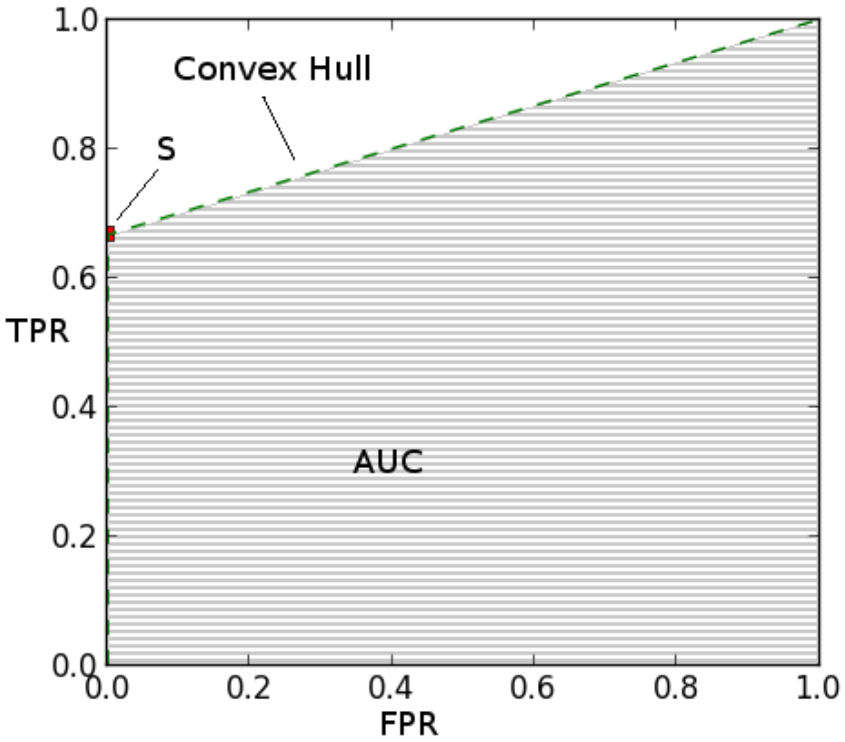
Opgave 7. Case-Study

Medici doen onderzoek naar een nieuw virus dat in een ziekenhuis uitbreekt onder geopereerde patiënten. Er bestaat het vermoeden dat oudere patiënten gevoeliger zijn dan jongere. In totaal worden 536 patiënten getest. De mediaan van de leeftijd ligt bij 53,5 (precies tussen twee personen van 53 en 54 in), dus 'oud' wordt gedefinieerd als $l \equiv \textit{leeftijd} \geq 53{,}5$. 402 patiënten blijken het virus te hebben ($\textit{besmet} = T$), waarvan 2/3 'oud'.

(a) Teken de kruistabel (contingency table, confusion matrix) van l en *besmet*.

leeftijd \ besmet	ja	nee	totaal
oud	268	0	268
jong	134	134	268
totaal	402	134	536

(b) Bekijk de subgroup S van oude mensen. Teken de ROC ruimte, en geef daar in aan waar S te vinden is.



(c) Teken in deze ruimte ook de convex hull die opgespannen wordt door deze enkele subgroup S. Wat is de waarde van de Area Under the Curve (AUC)?

$$\text{AUC} = \frac{5}{6} = 0.833$$

(d) Geef de formule van WRAcc, en bereken de WRAcc van S.

$$\begin{aligned}\text{WRAcc} &= P(S,T) - (P(S) \times P(T)) \\ &= \frac{1}{2} - \left(\frac{1}{2} \times \frac{3}{4}\right) \\ &= \frac{1}{8} \\ &= 0.125\end{aligned}$$

(e) Naast leeftijd kan ook het geslacht een bepalende factor zijn. Geef aan waar de subgroup M van *oude mannen* kan liggen. Als je S rapporteert als interessante subgroup, heeft het dan nog zin om M te rapporteren? Kan M een hogere WRAcc hebben?
M is een deelverzameling van S, dus zal ergens linksonder S in de ROC ruimte liggen. Aangezien S op de y-as ligt, kan M alleen maar (op of) direct onder S liggen. Daarmee zal M dus even ver of verder van ROC-heaven liggen, waardoor het

een lagere WRAcc zal hebben. Hierdoor is M niet interessant om te rapporteren, als S al gerapporteerd is.

(f) Bereken de information gain van een splitsing op *besmet*. Gebruik eventueel de tabel met benaderde waarden voor entropie.

$$\begin{aligned} IG(T, besmet) &= H(T) - H(T|besmet) \\ &= H(3/4) - \left(\frac{1}{2} \times H(1)\right) - \left(\frac{1}{2} \times H(1/2)\right) \\ &= 0.81 - \left(\frac{1}{2} \times 0\right) - \left(\frac{1}{2} \times 1\right) \\ &= 0.31 \end{aligned}$$

Opgave 6. Decision Trees

(a) Pruning is een manier om een gebouwde beslisboom weer te snoeien, zodat overfitting voorkomen kan worden. Leg uit waarom in de meeste beslisboomalgoritmen eerst een te grote boom gebouwd wordt, terwijl er daarna toch vaak weer gesnoeid gaat worden.
Van tevoren kan niet ingeschat worden welke takken waardevolle informatie opleveren en welke niet, omdat combinaties van beslissingen meer kunnen opleveren dan individuele beslisssingen.

(b) Geef van de volgende statements aan of ze waar zijn

1. Information gain van een gegeven attribuut neemt altijd toe naarmate je dieper in de boom komt.
   Onwaar.
2. Information gain van een gegeven attribuut neemt altijd af naarmate je dieper in de boom komt.
   Onwaar.
3. Information gain van een gegeven attribuut kan toenemen naarmate je dieper in de boom komt.
   Waar.
4. Information gain van een gegeven attribuut is altijd ≥ 0.
   Waar.
5. Information gain van een gegeven attribuut is altijd ≤ 1.
   Onwaar, bijvoorbeeld bij een niet-binaire target.

Ga uit van een dataset met drie binaire attributen (A, B en C) en een binair target D. De dataset is als volgt:

A	B	C	D
0	1	0	1
0	0	1	1
1	0	0	1
0	1	0	0
1	1	0	0
1	1	1	1
0	1	1	0
1	1	1	0

(c) Geef een beslisboom van diepte 1 (dus maximaal één splitsing) zoals die door C4.5 op basis van information gain geproduceerd zal worden.

START

$$
\begin{array}{lll}
H(T_0) & = 1 & \\
H(T_0|\text{split op A}) & = (0.5 \times 1) + (0.5 \times 1) & = 1 \\
H(T_0|\text{split op B}) & = (0.25 \times 0) + (0.75 \times H(1/3)) & \approx 0.689 \\
H(T_0|\text{split op C}) & = (0.5 \times 1) + (0.5 \times 1) & = 1
\end{array}
$$

Dus de eerste split is op B, met $IG(T_0, B) \approx 0.311$.

T

$B = 0$ $\rightarrow D = 1$ $\quad$ $B = 1$ $\rightarrow D = 0$

(d) Stel dat we een nieuwe numeriek target E introduceren, dat functioneel afhankelijk is van $A, \ldots, D$, namelijk $E = A + B + C + D$. Leg uit hoe een model tree geïnduceerd op deze dataset er uit zal zien.
Bij de eerste mogelijke splitsing wordt bepaald dat een perfecte lineaire regressie van E mogelijk is, namelijk $E = A + B + C + D$, dus er wordt geen splitsing uitgevoerd, en de 'boom' is slechts een rootnode met lineair model.