Deel II

Probabilistische Modellen

Les 10 Entropie, informatie en afstanden van kansverdelingen

Het algemeen probleem in de patroonherkenning is, gegeven een aantal klassen $K_1, \ldots, K_n$ van mogelijke patronen, een nieuw patroon aan een van de klassen K_i toe te wijzen. Denk bij de klassen bijvoorbeeld aan letters in de handschriftherkenning, aan woorden of fonemen in de spraakherkenning of objecten in de beeldherkenning.

In het verleden is geprobeerd, regels te vinden waarmee de klasse van een nieuw patroon bepaald kan worden. Maar er is gebleken dat dit slechts zeer beperkt inzetbaar is en de beste methoden in de patroonherkenning gebruiken nu probabilistische modellen.

Er zijn verschillende mogelijkheden voor de rol die kansverdelingen hierbij kunnen spelen:

- Het nieuwe patroon wordt door een vector (of een rij vectoren) in de *feature space* weergegeven. De klassen zijn gerepresenteerd door kansverdelingen op de feature space die aangeven hoe groot de kans is dat een patroon met een zekere feature vector bij deze klasse hoort. Het patroon wordt dan aan de klasse toegewezen waarvoor deze kans maximaal is.
- Ook voor het patroon wordt een kansverdeling bepaald en er wordt de klasse gekozen, waarvoor deze kansverdeling het meeste op de kansverdeling van de klasse lijkt.

We zullen later zien, dat deze twee mogelijkheden niet eens zo heel verschillend zijn, maar we zullen nu eerst het tweede idee bekijken, omdat dit niet zo intuïtief is.

10.1 Afstanden tussen kansverdelingen

Als voorbeeld bekijken we het probleem van de automatische taalherkenning op geschreven tekst. Voor een mens is dat natuurlijk geen erg groot probleem, tenminste bij bekende talen of bij talen waar men iets over weet, maar de automatisering hiervan is al een stukje lastiger.

Onze aanpak is, de relatieve frequenties van de letters te gebruiken. Het is natuurlijk bekend dat de letters in het alfabet niet even vaak gebruikt worden, in het Nederlands is bijvoorbeeld de letter `E` de meest frequente. Het idee is dat de relatieve frequenties voor verschillende talen er verschillend uit zien en dat we hiermee de talen kunnen onderscheiden.

> Vanaf de 16de eeuw zijn de relatieve frequenties in de cryptanalyse gebruikt om versleutelingen met monoalfabetische substitutie (elke letter wordt door een andere letter vervangen, maar één letter steeds door dezelfde) te kraken. Tot op die tijd dacht men eigenlijk dat zo'n versleuteling niet te kraken was, omdat er veel te veel sleutels bestaan ($26! \approx 4.03 \cdot 10^{26}$) om alle te proberen. Maar als men al weet dat de meest frequente letter in de versleuteling een `E` is en de volgende waarschijnlijk een `N` kan men al gauw verdere letters gokken.

> Het idee dat de letters überhaupt verschillende frequenties hebben, is waarschijnlijk pas na de opkomst van de boekdrukkerij (door Gutenberg) ontdekt, omdat de loodletters verschillend snel versleten waren.

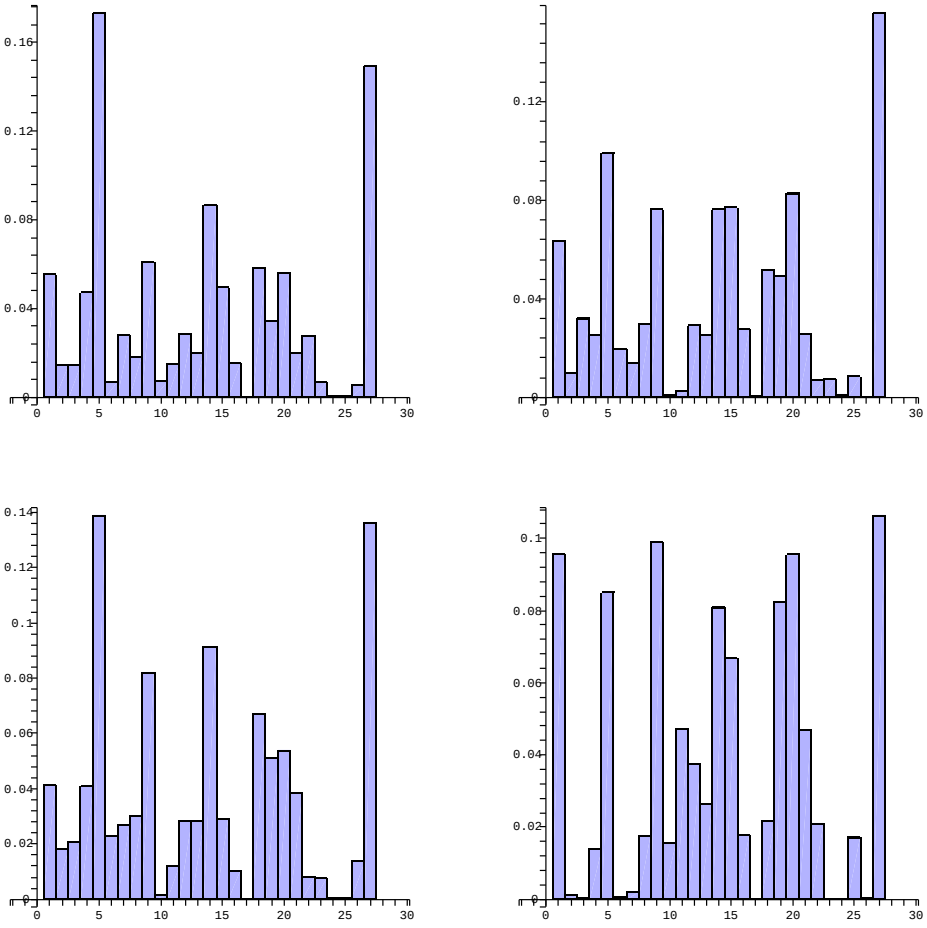
Voor een gegeven taal kan men nu op een grote achtergrondtekst de frequenties tellen en dit als kansverdeling van de stochast X die de letters beschrijft nemen. Men krijgt zo de kansen $p_1 := p(X = \texttt{A})$, $p_2 := p(X = \texttt{B})$, ..., $p_{26} := p(X = \texttt{Z})$, $p_{27} := p(X = \textit{spatie})$.

