Deel III

Probabilistische Modellen

Les 11 Onzekerheid, entropie en informatie

Als we erover nadenken hoe we conclusies trekken, komen we er snel achter dat dit meestal met het verkrijgen en verwerken van informatie te maken heeft. Vaak stellen we hiervoor vragen of maken een meting, om de onzekerheid die we over benodigde gegevens hebben te overkomen of tenminste te verkleinen.

Als we nu op het gebied van de kunstmatige intelligentie een systeem willen bouwen, dat op grond van zekere informatie beslissingen neemt, moeten we voor de begrippen *informatie* of *onzekerheid* definities vinden, die het mogelijk maken om ook kwantitatieve uitspraken hierover te kunnen doen. Een cruciaal begrip in dit kader is de *entropie* van een kansverdeling, die in principe aangeeft hoeveel bits we minstens nodig hebben, om de uitkomsten van een kansexperiment te beschrijven.

11.1 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.

Voorbeeld: Stel bij een paardenrace doen 8 paarden mee die niet even sterk zijn, maar waarvoor de kansen om te winnen gegeven zijn door

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{4}, \quad p_3 = \frac{1}{8}, \quad p_4 = \frac{1}{16}, \quad p_5 = p_6 = p_7 = p_8 = \frac{1}{64}.$$

Als we gewoon het nummer van het winnende paard door willen geven, hebben we hiervoor 3 bits nodig (want $2^3 = 8$). Maar omdat de kansen niet uniform verdeeld zijn, kunnen dit aantal reduceren, door de paarden met een hogere kans op een kortere manier te coderen. Hierbij moeten we er wel op letten, dat de beginstukken van de langere coderingen zelfs geen coderingen zijn. Een mogelijke codering voor de nummers 1 t/m 8 van de paarden in het voorbeeld is (aangegeven met strings van bits):

1: `0`, 2: `10`, 3: `110`, 4: `1110`, 5: `111100`, 6: `111101`, 7: `111110`, 8: `111111`.

Als we voor deze codering het gemiddeld benodigde aantal bits berekenen (dus de verwachtingswaarde van het aantal bits), krijgen we $\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{16} \cdot 4 + 4 \cdot \frac{1}{64} \cdot 6 = 2$. We hebben het aantal benodigde bits dus van 3 op 2 kunnen reduceren, door de uitkomsten die we vaker verwachten korter te coderen.

De onzekerheid bij een kansexperiment is natuurlijk bepaald door de kansen die we aan de mogelijke uitkomsten toewijzen. We kunnen ons dus afvragen

hoe we voor een discrete 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.

Eisen aan een functie voor de onzekerheid van een kansverdeling

De eisen die Shannon heeft gesteld zijn als volgt:

(1) $H(P)$ 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.

(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) $H(\frac{1}{mn}, \ldots, \frac{1}{mn}) = H(\frac{1}{m}, \ldots, \frac{1}{m}) + H(\frac{1}{n}, \ldots, \frac{1}{n})$. 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.

(7) 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 volstrekt intuïtief, alleen de punten (6) en (7) stellen inhoudelijke eisen, namelijk hoe de onzekerheden van verschillende gebeurtenissen gecombineerde moeten worden.

Het interessante (en misschien verrassende) is nu, dat deze eisen zo sterk zijn dat er in principe alleen maar een enkele 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 toelichten dat de functie $H(P)$ aan de eisen (1)-(7) voldoet. Hierbij zijn de punten (1)-(4) rechtstreeks duidelijk, de andere punten gaan we even na. Omdat de constante λ geen enkel verschil in de argumenten maakt, werken we voor het gemak met $\lambda = 1$.

(5) Voor een uniforme verdeling U_n op n punten geldt

$$H(U_n) = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = \sum_{i=1}^{n} \frac{1}{n} \log(n) = \log(n)$$

en omdat $\log(x)$ een strikt stijgende functie is, is $H(U_n) = \log(n) < \log(n+1) = H(U_{n+1})$.

(6) Dit volgt ook uit het feit dat $H(U_n) = \log(n)$, omdat $\log(mn) = \log(m) + \log(n)$.

