Informatie theorie

Sebastiaan Polderman
0820738

Paul Sohier
0806122

11 februari 2011

Hoofdstuk 1

Informatie

1.1 Opdacht 1

Een roulettespel heeft een draaischijf met 38 genummerde vakjes: 18 rode, 18 zwarte en 2 groene vakjes. Op de draaiende schijf wordt een balletje geworpen. Als de schijf tot rust komt, zal het balletje in één van de vakjes blijven liggen. Elk vakje heeft evenveel kans om het balletje te vangen. De zwarte vakjes zijn oneven genummerd van 1, 3, 5, . . . , 35, de rode vakjes zijn even genummerd van 2, 4, 6, . . . , 36 en de twee groene vakjes hebben de 'nummers' 0 en 00. Zodra de croupier de kleur of het nummer van het winnende vakje genoemd heeft, mag er niet meer ingezet worden.

(a) *Hoe groot is de selectieve informatie over elke kleur van het vakje?* 1 bit
$ld(1/p)$ waar p de kans op een kleur is.

Rood	$ld(38/18)$	1,078 bits
Zwart	$ld(38/18)$	1,078 bits
Groen	$ld(38/2)$	4,248 bits

(b) *Hoe groot is de gemiddelde informatie over de kleur van het vakje?* 1 bit
Het complement van de som $p_i \cdot ld(p_i)$ voor iedere i

$$(18/38) \cdot ld(18/38) + (18/38) \cdot ld(18/38) + (2/38) \cdot ld(2/38)) = 1.245 bits$$

(c) *Hoe groot is de gemiddelde informatie over het nummer van het vakje?*
$(2 * 6 + 16 * 5 + 8 * 4 + 4 * 3 + 2 * 2 + 1 * 1)/38 = 3.71 bit$

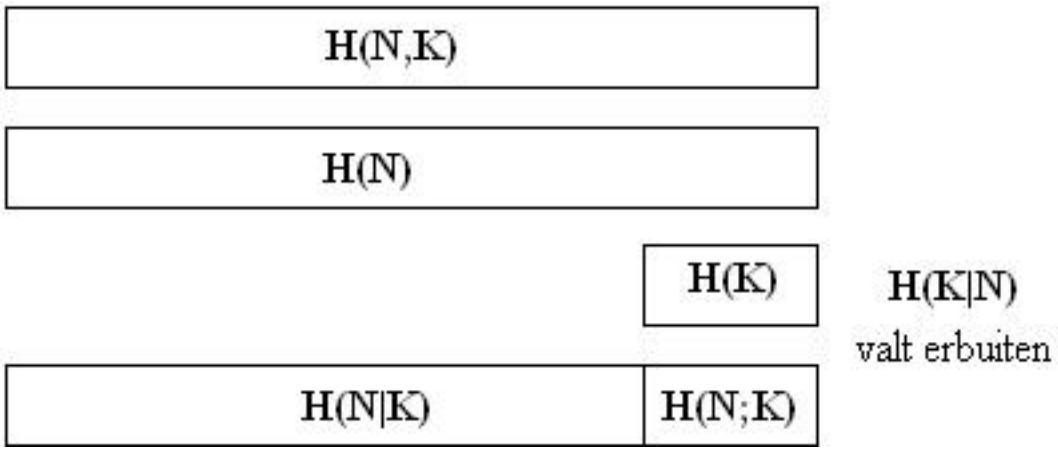
Het complement van de som $p_i \cdot ld(p_i)$ voor idere i.

Er zijn 38 mogelijke getallen, met iedere een gelijke kans: $(-(1/38) \cdot ld(1/38)) * 38 = ld(38) = 5.248 bits$

(d) *Hoe groot is de conditionele entropie (equivocatie) van de kleur als het nummer van het vakje al bekend is?*

Als je het nummer weet, weet je ook de kleur. in andere woorden: De kleur is volledig afhankelijk van het getal. Conditionele entropie is daarom 0

(e) *Hoe groot is de conditionele entropie (equivocatie) van het nummer als de kleur van het vakje al bebend is?*



1.2 Opdracht 4

Volgens bijlage D komen in normale teksten korte woorden frequenter voor dan lange woorden. In het 'Groene boekje' woordenlijst van de Nederlandse taal) blijkt dat niet te kloppen. Geef hier een verklaring voor.