De volgende tabel geeft deze kansverdelingen voor de vier talen *Nederlands*, *Engels*, *Duits* en *Fins* weer. De gebruikte achtergrondtekst is een tekst van de Europese Unie die in de verschillende talen vertaald is en ongeveer 50000 letters bevat.

letter	Nederlands	Engels	Duits	Fins
A	5.55%	6.37%	4.14%	9.57%
B	1.45%	0.99%	1.82%	0.10%
C	1.45%	3.20%	2.09%	0.05%
D	4.72%	2.56%	4.09%	1.40%
E	17.31%	9.93%	13.89%	8.50%
F	0.68%	1.95%	2.28%	0.07%
G	2.79%	1.41%	2.67%	0.19%
H	1.83%	3.00%	3.00%	1.77%
I	6.09%	7.62%	8.22%	9.90%
J	0.70%	0.10%	0.14%	1.57%
K	1.51%	0.27%	1.21%	4.74%
L	2.87%	2.93%	2.83%	3.75%
M	1.98%	2.52%	2.81%	2.65%
N	8.67%	7.63%	9.14%	8.08%
O	4.94%	7.73%	2.92%	6.68%
P	1.53%	2.78%	1.03%	1.78%
Q	0.01%	0.04%	0.01%	0.01%
R	5.81%	5.15%	6.69%	2.16%
S	3.44%	4.92%	5.10%	8.24%
T	5.63%	8.30%	5.40%	9.54%
U	2.01%	2.57%	3.85%	4.70%
V	2.77%	0.70%	0.80%	2.10%
W	0.67%	0.75%	0.77%	0.02%
X	0.05%	0.12%	0.05%	0.01%
Y	0.04%	0.84%	0.06%	1.71%
Z	0.55%	0.02%	1.36%	0.05%
spatie	14.94%	15.61%	13.63%	10.64%

Uit deze tabel kan men concluderen dat de kansverdelingen voor Nederlands, Engels en Duits enigszins op elkaar lijken, terwijl de verdeling voor Fins er behoorlijk anders uit ziet. Bijvoorbeeld bepaalt de relatieve frequentie van de *spatie* de gemiddelde lengte van de woorden (namelijk door $len = \frac{1}{p} - 1$) en men ziet dat de woorden in het Fins duidelijk langer zijn dan in de andere talen.

Een goed idee van de frequentieverdelingen krijgt men door de verdelingen als histogrammen te plotten, zo als in Figuur II.1 te zien.



Figuur II.1: Letter-frequentieverdelingen voor Nederlands (links boven) en Engels (rechts boven), Duits (links onder) en Fins (rechts onder).

Als men nu een nieuwe tekst krijgt waarvan men de taal wil bepalen, berekent men de frequentieverdeling voor deze tekst en vergelijkt deze met de bekende kansverdelingen van de verschillende talen. De aanname is dan, dat de tekst bij die taal hoort waarvoor de kansverdelingen het meeste op elkaar lijken.

De vraag is nu hoe men objectief bepaald, dat een kansverdeling meer op een dan op een andere lijkt.

Om een eenvoudige notatie te krijgen, beschrijven we een discrete kansverdeling P op de verzameling $\Omega = \{1, \ldots, n\}$ door de vector van kansen $p_i := p(i)$, dus $P = (p_1, p_2, \ldots, p_n)$. Voor een tweede kansverdeling $Q = (q_1, q_2, \ldots, q_n)$ op dezelfde verzameling Ω willen we nu een afstand tussen P en Q definiëren.

Een voor de hand liggende idee is, de euclidische afstand van de vectoren P

en Q in de n-dimensionale ruimte te nemen, dit geeft

$$d_2(P,Q) = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{\frac{1}{2}}.$$

Maar net zo goed zouden we in plaats van de kwadraten van de verschillen tussen p_i en q_i ook de absolute waarden van de verschillen kunnen optellen:

$$d_1(P,Q) = \sum_{i=1}^{n} |p_i - q_i|.$$

We kunnen zelfs heel algemeen een macht van de verschillen tussen p_i en q_i optellen, dit geeft

$$d_r(P,Q) = \left(\sum_{i=1}^{n} |p_i - q_i|^r \right)^{\frac{1}{r}}.$$

Hierbij hoeft r niet eens een geheel getal te zijn, we kunnen een willekeurige r met $0 < r < \infty$ kiezen. De reden dat we bij een r-de macht ook weer een r-de wortel trekken, heeft ermee te maken dat men graag wil dat een vermenigvuldiging van de vectoren met een constante factor tot een vermenigvuldiging van de afstand met dezelfde factor leidt.

Voor de volledigheid noemen we nog een verdere afstand, die we formeel kunnen krijgen als we bij $d_r(P,Q)$ de $r \to \infty$ laten lopen. Dan krijgen we namelijk de afstand

$$d_\infty(P,Q) = \max_i |p_i - q_i|$$

die gewoon het grootste verschil in een van de componenten aangeeft. Maar als we naar vectoren van kansverdelingen kijken, is dit meestal geen bijzonder nuttige afstand.

De vraag welke afstand nu een slimme keuze is, heeft helaas geen eenvoudig antwoord. Het hangt namelijk van het probleem af. Hoe groter de waarde van de parameter r is hoe groter is relatief het gewicht van de grotere verschillen en hoe kleiner de invloed van kleine verschillen. Als r heel groot wordt, speelt inderdaad alleen maar het grootste verschil nog een rol. In sommige problemen is het misschien wenselijk, kleine verschillen te onderdrukken, maar soms ligt de informatie juist in de componenten met kleine verschillen.