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

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

Als we ons afvragen, bij welke kansverdeling met n mogelijke uitkomsten we de grootste onzekerheid hebben, ligt het voor de hand dat dit bij een uniforme verdeling het geval is, want in dit geval hebben we geen reden om een voorkeur aan een of andere uitkomst te geven. Als $H(P)$ een maat voor de onzekerheid is, zouden we dus verwachten dat de waarde van de functie $H(P)$ voor een uniforme verdeling maximaal is en dit laat zich inderdaad bewijzen.

Stelling: Van alle kansverdelingen P op n mogelijke uitkomsten geeft de uniforme verdeling met $p_i = \frac{1}{n}$ de maximale waarde van de functie $H(P)$.

Omdat het bewijs van deze stelling niet moeilijk is en belangrijke inzichten geeft, gaan we het even na:

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

$$\log(x) \leq x - 1 \text{ met gelijkheid alleen maar voor } x = 1.$$

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

We hebben inmiddels twee belangrijke inzichten gewonnen, die we nog eens expliciet willen aangeven:

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

(II) Voor twee kansverdelingen P en Q is $-\sum p_i \log(q_i) \geq H(P)$ en er geldt

$$D(P, Q) := \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i}) \geq 0$$

want we hadden gezien dat $\sum_{i=1}^{n} p_i \log(p_i) \geq \sum_{i=1}^{n} p_i \log(q_i)$ en hieruit volgt $\sum_{i=1}^{n} p_i(\log(p_i) - \log(q_i)) = \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i}) \geq 0$.

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. Wegens ${}^2\log(x) = \frac{\log(x)}{\log(2)}$ geeft dit alleen maar een verschil van de constante factor $\log(2)$.

Definitie: 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

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

Voorbeeld: Zij X de stochast van een Bernoulli experiment met kans p op succes (uitkomst 1) en kans $1 - p$ op mislukken (uitkomst 0). Er geldt

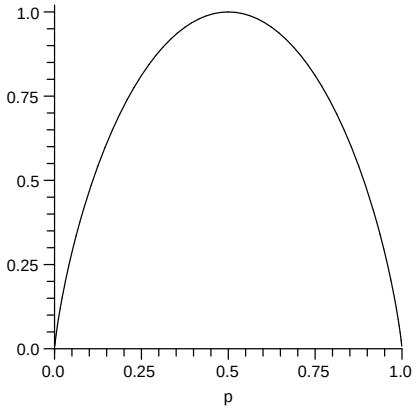
$$\begin{aligned} H(X) &= -p(X = 1) \ {}^2\log(p(X = 1)) - p(X = 0) \ {}^2\log(p(X = 0)) \\ &= -p \ {}^2\log(p) - (1 - p) \ {}^2\log(1 - p). \end{aligned}$$

In Figuur III.1 is duidelijk te zien dat de entropie maximaal wordt voor $p = 0.5$, dus voor een uniforme verdeling en dat in dit geval de entropie juist 1 *bit* is.

Relatieve entropie en de Kullback-Leibler afstand

We hebben gezien dat voor twee kansverdelingen $P = (p_1, \ldots, p_n)$ en $Q = (q_1, \ldots, q_n)$ geldt dat

$$D(P, Q) := \sum p_i({}^2\log(p_i) - {}^2\log(q_i)) = \sum p_i \ {}^2\log(\frac{p_i}{q_i}) \geq 0$$



Figuur III.1: Entropie van een Bernoulli experiment afhankelijk van de kans p op succes.

met gelijkheid alleen maar als $p_i = q_i$ voor alle i. Men noemt $D(P,Q)$ de *relatieve entropie* of *Kullback-Leibler afstand* tussen P en Q.

De relatieve entropie $D(P,Q)$ geeft aan, hoe veel bits we gemiddeld extra nodig hebben, omdat we de codering van de gegevens op grond van de (verkeerde) kansverdeling Q in plaats van P hebben gekozen. Er geldt namelijk

$$H(P) + D(P,Q) = -\sum_{i=1}^{n} p_i \ {}^2\log(p_i) + \sum_{i=1}^{n} p_i \ {}^2\log(\frac{p_i}{q_i}) = -\sum_{i=1}^{n} p_i \ {}^2\log(q_i)$$

en dit is juist de verwachtingswaarde van het aantal benodigde bits op grond van de kansverdeling Q.