De drie bestaande lidwoorden zijn veel gebruikte woorden in de Nederlandse taal. Lidwoorden worden vaker gebruikt als "normale" woorden, waardoor de verhouding hierdoor al niet klopt. Maar bijvoorbeeld ook werkwoorden en ook persoonsvormen zijn gemiddeld genomen korter. Er zijn dus wel meer verschillende lange woorden, maar die worden minder "hergebruikt" als de eerder genoemde korte woorden.

1.3 Opdracht 5

*Binare Coded Decimals ('BCD') is een code waarbij een getal van 0, 1, 2, . . . , 99 in een 8 bit-woord (een '*byte*') gecodeerd wort. Hoeveel redundantie bevat zo'n BCD-woord?*

100 mogelijke BCD waarde, $H(X) = ld(100) = 6.64$ 256 combinaties met een byte, ofwel 8 bits.

$8 - 6.64 = 1.36$ bits

Hoofdstuk 2

Codesystemen voor storingsvrije omgevingen

2.1 Opdracht 1

Een bron genereert een onafhankelijke rij symbolen uit het alfabet $\{0, 1\}$*. De kans op een 0 is* $0,9$*, ed kans op een 1 is* $0,1$*. De rij getallen wordt met een 'Run-Length codering' gecodeerd tot een rij met symbolen uit het alfabet* $\{a, b, c, d, e, f, g, h, i\}$*. Tenslotte wordt deze 'run length' codewoorden weer gecodeerd met een hoffmancode.*

b-code	r-code
1	a
01	b
001	c
0001	d
00001	e
000001	f
0000001	g
00000001	h
00000000	i

(a) Bereken de entropie van de bron.
$H(X) = -0.9 * ld(0.9) - 0.1 * ld(0.1) = 0,469$

(b) Bereken het gemiddeld aantal bronsymbolen per run-length codewoord.
$45/9 = 5$ bronsymbolen

(c) Bereken het gemiddeld aantal bit van de Huffmancode per run-length woord. $ld(9) = 3.2$ bits

(d) Bereken het gemiddelde aantal bronsymbolen b *per Huffmansymbool* $5/ld(9) = 1.577$ per huffmansymbool.

2.2 Opdracht 4

*Waarom zouden de makers van '*gzip*' de blokverwijzingen beperkt hebben?*

Wanneer de blokverwijzing groter zou zijn dan zou de redundantie in de blokverwijzing optreed zo groot kunnen worden dat de gecomprimeerde data groter is als het orgineel.

2.3 Opdracht 5

In het DNA van de bacterie Micrococcus Lysodeiktus *hebben de basen* A, C, T, G *de volgende waarschijnlijkheid:* $P(A) = P(T) = 29/200$ *en* $P(C) = P(G) = 71/200$. *Bij de bacterie E.Coli is deze verdeling:* $P(A) = P(T) = P(C) = P(G) = 1/4$. *Welke bacterie zou van de twee het meest complexe organisme zijn?*

Door de gelkere verdeling van E.coli lkt erop dat het DNA meer gecomprimeerd is waardoor de gegevens complexer zn.

2.4 Opdracht 6

*Het decompressie-algoritme van het programma '*bzip2*' maakt gebruik van een inversie permutatie* $T(i)$ *om uit de laatste kolom de bronrij terug te vinden. Wat is er fout aan de volgende redenatie om de bronrij terug te vinden uit tabel 2.5, gegeven dat de rij-index van de bronrij de waarde 3 heeft?*

> Men kan uit de tabel 2.5 aflezen dat de onbekende bronrij moet beginnen met een '*o*' (Rij 3, kolom 2) en eindigen met een '*b*' (rij 3 kolom 1). Vervolgens blijkt de rij bronsymbolen uit alle andere tweetallen te zijn opgebouwd voor de tussenliggende sumbolen: '*ob*', '*ro*', '*oo*', '*rr*', '*or*'. Wij weten alleen niet in welke volgorde. Wel is bekend dat elk tweetal exact één keer gebruikt moet worden.
>
> Met deze gegevens kunnen wij de bronrij herstellen: Start met '*o*' dan zijn er twee volgende letters mogelijk: '*b*' en '*o*' (vanwege '*bo*' en '*oo*') De rij '*ob*' heeft geen opvolger en is dus geen oplossing.