In een iets algemenere opzet zou men voor elke component een functie $d_i(p_i, q_i)$ definiëren, die de afstand in deze component aangeeft. Als afstand krijgt men dan

$$d(P,Q) = \sum_{i=1}^{n} d_i(p_i, q_i).$$

Hierbij kan de functie d_i aan de ene kant ervoor zorgen, dat componenten met belangrijkere informatie een hoog gewicht krijgen, maar ook dat afhankelijk van de kansen een hoger of lager gewicht toegewezen wordt.

Een eenvoudig voorbeeld hiervan is het toewijzen van gewichten aan de enkele componenten, dus bijvoorbeeld

$$d(P,Q) = \sum_{i=1}^{n} w_i|p_i - q_i| \quad \text{of} \quad d(P,Q) = \sum_{i=1}^{n} w_i p_i q_i.$$

Het laatste is een inproduct van de twee vectoren P en Q en geeft weer dat we in principe ook de hoek tussen twee vectoren als een soort afstand kunnen interpreteren, zeker als de lengte van de vectoren genormeerd is.

Deze methode wordt bijvoorbeeld in (eenvoudige) zoekmachines gebruikt, de gewichten zijn dan bijvoorbeeld de negatieve logaritmen van de relatieve frequenties van de woorden. Zo houdt men rekening ermee, dat frequente woorden weinig informatie over een document geven, terwijl minder frequente woorden vaak een belangrijke hint zijn.

De afstanden die we tot nu toe hebben bekeken, hebben op zich weinig met kansverdelingen te maken, want we hebben eigenlijk alleen maar naar de vectoren gekeken. Het enige wat van de kansverdelingen over blijft, is dat $\sum_{i=1}^{n} p_i = 1$. We zullen nu naar een alternatieve aanpak kijken, die geïnspireerd is van de communicatie- en informatietheorie.

10.2 Onzekerheid

Als we een experiment of gebeurtenis door een kansverdeling beschrijven, drukken we hiermee uit dat we niet zeker over de uitkomst zijn. Maar we hebben ook een intuïtieve idee dat de onzekerheid soms groter is dan in andere gevallen. Bijvoorbeeld zijn we onzekerder over de uitkomst bij het werpen van een dobbelsteen dan bij het werpen van een munt, omdat er in het ene geval 6 mogelijke uitkomsten zijn, maar in het andere geval slechts 2. Ook bij een sportwedstrijd hangt onze onzekerheid ervan af hoe we de kansen voor de uitkomst inschatten: Als alleen maar de KI-studenten onderling een zwemwedstrijd uitvechten is de onzekerheid waarschijnlijk groter dan als Pieter van den Hoogenband ook meedoet.

Het is duidelijk dat de onzekerheid bepaald wordt door de kansen die we aan de mogelijke uitkomsten toewijzen. We kunnen ons dus afvragen hoe we voor een kansverdeling $P = (p_1, \ldots, p_n)$ een waarde voor de onzekerheid kunnen berekenen. Het idee dat we hiervoor hebben, is een functie

$$H(P) = H(p_1, \ldots, p_n)$$

te vinden, die de onzekerheid weergeeft. Omdat we intuïtief wel een idee van de onzekerheid bij een kansverdeling hebben, moet zo'n functie zekere eigenschappen hebben. In het jaar 1948 is hiervoor door C.E. Shannon (dezelfde Shannon als bij het sampling theorema) in het kader van de communicatietheorie een voorstel gedaan aan welke eisen zo'n functie $H(P)$ zou moeten voldoen. De link tussen communicatietheorie en kansrekening bestaat erin, dat communicatie als transmissie (van bit-strings, dus van ketens van 0en en 1en) via kanalen gemodelleerd wordt, waarbij er toevallig fouten kunnen optreden. De vraag is dan, hoe veel onzekerheid in het ontvangen signaal ligt.

De eisen van Shannon zijn als volgt:

(1) Voor kansverdelingen P op n punten is $H(P)$ maximaal als P de uniforme verdeling is met $p_i = \frac{1}{n}$ voor alle i. Dit zegt dat we bij n mogelijke uitkomsten de grootste onzekerheid hebben, als elke optie dezelfde kans heeft.

(2) De onzekerheid hangt alleen maar van de kansen p_i, maar niet van hun volgorde af, dus geldt $H(p_1, \ldots, p_n) = H(p_{\pi(1)}, \ldots, p_{\pi(n)})$ voor elke permutatie π van de indices.

(3) $H(P) \geq 0$ en $H(P) = 0$ alleen maar als één van de $p_i = 1$ is (en de anderen dus 0). Dit betekent dat we altijd onzeker zijn, behalve als een uitkomst kans 1 heeft en dus zeker gaat gebeuren.

(4) $H(p_1, \ldots, p_n) = H(p_1, \ldots, p_n, 0)$, dus de onzekerheid verandert niet, als we de kansverdeling uitbreiden tot meer mogelijke gebeurtenissen, maar de nieuwe opties kans 0 hebben en dus nooit kunnen gebeuren.

(5) $H(\frac{1}{n}, \ldots, \frac{1}{n}) \leq H(\frac{1}{n+1}, \ldots, \frac{1}{n+1})$, d.w.z. de onzekerheid bij een uniforme verdeling met $n+1$ mogelijke uitkomsten is groter dan bij n mogelijke uitkomsten.

(6) $H(P)$ is een continue functie in de argumenten $p_1, \ldots, p_n$, want als we de kansen maar heel weinig veranderen, verandert ook de onzekerheid nauwelijks.

(7) $H(\frac{1}{mn}, \ldots, \frac{1}{mn}) = H(\frac{1}{m}, \ldots, \frac{1}{m}) + H(\frac{1}{n}, \ldots, \frac{1}{n})$. Als we twee onafhankelijke experimenten met uniforme verdelingen tot een gezamenlijk experiment combineren, willen we dat de onzekerheid van het gecombineerde experiment juist de som van de onzekerheden bij de enkele experimenten is.