Merk op: De naam Kullback-Leibler *afstand* voor de relatieve entropie is een beetje misleidend, omdat we het niet met een afstand zo als de gewone Euclidische afstand in het vlak of in de ruimte te maken hebben.

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 relatieve entropie heeft alleen maar de eerste van deze drie eigenschappen. Maar met een eenvoudig trucje kunnen we van de relatieve entropie wel een symmetrische functie maken, namelijk door

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

Ook dit heet meestal de Kullback-Leibler afstand van P en Q, soms iets duidelijker de *symmetrische Kullback-Leibler afstand*.

Ook al voldoet de Kullback-Leibler niet aan de driehoeksongelijkheid, zijn $D(P,Q)$ of $d_{KL}(P,Q)$ toch vaak handig om te kwantificeren hoe sterk verschillende kansverdelingen Q op een vaste (doel-)kansverdeling P lijken.

11.2 Entropie van continue kansverdelingen

We hebben ons tot nog toe tot discrete kansverdelingen beperkt. 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:

$$H(X) := -\int_{-\infty}^{\infty} f(x)\ {}^2\log(f(x))\ dx.$$

> Om duidelijk te maken dat het om de entropie van een continue variabel gaat, spreekt men vaak ook van *differentiële entropie*. Het idee achter deze naam is, de variabel x te discretiseren door de waarden in het interval $[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}]$ aan de discrete waarde x_i toe te wijzen en de kans over dit interval als kans $p_i := \int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}} f(x)\ dx$ te definiëren. Met de overgang $\Delta x \to 0$ komt men dan naar de continue versie van de entropie.

Ook de *relatieve entropie* of *Kullback-Leibler afstand* van twee stochasten X en Y met dichtheidsfuncties $f(x)$ en $g(x)$ wordt analoog met het discrete geval gedefinieerd, namelijk door

$$D(X,Y) := \int_{-\infty}^{\infty} f(x)\ {}^2\log(\frac{f(x)}{g(x)})\ dx.$$

Met hetzelfde argument als bij de discrete kansverdelingen geldt weer

$$D(X,Y) \geq 0 \quad \text{en} \quad \int_{-\infty}^{\infty} f(x)\ {}^2\log(g(x))\ dx = H(X) + D(X,Y).$$

Men ziet makkelijk in dat de entropie van een stochast X onafhankelijk van de verwachtingswaarde $\mu := E[X]$ is, want met de substitutie $x' = x + a$ volgt dat de verschoven stochast $X + a$ dezelfde entropie als X heeft. Aan de andere kant heeft de variantie $Var(X) = E[X^2 - E[X]^2]$ zeker een invloed op de entropie, want hoe sterker de resultaten vaan X verspreid zijn, hoe onzekerder zijn we over de uitkomsten van X.

Bij discrete kansverdelingen hadden we gezien, dat onder de verdelingen met n mogelijke uitkomsten de uniforme verdeling de hoogste entropie heeft. De equivalente vraag voor continue kansverdelingen is, welke verdeling met gegeven variantie σ^2 de grootste entropie heeft.

Het lijkt misschien enigszins verrassend dat we ook deze vraag kunnen beantwoorden, want we moeten een uitspraak over alle mogelijke dichtheidsfuncties maken. Maar er laat zich aantonen dat bij gegeven variantie de normale

verdeling de maximale entropie heeft, dus dat we bij de normale verdeling de grootste onzekerheid over de mogelijke uitkomsten hebben.

Stelling: Onder alle continue kansverdelingen met variantie σ^2 heeft de normale verdeling

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

de maximale entropie.

> Het idee onder zekere randvoorwaarden de kansverdeling met maximale entropie te bepalen, geeft aanleiding tot een alternatieve manier om parameters van een probabilistisch model te schatten. In Wiskunde 1 hadden we hiervoor al de *maximum likelihood* methode leren kennen, waarbij de parameters zo bepaald worden dat de kans op de waargenomen resultaten maximaal is. Bij de toegang middels maximale entropie worden de parameters zo gekozen, dat de entropie maximaal wordt, dus het wordt het meest algemene model verondersteld dat de waarnemingen verklaart.
>
> Vaak is een algemeen model ook eenvoudiger dan een speciaal model en heeft het voordeel enigszins robuust tegen uitschieters in het training materiaal te zijn. Het principe om onder gegeven randvoorwaarden het eenvoudigste model te kiezen staat ook bekend onder de naam *Ockham's razor* (na de filosoof William van Ockham (1285-1349): 'The simplest explanation is the best.').