De rij '*oo*' heeft als uitbreiding '*oor*' (Vanwege '*or*'). Dit passen en meten kunnen wij herhalen tot wij de orginele bronrij '*oorrob*' gevonden hebben.

In deze omschrijving word bij de eerste stap de mogelijkheid '*or*' vermeden waardoor er niet uit blijkt dat er 2 manieren zijn van oplossen namelijk de oplossing '*oorrob*' waar op uitgekomen is dmv deze uitleg en de oplossing '*orroob*' als je ook '*or*' als optie geeft bij mogelijkheid 1.

Hoofdstuk 3

Coderingen voor storingsrijke omgevingen

3.1 Opdracht 1 *

Een BSC heeft een Binary Error Rate *van 0,3. Wat is de discrete kanaalcapaciteit van dit kanaal?* $1 + 0,3 * ld(0,3) + (1 - 0,3) * ld(1 - 0,3) = 0,11870910\ldots$

3.2 Opdracht 2

Noem 3 decodeerprincipes en hun eigenschappen.

MAP Maximum-a-Posteriori is een decodeerprincipe die gebruik maakt van over het algemeen zo laag mogelijke decodeer fout te houden. Het decodeerprincipe kiest een zo grootmogelijke x_i uit $P(X = x_i|Y = y_i)$ waardoor het codewoord y_i het meest in de buurt komt.

ML Maximum Likelihood is een decodeerprincipe die voornamelijk gebruikt wordt bij het decoderen van uniform verdeelde bronworden. Om dit te doen dient je een x_i zo hoog mogelijk te hebben bij $P(Y = y_i|X = x_i)$ (y_i is in dit geval het code woord)

MD Minimum Distance is een decodeerprincipes kiest een codewoord x_i wat zo dicht mogelijk bij y_i ligt qua symbolen. Het wordt voornamelijk gebruikt bij cyberteksten waar genoeg redudantie in is hierdoor kan er voldoende afstand tussen de code woorden worden bereikt.

3.3 Opdracht 3

De *Soundex codering* is een codering die bronwoorden uit een West Europesche spreektaal vertaalt in codewoorden van één letter gevolgd door drie cijfers. Het voordeel van deze codering is dat de woorden die veel op elkaar lijken qua uitspraak, dezelfde code krijgen. Soundex wordt veel toegepast in spellingscontrole in woordproccesssoren, reisplanners en reserveringssystemen etc. Bijvoorbeeld, de namen *'brok', 'brock', 'broek'* geven dezelfde code B620. Daarentegen geven de namen *'jansen', 'janssen', 'jansens'* de *J525*.

Het soundex-algoritme werkt als volgt:

(a) *De eerste letter van het bronwoord wordt de beginletter in het Soundex codewoord*

(b) *De volgende 3 cijfers komen uit de volgende tabel. Zijn worden in volgorde opgebouwd in volgorde van de letters in het bronwoord. Als het bronwoord te kort is voor een volledig Soundex codewoord, wordt het Soundex codewoord aangevuld met nullen. Te lange codewoorden worden afgebroken na drie cijfers.*

code	letter	uitspraakorgaan
1	b p f v	lippen
2	c s k g j q x z	keel
3	d t	tanden
4	l	tong voor
5	m n	neus
6	r	tong achter
geen	a e h i o u y w	

Deze tabel geeft de cijfercodering van de letters aan. De meeste medeklinkers zijn volgens uitspraak gegroepeerd. De klinkers en de zachte medeklinkers krijgen geen cijfercode. Hoewel het Soundex-algoritme goed werkt voor West-Europese talen, is het niet voor andere spreektalen geschikt.