(8) We splitsen de verzameling $\Omega = \{1, \ldots, n\}$ op in de twee deelverzamelingen $\Omega_1 = \{1, \ldots, r\}$ en $\Omega_2 = \{r+1, \ldots, n\}$. De totale kans voor de uitkomsten in Ω_1 is $q_1 = p_1 + \ldots + p_r$ en de kans voor Ω_2 is $q_2 = p_{r+1} + \ldots + p_n$. De onzekerheid of een uitkomst in Ω_1 of Ω_2 ligt, is $H(q_1, q_2)$, de onzekerheid over een uitkomst in Ω_1 is $H(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1})$, omdat $(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1})$ juist de kansverdeling op Ω_1 is. Net zo is $H(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2})$ de onzekerheid over een uitkomst in Ω_2. De totale onzekerheid over de uitkomst van P is samengesteld uit de onzekerheden in welke deelverzameling een uitkomst ligt en de onzekerheden van de twee deelverzamelingen, die met hun kansen q_1 en q_2 gewogen zijn, dus moet gelden:

$$H(p_1, \ldots, p_n) = H(q_1, q_2) + q_1 H(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1}) + q_2 H(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2}).$$

De meeste van deze punten zijn redelijk vanzelfsprekend, alleen maar de punten (7) en (8) stellen inhoudelijke eisen, namelijk hoe de onzekerheden van verschillende gebeurtenissen gecombineerde moeten worden. Het interessante is

nu, dat deze eisen zo sterk zijn dat er in principe alleen maar een functie $H(P)$ bestaat die aan de eisen voldoet, namelijk de functie:

$$H(P) = H(p_1, \ldots, p_n) = -\lambda \sum_{i=1}^{n} p_i \log(p_i)$$

met $\lambda > 0$, waarbij de som alleen maar over de p_i met $p_i \neq 0$ loopt. We zullen dit hier niet bewijzen, maar wel nagaan dat de functie $H(P)$ aan de eisen (1)-(8) voldoet. Hierbij zijn de punten (2), (3), (4) en (6) duidelijk.

(1) In het punt $x = 1$ is $\log(x) = 0$ en $\log'(x) = 1$, dus is de lijn met vergelijking $y = x - 1$ de raaklijn aan de grafiek van de logaritme in het punt $x = 1$. Omdat $\log''(x) = -\frac{1}{x^2} < 0$, blijft de logaritme steeds onder deze raaklijn, daarom geldt $\log(x) \leq x - 1$. Voor twee kansverdelingen $P = (p_1, \ldots, p_n)$ en $Q = (q_1, \ldots, q_n)$ volgt hieruit dat

$$\sum_{i=1}^{n} p_i \log(\frac{q_i}{p_i}) \leq \sum_{i=1}^{n} p_i (\frac{q_i}{p_i} - 1) = \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0.$$

Wegens $\log(\frac{q_i}{p_i}) = \log(q_i) - \log(p_i)$ volgt hieruit dat

$$-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i).$$

Als we nu voor Q speciaal de uniforme verdeling met $q_i = \frac{1}{n}$ kiezen, volgt hieruit aan de ene kant dat

$$H(P) \leq -\sum_{i=1}^{n} p_i \log(\frac{1}{n}) = \sum_{i=1}^{n} p_i \log(n) = \log(n).$$

Maar aan de andere kant is $H(Q) = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = \log(n)$, dus is de waarde voor de uniforme verdeling inderdaad maximaal.

(5) Dit volgt nu meteen uit deel (1), omdat voor een uniforme verdeling geldt dat $H(\frac{1}{n}, \ldots, \frac{1}{n}) = \log(n)$ en $\log(n) < \log(n+1)$.

(7) Dit volgt ook uit deel (1), omdat $\log(mn) = \log(m) + \log(n)$.

(8) Uit $q_1 = \sum_{i=1}^{r} p_i$ en $q_2 = \sum_{i=r+1}^{n} p_i$ volgt

$$\begin{aligned}
&H(q_1, q_2) + q_1 H(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1}) + q_2 H(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2}) \\
&= -q_1 \log(q_1) - q_2 \log(q_2) - q_1 \sum_{i=1}^{r} \frac{p_i}{q_1} \log(\frac{p_i}{q_1}) - q_2 \sum_{i=r+1}^{n} \frac{p_i}{q_2} \log(\frac{p_i}{q_2}) \\
&= -\sum_{i=1}^{r} p_i \log(q_1) - \sum_{i=r+1}^{n} p_i \log(q_2) - \sum_{i=1}^{r} p_i (\log(p_i) - \log(q_1)) \\
&\quad - \sum_{i=r+1}^{n} p_i (\log(p_i) - \log(q_2)) = -\sum_{i=1}^{n} p_i \log(p_i) = H(p_1, \ldots, p_n).
\end{aligned}$$

We hebben bij punt (1) twee belangrijke resultaten bewezen, die we nog eens expliciet willen noemen:

(I) Voor een uniforme verdeling P op n punten is $H(P) = \log(n)$.

(II) Voor twee kansverdelingen P en Q is $-\sum p_i \log(q_i)$ minimaal voor $Q = P$.

Omdat de ideeën voor het formaliseren van onzekerheid uit de communicatietheorie komen waar men het over bit-strings heeft, is het gebruikelijk de functie $H(P)$ niet met behulp van de natuurlijke logaritme (met basis e) maar met de logaritme met basis 2 te formuleren. Omdat ${}^2\log(x) = \frac{\log(x)}{\log(2)}$ geeft dit alleen maar een verschil van de constante factor $\log(2)$. De functie

$$H(P) = H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ {}^2\log(p_i)$$

heet de *entropie* van de kansverdeling P.

> Het begrip *entropie* speelt ook in de natuurkunde, vooral in de thermodynamica, een belangrijke rol. Hier geeft de entropie een maat voor de wanorde in een systeem. De tweede hoofdstelling van de thermodynamica zegt (in het grof) dat in een gesloten systeem de entropie nooit afneemt, d.w.z. dat zonder invloed van buiten de wanorde in een systeem steeds toeneemt. (Dit is natuurlijk ook een alledaagse ervaring.)