Voor het bewijs van de stelling dat de normale verdeling de maximale entropie heeft, is een techniek namens *variatierekening* nodig. Hierbij gaat het om het vinden van extrema van functies, die niet van een of meerdere variabelen afhangen maar van een continue hoeveelheid variabelen, anders gezegd om maxima en minima van functies, die zelfs ook weer van functies afhangen. We zullen in deze cursus geen variatierekening behandelen, maar schetsen wel even het idee.

In ons geval willen we een maximum van de functie

$$H(f) = -\int_{-\infty}^{\infty} f(x)\ {}^2\log(f(x))\ dx$$

vinden, die van de dichtheidsfunctie $f = f(x)$ afhangt. Hierbij moet $f(x)$ aan zekere randvoorwaarden voldoen, namelijk dat het een dichtheidsfunctie is, dat de variantie σ^2 is, en we mogen nog veronderstellen dat de verwachtingswaarde $\mu = 0$ is. We moeten dus een maximum van $H(f)$ vinden onder de randvoorwaarden:

(i) $f(x) \geq 0$;

(ii) $\int_{-\infty}^{\infty} f(x)\ dx = 1$;

(iii) $\int_{-\infty}^{\infty} x f(x)\ dx = 0$;

(iv) $\int_{-\infty}^{\infty} x^2 f(x)\, dx = \sigma^2$.

Dit is natuurlijk typisch een situatie voor Lagrange multiplicatoren, we definiëren daarom de Lagrange functie

$$L(f) = -\int_{-\infty}^{\infty} f(x)\log(f(x))\, dx + \lambda_0(\int_{-\infty}^{\infty} f(x)\, dx - 1) + \lambda_1(\int_{-\infty}^{\infty} xf(x)\, dx) + \lambda_2(\int_{-\infty}^{\infty} x^2 f(x)\, dx - \sigma^2).$$

Hierbij vergeten we even de randvoorwaarde $f(x) \geq 0$, die zal uiteindelijk van zelfs goed komen. We werken met de natuurlijke logaritme $\log(x)$ in plaats van de logaritme met basis 2, omdat dit voor het bepalen van de afgeleiden handiger is.

Om de kritieke punten van de Lagrange functie $L(f)$ te vinden, moeten we nu de partiële afgeleiden naar de variabelen bepalen, dus naar de functiewaarden $f(x)$ van de dichtheidsfunctie. Merk op dat x in dit geval een constante en geen variabel is, de variabelen zijn juist de functiewaarden op gegeven punten x. We moeten nu $L(f)$ voor een vaste x naar $f(x)$ afleiden en dit gelijk aan 0 zetten. Omdat we hierbij alleen maar naar een enkele waarde van x kijken, mogen we de integralen in $L(f)$ meteen weglaten. We krijgen

$$\begin{aligned}\frac{\partial L}{\partial f(x)} &= -\log(f(x)) - f(x)\cdot\frac{1}{f(x)} + \lambda_0\cdot 1 + \lambda_1\cdot x + \lambda_2\cdot x^2 \\ &= -\log(f(x)) - 1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2.\end{aligned}$$

Uit $\frac{\partial L}{\partial f(x)} = 0$ volgt nu $\log(f(x)) = -1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2$ en dus

$$f(x) = e^{-1+\lambda_0+\lambda_1 x+\lambda_2 x^2}.$$

Maar dit betekent dat $f(x)$ juist een normale verdeling is en volgens de randvoorwaarden moeten de constanten λ_0, λ_1, λ_2 zo gekozen worden dat de verwachtingswaarde 0 en de variantie σ^2 wordt, en dit is juist voor

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

het geval.