(a) *Welke aspecten maken Soundex codering geschikt voor Weste-Europese talen?* Deze coderingen is alleen geschikt voor westerse talen omdat het gebruik maakt van klanken die over het algemeen voor komen in de westerse omgeving. In andere gebieden kunnen minder klanken zijn of zijn er veel meer letters met de zelfde klank.

(b) Een andere manier om alternatieve woorden te vinden is de '*Edit Distance*':

- Spatie tussenvoegen;

- Twee buurletters verwisselen;
- Een letter vervangen door een andere;
- Het verwijderen van een letter;
- Een letter toevoegen.

*(*Wat is het belanrijkste verschil tussen de 'Edit Distance' en Soundex?) Het belangrijkste verschil is in dit geval dat Soundex een vaste lengte heeft. bij de andere manieren die hier boven staan is er geen vaste lengte gespecificeert.

3.4 Opdracht 4

Gegeven een taal met 8 bronwoorden van 3 bit.

(A) *Hoeveel redundantie moeten de codewoorden bevatten om een taal met 8 bronwoorden van ieder 3 bit intolerant voor 2 bit fouten te maken?*

Bronwoord	Tolerantie toevoeging	Codewoord
000	1111	0001111
001	0011	0010011
010	1010	0101010
011	1100	0111100
100	0000	1000000
101	0110	1010110
110	0101	1100101
111	1001	1111001

(B) *Welke CRC-polynoom zou in aanmerking komen om de bronwoorden te beschermen tegen 1 en 2 bit fouten?* Een CRC-6 is voldoende want er dienen alleen maar 1 en 2 bit foute gedetecteert te worden.

3.5 Opdracht 5

Een geheugenloos kanaal tussen X *en* Y *heeft de volgende eigenschappen* $P(Y = y|X = x) = 0,5$

(A) *Bepaal de discrete kanaalcapaciteit van het kanaal in figuur 3.5?* $H(Y|X) = -8 * (0.25 * 0.25 * ld(0.5)) = 0.5\text{bit}$

(B) *Ontwerp een transmissiecode waarmee via dit kanaal foutloze transmisse plaats kan vinden.*

Als je het kanaal in 2 delen opdeeld kan je makkelijk een 1 en 2 verzenden, waardoor binair geen probleem is.

Hoofdstuk 4

Cryptografie

4.1 Opdracht 1

Probeer het volgende Ceasargecodeerde bericht teontcijferen:
LNNZDPLNCDJHQLNRYHUZRQ

Het ontcijferen van dit Ceasargecodeerde bericht kan onder andere op de volgende manier door alle mogelijke uitkomsten op te schrijven tot dat je een leesbaar bericht tegen komt.

0	LNNZDPLNCDJHQLNRYHUZRQ
1	KMMYCOKMBCIGPKMQXGTYQP
2	JLLXBNJLABHFOJLPWFSXPO
3	IKKWAMIKZAGENIKOVERWON

In dit geval is de meest logische uitkomst: Ik kwam ik zag en ik overwon

4.2 Opdracht 2

Geef enkele voorbeelden waaruit blijkt dat berichten in het algemeen niet uniform verdeeld zijn. Welke oplossing is geschikt om deze berichten uniform verdeeld te maken?

Om een bericht uniform te verdelen is er onderandere een mogelijkheid om het bericht in te pakken ofwel comprimeren dit heeft als gevolg dat elk teken of teken reeks gemiddeld even vaak voorkomt.

4.3 Opdracht 3

Een andere manier om geheime boodschappen te versturen is steganografie.

(a) *Wanneer zouden partijen steganografie gebruiken?*
Als er een open medium is waar het bericht over verstuurt kan worden waardoor de daadwerkelijke boodschap geheim dient te blijven. Dit is veel gebruikt in het engelse verzet tijdens de oorlog. (Er zijn destijds een hoop geiten gemolken)

(b) *Wat is het nadeel van steganografie?*
Het nadeel van steganografie is dat als het code boek uitlekt iedereen weet wat de berichten betekenen en je beperkt ben tot een code boek om te achterhalen wat een code betekent. je kan namelijk niet alle commado's kwijt in een code boek.