We hebben tot nu toe de entropie alleen maar voor een kansverdeling gedefinieerd. Vaak spreekt men immers ook van de entropie van een stochast X. Hiermee is de entropie van de kansverdeling van de mogelijke uitkomsten van X bedoelt. Stel een stochast X heeft de mogelijke uitkomsten $x_1, \ldots, x_n$, dan geeft $p_i := p(X = x_i)$ de kans op de i-de mogelijke uitkomst en de kansverdeling $P = (p_1, \ldots, p_n)$ beschrijft de kansen van de mogelijke uitkomsten van X. We definiëren dus de de entropie van een stochast X met mogelijke uitkomsten $x_1, \ldots, x_n$ door

$$H(X) := -\sum_{i=1}^{n} p(X = x_i) \ {}^2\log(p(X = x_i)).$$

> We zullen het in deze les alleen maar over discrete kansverdelingen hebben. De overgang tot continue kansverdeling is echter geen probleem: In plaats van de kansen p_i krijgen we een dichtheidsfunctie $f(x)$ voor de kansverdeling en de som over de mogelijke uitkomsten wordt de integraal over de continue variabel x. Voor de entropie van een stochast X met dichtheidsfunctie $f(x)$ krijgt men zo:
>
> $$H(X) := -\int_{-\infty}^{\infty} f(x) \ {}^2\log(f(x)) \ dx.$$

We komen even terug op het voorbeeld van de frequentieverdelingen voor de letters in verschillende talen. Als we voor deze kansverdelingen de entropieën uitrekenen, krijgen we:

$$H(\text{Nederlands}) = 4.019, \qquad H(\text{Engels}) = 4.070,$$
$$H(\text{Duits}) = 4.109, \qquad H(\text{Fins}) = 3.982.$$

We zullen later nog zien hoe we deze waarden moeten interpreteren. Het enige wat we nu al kunnen zeggen is dat de onzekerheid in het Duits het grootst en in het Fins het kleinste lijkt. Om in te kunnen schatten, hoe significant de verschillen tussen de talen zijn, vergelijken we de waarden met de entropie van de uniforme verdeling op 27 letters, deze is ${}^2\log(27) \approx 4.755$. Dit betekent dat de entropieën tot op ongeveer 0.7 na bij de maximaal mogelijke waarde liggen, en het verschil van meer dan 0.1 tussen Duits en Fins lijkt dus redelijk groot.

10.3 Voorwaardelijke entropie

Een belangrijke vraag is hoe zich de entropie van verschillende stochasten gedraagt als we deze combineren. We zouden verwachten, dat voor twee onafhankelijke stochasten X en Y de entropie van de combinatie van X en Y de som van de entropieën van X en Y is. Voor stochasten X, Y met uniforme verdelingen is dit juist eis (7) in onze lijst. Voor twee stochasten X en Y geldt inderdaad de stelling:

$H(X,Y) \leq H(X) + H(Y)$ en
$H(X,Y) = H(X) + H(Y)$ alleen maar als X en Y onafhankelijk zijn.

Dit zien we als volgt in: We definiëren de kansen voor de stochasten als $p_i := p(X = x_i)$ voor $1 \leq i \leq n$, $y_j := p(Y = y_j)$ voor $1 \leq j \leq m$ en de gecombineerde kans als $r_{ij} := p(X = x_i, Y = y_j)$. Als we voor vaste i de kansen r_{ij} voor alle j optellen, krijgen we de kans op x_i, dus geldt $p_i = \sum_{j=1}^{m} r_{ij}$ en evenzo $q_j = \sum_{i=1}^{n} r_{ij}$. We hebben dus

$$\begin{aligned}
H(X) + H(Y) &= -\sum_{i=1}^{n} p_i \ {}^2\log(p_i) - \sum_{j=1}^{m} q_j \ {}^2\log(q_j) \\
&= -\sum_{i=1}^{n} (\sum_{j=1}^{m} r_{ij}) \ {}^2\log(p_i) - \sum_{j=1}^{m} (\sum_{i=1}^{n} r_{ij}) \ {}^2\log(q_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} ({}^2\log(p_i) + {}^2\log(q_j)) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \ {}^2\log(p_i q_j) \\
&\geq -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \ {}^2\log(r_{ij}) = H(X,Y).
\end{aligned}$$

De ongelijkheid $-\sum\sum r_{ij} \ {}^2\log(p_i q_j) \geq -\sum\sum r_{ij} \ {}^2\log(r_{ij})$ volgt hierbij weer uit de eigenschap (II) die we boven hebben bewezen, omdat ook $p_i q_j$ een kansverdeling op $\{1, \ldots, n\} \times \{1, \ldots, m\}$ is.

We zien dat $H(X) + H(Y)$ alleen maar geldt als $p_i q_j = r_{ij}$ voor alle paren (i, j), dus als $p(X = x_i) \cdot p(Y = y_j) = p(X = x_i, Y = y_j)$, maar dit is precies de uitspraak dat X en Y onafhankelijk zijn.

Als we stochasten combineren, moeten we het natuurlijk ook over voorwaardelijke kansen hebben. Maar voorwaardelijke kansen zijn ook gewoon kansverdelingen: Als we de kans op een uitkomst x_i voor de stochast X onder de voorwaarde A weer als $p_i := p(X = x_i \mid A)$ beschrijven, is $P = (p_1, \ldots, p_n)$ een kansverdeling en $\sum_{i=1}^{n} p_i = 1$. We definiëren daarom de *voorwaardelijke entropie* $H(X \mid A)$ door

$$H(X \mid A) := -\sum_{i=1}^{n} p(X = x_i \mid A)\ {}^2\log(p(X = x_i \mid A)).$$