We hebben tot nu toe alleen maar aangetoond dat de normale verdeling een kritieke waarde voor de Lagrange functie is. Maar als we nu veronderstellen dat $g(x)$ de dichtheidsfunctie van een stochast Y met verwachtingswaarde 0 en variantie σ^2 is, kunnen we aantonen dat $H(Y) \leq H(X)$ is, dus dat de entropie

voor de normale verdeling inderdaad een maximum aanneemt:

$$\begin{aligned}
H(Y) &= -\int_{-\infty}^{\infty} g(x)\log(g(x))\,dx = -\int_{-\infty}^{\infty} g(x)\log(\frac{g(x)}{f(x)}\cdot f(x))\,dx \\
&= -\int_{-\infty}^{\infty} g(x)\log(\frac{g(x)}{f(x)})\,dx - \int_{-\infty}^{\infty} g(x)\log(f(x))\,dx \\
&= -D(Y,X) - \int_{-\infty}^{\infty} g(x)\log(f(x))\,dx \\
&\leq_{(*)} -\int_{-\infty}^{\infty} g(x)(-1+\lambda_0+\lambda_1 x+\lambda_2 x^2)\,dx \\
&=_{(**)} -\int_{-\infty}^{\infty} f(x)(-1+\lambda_0+\lambda_1 x+\lambda_2 x^2)\,dx \\
&= -\int_{-\infty}^{\infty} f(x)\log(f(x))\,dx = H(X).
\end{aligned}$$

Bij $(*)$ hebben we toegepast dat de relatieve entropie $D(Y,X) \geq 0$ is, en bij $(**)$ dat X en Y kansverdelingen met dezelfde verwachtingswaarde en variantie hebben, dus dat $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} g(x)\,dx = 1$, $\int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{\infty} xg(x)\,dx = 0$ en $\int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 g(x)\,dx = \sigma^2$.

Als voorbeelden vergelijken we de entropie van een normale verdeling met variantie σ^2 met de entropie van een uniforme verdeling met dezelfde variantie.

Entropie van de normale verdeling

Zij X een normaal verdeelde stochast met verwachtingswaarde μ en variantie σ^2, dan heeft X de dichtheidsfunctie $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Voor de entropie van X geldt

$$\begin{aligned}
H(X) &= -\int_{-\infty}^{\infty} f(x)\ {}^2\log(f(x))\,dx \\
&= -\int_{-\infty}^{\infty} f(x)({}^2\log(\frac{1}{\sqrt{2\pi}\,\sigma}) + {}^2\log(e^{-\frac{(x-\mu)^2}{2\sigma^2}}))\,dx \\
&= -{}^2\log(\frac{1}{\sqrt{2\pi}\,\sigma})\underbrace{\int_{-\infty}^{\infty} f(x)\,dx}_{=1} - \frac{1}{\log(2)}\int_{-\infty}^{\infty} f(x)\log(e^{-\frac{(x-\mu)^2}{2\sigma^2}})\,dx \\
&= {}^2\log(\sqrt{2\pi}\,\sigma) - \frac{1}{\log(2)}\int_{-\infty}^{\infty} f(x)(-\frac{(x-\mu)^2}{2\sigma^2})\,dx \\
&= {}^2\log(\sqrt{2\pi}\,\sigma) + \frac{1}{\log(2)}\frac{1}{2\sigma^2}\underbrace{\int_{-\infty}^{\infty} f(x)(x-\mu)^2\,dx}_{=Var(X)=\sigma^2} \\
&= {}^2\log(\sqrt{2\pi}\,\sigma) + \frac{1}{2\log(2)} = {}^2\log(\sqrt{2\pi}\,\sigma) + {}^2\log(\sqrt{e}) \\
&= {}^2\log(\sqrt{2\pi e}\,\sigma)
\end{aligned}$$

Entropie van de uniforme verdeling

Zij X een stochast met uniforme verdeling op het interval $[-a,a]$, dus met dichtheidsfunctie $f(x)=\frac{1}{2a}$ voor $x\in[-a,a]$ en $f(x)=0$ voor $x\notin[-a,a]$. We moeten eerst de variantie van X berekenen, hiervoor geldt

$$Var(X)=\int_{-a}^{a}\frac{1}{2a}x^2\,dx=\frac{1}{2a}\cdot\frac{x^3}{3}\Big|_{-a}^{a}=\frac{1}{2a}\,\frac{2a^3}{3}=\frac{a^2}{3}.$$

De variantie is dus voor $a=\sqrt{3}\,\sigma$ gelijk aan σ^2.

