Statistiek: Entropie

12/30/2013

dr. Brenda Casteleyn

1. Theorie

Vooraf: regels van logaritmes: dit is de macht waartoe een bepaald grondgetal verheven is.

$\log_a x = b$ als $x^b = a$ (want b is de macht waartoe je x moet verheffen om het grondgetal a te bekomen)

Rekenregels:

$\log_a (x.y.z) = \log_a x + \log_a y + \log_a z$

$\log_a(\frac{x}{y}) = \log_a x - \log_a y$

$\log_a x^n = n.\log_a x$

$\log_a \sqrt[n]{x} = 1/n.\ \log_a x$

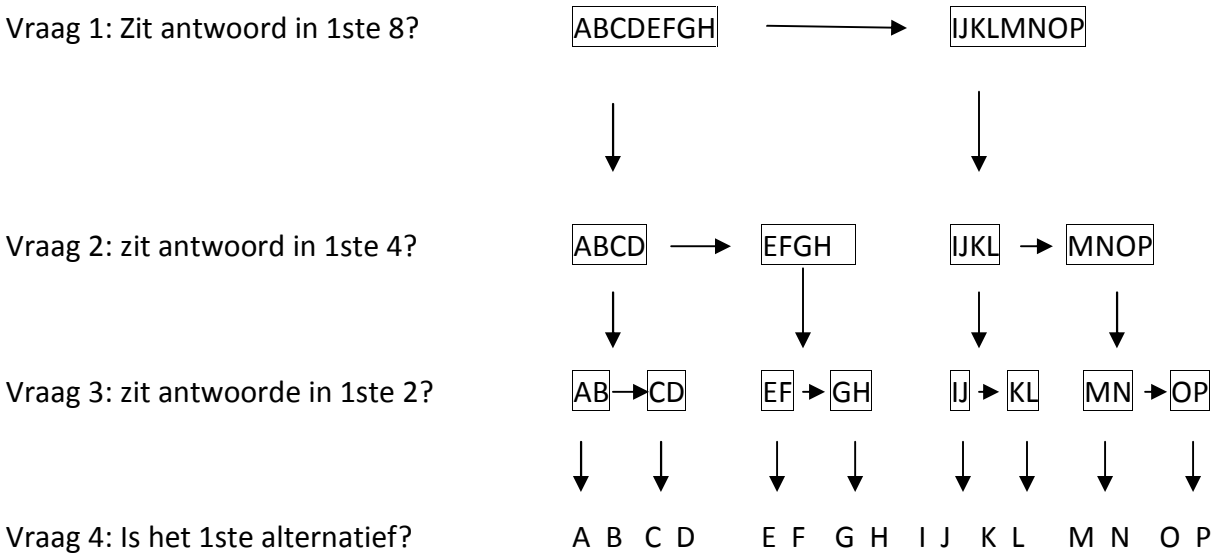
$\log_a x = \frac{1}{\log_x a}$

Omzetting naar ander grondgetal:

$\log_a x = \frac{\log_b x}{\log_b a}$

Entropie in informatietheorie

= hoeveel vragen moet je stellen om uit een aantal alternatieven het antwoord te vinden. B.V. je hebt 16 alternatieven, nl. ABCDEFGHIJKLMNOP. Je moet 4 vragen stellen om te weten welk alternatief er gekozen is (pijltjes naar rechts betekenen 'nee', pijltjes naar onder bekekenen 'ja').



Of in termen van logaritmes: $\log_2 16 = 4$

Algemeen: O[X]= $\log_2 k$ (en k is het aantal categorieën)

Wanneer niet elk alternatief evenveel kans heeft om gekozen te worden, maken we het gewogen gemiddelde van de onzekerheden die bij de verschillende categorieën horen. Als wegingscoëfficiënten nemen we de relatieve frequenties. Formule voor entropie H:

$$H = -\sum_{i=1}^{k}(f_i^* . log_2 f_i^*)$$

En omdat natuurlijke logaritmen gemakkelijker zijn om mee te werken vervangen we door ln met behulp van formule om van basis te veranderen. We krijgen dan:

$$H = \frac{-\sum_{i=1}^{k}(f_i^* . ln f_i^*)}{\ln 2}$$

Dit kunnen we echter niet goed interpreteren, dus we willen een minimum en een maximum bepalen. De entropie is maximaal als elke categorie evenveel kans heeft:

$$H = -\sum_{i=1}^{k}\frac{1}{k} . log_2 \frac{1}{k} = \log_2 k$$

(dus: ons oorspronkelijke formule H = $\log_2 k$. Daarom delen we dus door deze waarde om een genormeerde entropie te vinden, die varieert tussen 0 en 1):

$$H = \frac{-\sum_{i=1}^{k}(f_i^* . log_2 f_i^*)}{\log_2 k}$$

En omdat natuurlijke logaritmen gemakkelijker zijn om mee te werken dan logaritme met grondgetal 2 vervangen we $\log_2$ door ln door gebruik te maken van formule om van basis te veranderen. Na wat rekenwerk vinden we:

$$H = \frac{-\sum_{i=1}^{k}(f_i^* . ln f_i^*)}{\ln k}$$

2. Oefeningen

Examen Thijssen 2012 Vraag 1

Wat is de maximale waarde van de entropie met k categorieën?

A. $-\log_2 k$

B. $\log_2 k$

C. ln 2

D - ln 2

Fictieve oefening

Bereken de entropie voor gegevens weergegeven in de volgende tabel:

x_i	f_i
10	5
20	10
30	3
40	7
Totaal	25

3. Oplossingen

Examen Thijssen 2012 Vraag 1

Wat is de maximale waarde van de entropie met k categorieën?

➔ Antwoord B

Fictieve oefening

In de onderstaande tabel werd f_i en $\ln(f_i)$ uitgewerkt: f_i^* vinden we door f_i te delen door N (= 25).

x_i	f_i	f_i^*	$\ln(f_i^*)$	$f_i \ln(f_i^*)$	$(f_i^{*\wedge} f_i^*)$
10	5	0,2	-1,60944	-0,32189	0,72478
20	10	0,4	-0,91629	-0,36652	0,693145
30	3	0,12	-2,12026	-0,25443	0,775357
40	7	0,28	-1,27297	-0,35643	0,700171
Totaal	25			-1,29927	0,272732
					-1,29927

Om de entropie te berekenen hebben we volgende formule nodig:

$$H = \frac{-\sum_{i=1}^{k}(f_i^*.lnf_i^*)}{\ln k}$$

De teller vinden we door de som te maken van de voorlaatste kolom: 1,29927

De noemer is ln(4) = 1,386294

H = 1,29927/1,386294 = 0,93722

In de laatstekolom zien we de berekening van $f_i^{f_i}$ en onderaan ln van het product: $\ln(\prod_{i=1}^{k} f_i^{f_i})$ Deze uitkomst is gelijk aan de som $\sum_{i=1}^{k}(f_i^*.lnf_i^*)$ (door gebruik te maken van de derde rekenregel). In principe kan je dus kiezen of de berekening maakt via de vierde kolom of de derde kolom.