Nog algemener kunnen we ook de voorwaardelijke entropie van een stochast X, gegeven een andere stochast Y definiëren. Het idee hierbij is, dat de uitkomsten van de stochast Y de onzekerheid over de stochast X kunnen veranderen. We lopen dus over alle mogelijke uitkomsten y_j van de stochast Y, berekenen voor deze uitkomsten de voorwaardelijke entropie $H(X \mid y_j)$ en tellen deze entropieën op, met de kansen op de enkele y_j als gewichten, dus:

$$\begin{aligned} H(X \mid Y) &:= \sum_{j=1}^{m} H(X \mid Y = y_j) p(Y = x_j) \\ &= -\sum_{j=1}^{m} \sum_{i=1}^{n} p(X = x_i \mid Y = y_j)\ {}^2\log(p(X = x_i \mid Y = y_j)) \cdot p(Y = y_j). \end{aligned}$$

Dat deze definitie enigszins zinvol is, zien we aan de twee extreme gevallen $Y = X$ en X en Y onafhankelijk:

Als $Y = X$ is, dan is $p(X = x_i \mid X = x_j) = 1$ als $i = j$ en 0 als $i \neq j$. Maar dan geldt $H(X \mid X) = -\sum_{i=1}^{n} p(X = x_i \mid X = x_i)\ {}^2\log(p(X = x_i \mid X = x_i))p(X = x_i) = -\sum_{i=1}^{n} 1 \cdot 0 \cdot p(X = x_i) = 0$. Er geldt dus

$$H(X \mid X) = 0.$$

Dit zegt dat er geen onzekerheid over X meer bestaat, als we de uitkomsten van X al kennen.

Omgekeerd, als X en Y onafhankelijk zijn, dan geldt $p(X = x_i \mid Y = y_j) = p(X = x_i)$, en hieruit volgt $H(X \mid Y) = -\sum_{j=1}^{m} \sum_{i=1}^{n} p(X = x_i)\ {}^2\log(p(X = x_i))p(Y = y_j) = -\sum_{i=1}^{n} p(X = x_i)\ {}^2\log(p(X = x_i)) = H(X)$. Voor onafhankelijke stochasten X en Y geldt dus dat

$$H(X \mid Y) = H(X).$$

Dit betekent, dat de kennis over Y de onzekerheid bij X niet reduceert, en dat is precies wat we bij onafhankelijke stochasten zouden verwachten.

We kunnen nu ook de precieze samenhang tussen de voorwaardelijke entropie $H(X \mid Y)$ en de entropie van de combinatie van X en Y aangeven, er geldt namelijk

$$H(X, Y) = H(Y) + H(X \mid Y) \quad \text{of te wel} \quad H(X \mid Y) = H(X, Y) - H(Y).$$

Dit zien we als volgt in: We schrijven weer $r_{ij} := p(X = x_i, Y = y_j)$ voor de gecombineerde kans op x_i en y_j. Volgens de definitie van de voorwaardelijke kans geldt dat $p(X = x_i \mid Y = y_j) = \frac{r_{ij}}{q_j}$ en dus $r_{ij} = p(X = x_i \mid Y = y_j)q_j$, waarbij we weer $q_j := p(Y = y_j)$ schrijven. Er geldt dus:

$$\begin{aligned}
H(X,Y) &= -\sum_{i,j} r_{ij}\ {}^2\log(r_{ij}) = -\sum_{i,j} r_{ij}\ {}^2\log(p(X = x_i \mid Y = y_j)q_j) \\
&= -\sum_{i,j} r_{ij}\ {}^2\log(p(X = x_i \mid Y = y_j)) - \sum_{i,j} r_{ij}\ {}^2\log(q_j) \\
&= -\sum_{i,j} r_{ij}\ {}^2\log(p(X = x_i \mid Y = y_j)) - \sum_{j=1}^{m} q_j\ {}^2\log(q_j) \\
&= -\sum_{i,j} p(X = x_i \mid Y = y_j)q_j\ {}^2\log(p(X = x_i \mid Y = y_j)) - H(Y) \\
&= H(X \mid Y) + H(Y).
\end{aligned}$$

Hieruit volgt in het bijzonder dat

$$H(X \mid Y) \leq H(X),$$

want $H(X \mid Y) = H(X,Y) - H(Y) \leq H(X) + H(Y) - H(Y) = H(X)$, en dus is de voorwaardelijke entropie van een stochast nooit groter dan zijn absolute entropie. Ook dit is een eigenschap die we van een redelijke maat voor onzekerheid hadden kunnen verwachten, want door aanvullende informatie zouden we niet onzekerder over de uitkomsten van X worden.

10.4 Informatie

We hebben bij de voorwaardelijke entropie gezien, dat kennis over een stochast Y de onzekerheid over de stochast X kan reduceren. Het verschil van de entropieën $H(X) - H(X \mid Y)$ kunnen we dus zien als de informatie die Y aan onze kennis over X bijdraagt. Dit lijdt tot een precieze definitie van het begrip *informatie*, die we nu gaan bespreken.

Net als bij de entropie geven we ook bij de informatie aan, wat we van een functie verwachten, die de informatie van een gebeurtenis beschrijft. We schrijven $I(X = x_i)$ voor de informatie die de uitkomst x_i van de stochast X oplevert. Maar eigenlijk mag een abstracte definitie van informatie niet van de specifieke uitkomst afhangen, maar alleen maar van de kans op deze uitkomst. We willen dus dat $I(X = x_i) = I(p_i)$ voor $p_i = p(X = x_i)$. Een verdere eigenschap eisen we voor de informatie van onafhankelijke gebeurtenissen: Als X en Y onafhankelijke stochasten zijn, geldt met $p_i = p(X = x_i)$ en $q_j = p(Y = y_j)$ dat $p(X = x_i, Y = y_j) = p_i q_j$. Maar het ligt voor de hand dat de informatie die in de uitkomst $X = x_i$ en $Y = y_j$ zit, de som van de informaties van de enkele uitkomsten is. Dit geeft de eis $I(p_i q_j) = I(p_i) + I(q_j)$. Met een soortgelijke (maar eenvoudigere) redenering als bij de entropie kan men nu aantonen dat de functie I noodzakelijk van de vorm $I(p) = -\lambda \log(p)$ is, en ook hier kiest men voor de logaritme met basis 2, dus definieert men:

De informatie van een uitkomst $X = x$ met $p(X = x) = p$ is

$$I(p) := -\,{}^2\log(p).$$

Deze definitie van informatie klopt ook met onze intuïtie dat een gebeurtenis met een kleine kans meer informatie oplevert dan een gebeurtenis met een grote kans, namelijk het gewone.