Voor de entropie geldt nu

$$H(X)=-\int_{-a}^{a}\frac{1}{2a}\ {}^2\log(\frac{1}{2a})\,dx=-\ {}^2\log(\frac{1}{2a})=\ {}^2\log(2a).$$

Voor een uniforme stochast met variantie σ^2, dus met $a=\sqrt{3}\,\sigma$, krijgen we dus de entropie

$$H(X)=\ {}^2\log(\sqrt{12}\,\sigma).$$

Omdat $\sqrt{12}\approx 3.464<4.132\approx\sqrt{2\pi e}$ is de entropie bij een uniforme verdeling inderdaad kleiner dan bij een normale verdeling met dezelfde variantie.

11.3 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:

$H(X,Y)\le H(X)+H(Y)$ 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 $1\le i\le n$, $y_j:=p(Y=y_j)$ voor $1\le j\le m$ 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

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

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

We zien dat $H(X)+H(Y)$ alleen maar geldt als $p_iq_j = r_{ij}$ 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.

Definitie: De *voorwaardelijke entropie* van de stochast X onder de voorwaarde van de stochast Y is gedefinieerd door

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

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

(1) 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

$$\begin{aligned} H(X|X) &= -\sum_{j=1}^{n}\sum_{i=1}^{n} p(X{=}x_i|X{=}x_j)\ {}^2\log(p(X{=}x_i|X{=}x_j))p(X{=}x_j) \\ &= -\sum_{i=1}^{n} 1\cdot 0\cdot p(X=x_i) = 0. \end{aligned}$$

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.

(2) Als X en Y onafhankelijk zijn, dan geldt $p(X = x_i \mid Y = y_j) = p(X = x_i)$, en hieruit volgt

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

Voor onafhankelijke stochasten X en Y geldt dus

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

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.

11.4 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 behandelen.

Net als bij de entropie stellen we ook bij de informatie eisen aan een functie 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. Dit geeft aanleiding tot de eerste eis die we aan een functie voor de informatie hebben:

(1) Er geldt $I(X = x_i) = I(p_i)$ voor $p_i = p(X = x_i)$.

Verder bekijken we 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:

(2) Voor *onafhankelijke* stochasten X en Y met $p_i = p(X = x_i)$ en $q_j = p(Y = y_j)$ geldt $I(p_i q_j) = I(p_i) + I(q_j)$.

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:

Definitie: Voor een stochast X is de *informatie* van de uitkomst $X = x$ met $p(X = x) = p$ gegeven door

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

> Deze definitie van informatie is in ieder geval ook in overeenstemming met onze intuïtie dat het optreden van een gebeurtenis met een kleine kans meer informatie oplevert dan een gebeurtenis met een grote kans, dus van *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.

Verband tussen informatie en entropie

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. Maar in de taal van de kansrekening is dit gewogen gemiddelde juist de verwachtingswaarde:

Merk op: De entropie $H(X)$ van een stochast X is de verwachtingswaarde van de informatie van de enkele uitkomsten van de stochast.

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 $n = 2^I$, waarbij I de informatie is die in een enkele uitkomst zit. Maar we hebben net 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. Dit kunnen we ook als volgt formuleren:

Merk op: De onzekerheid bij een stochast X is even groot als de onzekerheid bij een uniforme verdeling met $2^{H(X)}$ mogelijke uitkomsten. Anders gezegd is $2^{H(X)}$ het gemiddelde aantal alternatieven, dat we bij een kansexperiment voor de stochast X verwachten.

We hebben in het begin van deze sectie gesteld, dat het verschil van de entropieën $H(X) - H(X \mid Y)$ de informatie is, die Y over X onthult. 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 dit is ook zinvol omdat kennis van X de onzekerheid over X precies moet compenseren. Aan de andere kant geldt voor onafhankelijke stochasten X en Y dat $I(X \mid Y) = 0$, want $H(X \mid Y) = H(X) + H(Y)$. Ook dit is juist wat we nodig hebben, want onafhankelijke stochasten mogen onderling geen informatie onthullen.

Bij de definitie van $I(X \mid Y)$ 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 gedefinieerd door

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

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 onthult de stochast X net zo veel informatie over Y als de stochast Y over X onthult.