4.4 Opdracht 4

Een bekende manier om geheime boodschappen te ontcijferen is gebruik te maken van letterfrequenties. Deze methode werkt bij systemen waarbij dfe letters simpelweg vervangen worden door andere tekens, zoals monoafabetische substituering. Methoden die gebruik maken van verwisselingen van letterposities zijn minder kwetsbaar voor deze methoden.

(a) *Welke principes zijn volgens Shannon noodzakelijk voor een betrouwbaar cryptografisch systeem?*

- Het systeem moet praktische zijn
- Het systeem moet niet geheim zijn
- Het moet werken met telegrafie
- Het moet portable zijn
- Het moet makkelijk in gebruik zijn

(b) *Als de letters uniform verdeeld zijn in een bericht dan hebben alle letters $i = 1 \ldots 26$ evenveel kans om op te treden $p(X = a) = p_a = 1/26$. Indien wij een ander bericht van gelijke lengte met willekeurig verdeelde letters op het eerste bericht leggen, dan is de kans dat een positie twee letters 'a' op elkaar liggen gelijk aan $P(X = a \cap x = a) = p_a^2 = (1/26)^2 = 0,0385$. Als wij deze methode per taal uitvoeren, blijkt dat deze coïncidentie per letter per taal veschilt. Om deze ereigenschap van een taal met een kental te beschrijven wordt zij gedefineerd als de* coïncidentie-index*: $i_c = \sum_{i=1}^{26} p_i^2$:*

taal	i_c
Engels	00661
Frans	0,0778
Duits	0,0762
Italiaans	0,0738
Japans	0,0819
Russische	0,0529
Random	0,0385

Bereken de coïncidentie-index voor de Nederlandse taal. Maak gebruik van de gegeven letterfrequenties in bijlage C.

i	Symbool	p_i	$\sum_{j=1}^{i}(p_j^2)$
1	E	0,190	0,0361
2	N	0,110	0,0482
3	A	0,066	0,052556
4	T	0,065	0,056781
5	D	0,063	0,06075
6	O	0,063	0,064719
7	R	0,060	0,068319
8	I	0,054	0,071235
9	L	0,042	0,072999
10	S	0,041	0,07468
11	G	0,038	0,076124
12	H	0,026	0,0768
13	V	0,025	0,077425
14	U	0,024	0,078001
15	K	0,020	0,078401
16	M	0,020	0,078801
17	B	0,015	0,079026
18	W	0,015	0,079251
19	Y	0,015	0,079476
20	C	0,013	0,079645
21	Z	0,013	0,079814
22	F	0,010	0,079914
23	Z	0,010	0,080014
24	J	0,001	0,080015
25	Q	0,001	0,080016
26	X	0,000	0,080016
Totaal			0,080016

de i_c van de nederlandse taal is 0,080016

(c) *Hoe zou de coïncidentie-index i_c gebruikt kunnen worden bij het kraken*

van een cipher-text?
als de taal van de tekst bekend is dan kan bij onder andere een Ceasar-gecodeerd bericht sneller gezien worden welke het is door een paar regels over elkaar te leggen.

4.5 Opdracht 5

Waarom moet de entropie $H(K|C)$ zo groot mogelijk zijn?
De entropie dient zo hoog mogelijk te zijn zodat als C bekent is dat dan de K niet makkelijk gevonden.

4.6 Opdracht 6

Een natuurlijke tekst in het Nederlands heeft een relatieve nulde-orde redundantie van 50%. De Nederlandse tekst wordt op karakterbasis met een Ceasarcode versleuteld.

(a) *Bereken de kritieke lengte van de cipher-text.*
$\frac{H(K)}{R(P)} = \frac{ld(26)}{50\% \cdot ld(26)} = 2$

(b) *Indien de Nederlandse tekst gecomprimeerd wordt met een code-efficiëntie van 70%, wat is dan de kritieke lengte van de Ceasar codering?*
$\frac{H(K)}{R(P)} = \frac{ld(26)}{30\% \cdot ld(26)} = 3\frac{1}{3}$