Een belangrijke rechtvaardiging van deze definitie van informatie vinden we weer in de communicatietheorie: Als we een bit-string van lengte n produceren door toevallig n keer een 0 of 1 te kiezen, heeft elke bit van de string de informatie $I(\frac{1}{2}) = -\,{}^2\log(\frac{1}{2}) = {}^2\log(2) = 1$ en de totale informatie in de string is dus $-n\ {}^2\log(\frac{1}{n}) = n$, omdat de keuzes van de bits onafhankelijk zijn. Het is daarom ook gebruikelijk, informatie (en entropie) in *bits* aan te geven.

Met behulp van het begrip van informatie kunnen we nu de entropie herinterpreteren. Er geldt

$$H(X) = -\sum p_i\ {}^2\log(p_i) = \sum p_i \cdot I(p_i)$$

dus is de entropie het gemiddelde van de informatie in de enkele uitkomsten, gewogen met de kansen van de uitkomsten. In de taal van de kansrekening is zo'n gemiddelde juist de verwachtingswaarde, de entropie van een stochast is dus de verwachtingswaarde van de informatie van de enkele uitkomsten.

Maar dit kunnen we ook nog iets anders formuleren: Een uitkomst met informatie $I = {}^2\log(n)$ heeft kans $p = \frac{1}{n}$. Als de uitkomst bij een uniforme verdeling hoort, is $\frac{1}{p} = n$ het aantal mogelijke uitkomsten. Dit betekent dat we voor een uniforme verdeling het aantal mogelijke uitkomsten kunnen schrijven als $n = 2^I$. Maar we hebben nu gezien dat de entropie de verwachtingswaarde van de informatie in de enkele uitkomsten is, dus kunnen we $2^{H(X)}$ interpreteren als het gemiddelde aantal alternatieven, dat we bij de stochast X kunnen verwachten, met andere woorden de onzekerheid bij onze stochast X is even groot als de onzekerheid bij een uniforme verdeling met $2^{H(X)}$ mogelijke uitkomsten.

Met deze interpretatie van de entropie kijken we nu nog eens naar het voorbeeld van de frequentieverdelingen. We hebben:

$$\begin{array}{ll} 2^{H(\text{Nederlands})} = 16.21, & 2^{H(\text{Engels})} = 16.80, \\ 2^{H(\text{Duits})} = 17.26, & 2^{H(\text{Fins})} = 15.80. \end{array}$$

Het gemiddelde aantal alternatieven, dat we in de verschillende talen voor een letter verwachten, ligt dus tussen 15.80 voor Fins en 17.26 voor Duits, terwijl we bij een uniforme verdeling 27 alternatieven zouden hebben.

We hebben in het begin van deze sectie gezegd, dat het verschil van de entropieën $H(X) - H(X \mid Y)$ de informatie is, die Y over X onthuld. Als notatie hiervoor gebruiken we

$$I(X \mid Y) := H(X) - H(X \mid Y).$$

Er geldt $I(X \mid X) = H(X)$, want $H(X \mid X) = 0$ en voor onafhankelijke stochasten X en Y is $I(X \mid Y) = 0$, omdat $H(X \mid Y) = H(X) + H(Y)$.

Bij deze definitie kijken we naar de gemiddelde reductie die de enkele uitkomsten van Y voor de entropie van X opleveren. We kunnen natuurlijk ook naar de informatie kijken, die een bepaalde uitkomst $Y = y$ voor de stochast Y over X oplevert, deze is

$$I(X \mid Y = y) = H(X) - H(X \mid Y = y).$$

Er bestaat een iets verrassende symmetrie voor het onthullen van informatie van een stochast over de andere. We hebben namelijk

$$\begin{aligned} I(X \mid Y) &= H(X) - H(X \mid Y) = H(X) - (H(X,Y) - H(Y)) \\ &= H(Y) + (H(X) - H(X,Y)) = H(Y) - H(Y \mid X) = I(Y \mid X), \end{aligned}$$

dus de stochast X onthuld over Y net zo veel informatie als de stochast Y over X.

10.5 Kullback-Leibler afstand

We komen nu nog eens terug op de afstanden tussen kansverdelingen. We hebben gezien dat $-\sum p_i \ {}^2\log(p_i) \leq -\sum p_i \ {}^2\log(q_i)$, dus

$$\sum p_i({}^2\log(p_i) - {}^2\log(q_i)) = \sum p_i \ {}^2\log(\frac{p_i}{q_i}) \geq 0$$

met gelijkheid alleen maar als $p_i = q_i$ voor alle i. Men kan dus $\sum p_i \ {}^2\log(\frac{p_i}{q_i})$ als een soort afstand tussen de kansverdelingen $Q = (q_1, \ldots, q_n)$ en $P = (p_1, \ldots, p_n)$ opvatten en men noemt

$$d_{KL}(P, Q) := \sum p_i \ {}^2\log(\frac{p_i}{q_i})$$

de *Kullback-Leibler* afstand die Q van P heeft. Merk op dat $d_{KL}(P, Q)$ niet symmetrisch in de argumenten P en Q is, d.w.z. in het algemeen is $d_{KL}(P, Q) \neq d_{KL}(Q, P)$. De Kullback-Leibler afstand wordt vaak gebruikt om de afstanden van verschillende kansverdelingen Q van een vaste (doel-)kansverdeling P te bepalen.