11.5 Toepassing: Automatische Taalherkenning

Als voorbeeld voor de toepassing van de concepten van entropie en informatie bekijken we het probleem van de automatische taalherkenning op geschreven tekst. Voor een mens is dit meestal nauwelijks een 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 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 $p_1 := p(X = \texttt{A})$, $p_2 := p(X = \texttt{B})$, $\ldots$, $p_{26} := p(X = \texttt{Z})$, $p_{27} := p(X = \textit{spatie})$.

Tabel III.1 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. 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 $l_{gem} = \frac{1}{p} - 1$) en men ziet dat de woorden in het Fins gemiddeld duidelijk langer zijn dan in de andere talen.

Een betere voorstelling van de frequentieverdelingen dan met de tabel krijgt men door de verdelingen als histogrammen te plotten, zo als in Figuur III.2 te zien. Hier valt bijvoorbeeld op, dat er in het Fins meer letters met een relatief hoge frequentie zijn, en dat in het Nederlands en Duits de letter `E` met duidelijke afstand de hoogste frequentie heeft.

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%

Tabel III.1: Letter frequenties voor vier verschillende talen

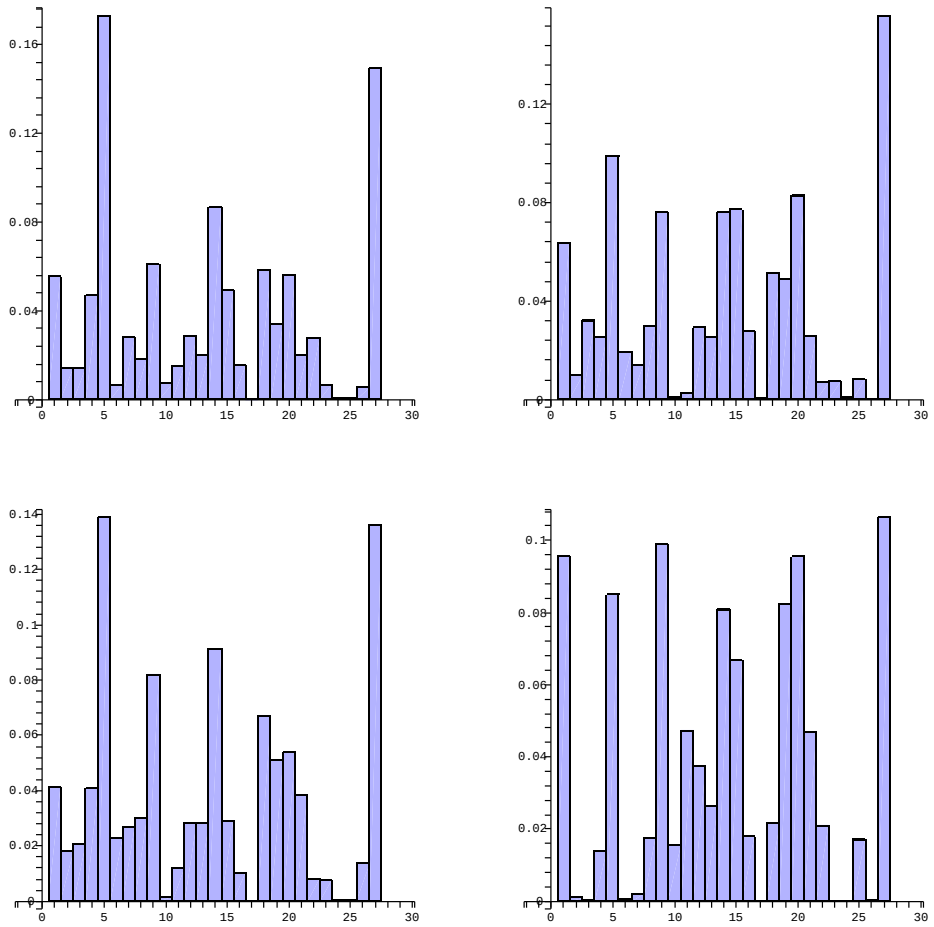
Als we de frequentieverdelingen als kansverdelingen opvatten, kunnen we voor de verschillende talen de entropieën van deze verdelingen uitrekenen, dit geeft de volgende waarden:

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

Met de interpretatie van de entropie met behulp van informatie geeft dit:

$$2^{H(\text{Nederlands})} = 16.21, \qquad 2^{H(\text{Engels})} = 16.80,$$
$$2^{H(\text{Duits})} = 17.26, \qquad 2^{H(\text{Fins})} = 15.80.$$

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.