Maar het is makkelijk om met behulp van $d_{KL}(P, Q)$ een afstand maken, die wel symmetrisch in de argumenten is, namelijk

$$d(P, Q) := \frac{1}{2}(d_{KL}(P, Q) + d_{KL}(Q, P)) = \frac{1}{2}\sum p_i \ {}^2\log(\frac{p_i}{q_i}) + q_i \ {}^2\log(\frac{q_i}{p_i}).$$

Ook dit heet meestal de Kullback-Leibler afstand, soms met het attribuut *symmetrisch* erbij. Let op dat de symmetrische Kullback-Leibler afstand geen afstandsfunctie in de gebruikelijke zin is, zo als de euclidische afstand bijvoorbeeld. Een echte afstandsfunctie moet namelijk de volgende drie eigenschappen hebben:

(i) $d(P, Q) \geq 0$ en $d(P, Q) = 0$ alleen maar als $P = Q$,

(ii) $d(P,Q) = d(Q,P)$ (symmetrie),

(iii) $d(P,Q) + d(Q,R) \geq d(P,R)$ (driehoeksongelijkheid).

De symmetrische Kullback-Leibler afstand heeft wel de eerste twee eigenschappen, maar voldoet niet aan de driehoeksongelijkheid.

De Kullback-Leibler afstand geeft het verschil tussen $-\sum p_i \, {}^2\log(q_i)$ en de entropie $H(P)$ van de kansverdeling P aan, er geldt dus $-\sum p_i \, {}^2\log(q_i) = H(P) + d_{KL}(P,Q)$. Als we nu $2^{H(P)}$ als gemiddelde alternatieven interpreteren, die we bij een stochast X met kansverdeling P verwachten, kunnen we ook de Kullback-Leibler afstand op deze manier interpreteren. Er geldt $2^{H(P)+d_{KL}(P,Q)} = 2^{H(P)} \cdot 2^{d_{KL}(P,Q)}$, dus is $2^{d_{KL}(P,Q)}$ de factor waarmee we het gemiddelde aantal alternatieven moeten vermenigvuldigen, omdat we de *verkeerde* kansverdeling Q in plaats van P toepassen.

De volgende tabellen geven links de Kullback-Leibler afstanden tussen de talen uit het voorbeeld met de frequentieverdelingen en rechts de factoren $2^{d_{KL}(P,Q)}$. Hierbij kan men een factor 1.138 interpreteren als een afwijking van 13.8% van het aantal verwachtte alternatieven bij de juiste kansverdeling.

taal	NL	EN	DU	FI
NL	-	0.186	0.091	0.471
EN	0.171	-	0.155	0.458
DU	0.090	0.177	-	0.610
FI	0.397	0.373	0.453	-

taal	NL	EN	DU	FI
NL	-	1.138	1.065	1.386
EN	1.126	-	1.114	1.373
DU	1.064	1.130	-	1.527
FI	1.317	1.295	1.368	-

Het is opvallend hoe sterk Duits en Fins van elkaar afwijken, terwijl Nederlands en Duits redelijk dicht bij elkaar liggen.

> De Kullback-Leibler afstand speelt een belangrijke rol bij het bepalen van de parameters van probabilistische modellen. Het idee is dat op een training set de kansen p_i bepaald worden en vervolgens een probabilistisch model gebouwd wordt, dat van enkele parameters afhangt. Dit kan bijvoorbeeld een normaalverdeling zijn, met als parameters de verwachtingswaarde en de variantie. Deze parameters kunnen meestal niet rechtstreeks berekent worden, maar worden in een iteratief proces benadert, waarbij de Kullback-Leibler afstand stapsgewijs kleiner wordt. Als geen verbetering meer bereikt wordt, worden deze parameters voor het model gekozen.

Belangrijke begrippen in deze les

- afstanden tussen kansverdelingen
- onzekerheid, entropie
- voorwaardelijke entropie

- informatie
- Kullback-Leibler afstand

Opgaven

44. Er vinden twee paardenraces plaats, het eerste met 7 paarden en het tweede met 8 paarden. In de eerste race hebben 3 paarden kans $\frac{1}{6}$ om te winnen, de andere 4 hebben kans $\frac{1}{8}$. In de tweede race hebben 2 paarden kans $\frac{1}{4}$ om te winnen en de andere 6 kans $\frac{1}{12}$. Maak eerst een gok in welk van de races de uitkomst onzekerder is (en geef een reden hiervoor), en bereken dan de entropieën voor de twee races.

45. Er wordt met een eerlijke dobbelsteen gedobbeld. De stochast X geeft het aantal ogen dat gedobbeld wordt, de stochast Y heeft de waarde 0 of 1, afhankelijk of het aantal ogen even of oneven is. Bereken $H(X)$, $H(Y)$ en $H(X \mid Y)$.

46. Voor een geheel getal N neemt de stochast X volgens een uniforme verdeling de waarden $1, 2, \ldots, 2N$ aan. De stochast Y is 0 als de waarde van X even is en Y is 1 als de waarde van X oneven is. Laat zien dat $H(X \mid Y) = H(X) - 1$ en dat $H(Y \mid X) = 0$.

47. De uitkomsten van twee (eerlijke) dobbelstenen worden door de stochasten X en Y beschreven, de som van de twee dobbelstenen door de stochast Z. Ga na dat voor de combinatie van de stochasten X en Y geldt dat $H(X,Y) = H(X) + H(Y)$ en dat $H(Z) < H(X,Y)$.

48. Een stochast X heeft een binomiale verdeling met parameters n en p, d.w.z. de kans op de i-de uitkomst is $p(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$. Laat zien dat $H(X) = -n(p\ {}^2\log(p) + (1-p)\ {}^2\log(1-p))$.

49. Waar zit meer informatie in, in een string van 10 letters uit $\{A, \ldots, Z\}$ of in een string van 26 cijfers uit $\{0, \ldots, 9\}$?

50. Er wordt met een eerlijke dobbelsteen gedobbeld. Wat is de informatie, die de kennis dat het aantal ogen niet door 3 deelbaar is, over het aantal ogen onthuld?

51. Uit onderzoek is gebleken dat 70% van de mannen donker haar hebben en 25% van de vrouwen blond zijn. Verder is bekend dat 80% van de blonde vrouwen met een donkerharig man trouwen. Hoeveel informatie over de haarkleur van de man onthuld de haarkleur van zijn vrouw?