Figuur III.2: Letter-frequentieverdelingen voor Nederlands (links boven) en Engels (rechts boven), Duits (links onder) en Fins (rechts onder).

Classificatie van patronen

Een typisch 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 ons voorbeeld van de automatische taalherkenning zijn de klassen natuurlijk de talen en het nieuwe patroon is een nieuwe tekst.

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, bijvoorbeeld (hidden) Markov modellen of/en neuronale netwerken.

Er zijn verschillende mogelijkheden voor de rol die kansverdelingen bij het classificeren van patronen kunnen spelen:

- Het nieuwe patroon wordt door een vector (of een rij vectoren) in de *kenmerkruimte* (feature space) weergegeven. De klassen zijn gerepresenteerd door kansverdelingen op de kenmerkruimte die aangeven hoe groot de kans is dat een patroon met een zekere 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 eerder berekende kansverdeling van de klasse lijkt.

We zullen de tweede insteek nu eens nader bekijken, omdat die minder voor de hand liggend lijkt als de eerste. In het voorbeeld van de automatische taalherkenning zijn de kansverdelingen gegeven door de relatieve frequenties van de letters. Voor een nieuwe tekst waarvan we de taal willen bepalen moeten we daarom ook de frequentieverdeling berekenen en vervolgens deze kansverdeling met de bekende kansverdelingen van de verschillende talen vergelijken. 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.

Afstanden tussen kansverdelingen

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 machtswortel 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.

> Het idee de afstand tussen kansverdelingen met behulp van een inproduct te berekenen 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 vectoren gekeken. Het enige wat van de kansverdelingen over blijft, is dat de som van de componenten 1 is, dus dat $\sum_{i=1}^{n} p_i = 1$.

Kullback-Leibler afstand

Maar natuurlijk hebben we eerder in deze les ook al een maat voor de afstand tussen kansverdelingen gezien, namelijk de Kullback-Leibler afstand (of relatieve entropie).

We hadden gezien dat de Kullback-Leibler afstand $D(P,Q)$ het verschil tussen $-\sum p_i \ {}^2\log(q_i)$ en de entropie $H(P)$ van de kansverdeling P aangeeft, dus dat

$$D(P,Q) = \sum_{i=1}^{n} p_i \ {}^2\log(\frac{p_i}{q_i}) = (-\sum_{i=1}^{n} p_i \ {}^2\log(q_i)) - H(P).$$

Als we nu $2^{H(P)}$ als het gemiddelde aantal alternatieven interpreteren, die we bij een stochast X met kansverdeling P verwachten, kunnen we ook de Kullback-Leibler afstand herinterpreteren: Er geldt

$$2^{H(P)+D(P,Q)} = 2^{H(P)} \cdot 2^{D(P,Q)},$$

dus is $2^{D(P,Q)}$ de factor waarmee we het gemiddelde aantal alternatieven moeten vermenigvuldigen, omdat we de *verkeerde* kansverdeling Q in plaats van P veronderstellen.

De volgende tabellen geven links de Kullback-Leibler afstanden tussen de talen uit het voorbeeld met de frequentieverdelingen en rechts de factoren $2^{D(P,Q)}$. Hierbij betekent bijvoorbeeld een factor 1.138 een afwijking van 13.8% van het aantal verwachte alternatieven bij de juiste kansverdeling. Merk op dat de tabellen niet symmetrisch zijn, omdat we de gewone Kullback-Leibler afstand $D(P,Q)$ en niet de symmetrische versie $d_{KL}(P,Q)$ toepassen.

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 zekere hoeveelheid training materiaal de kansen p_i worden bepaald en vervolgens een probabilistisch model gebouwd wordt, dat van enkele parameters afhangt. Dit kan bijvoorbeeld een normale verdeling 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

- onzekerheid, entropie
- relatieve entropie, Kullback-Leibler afstand
- entropie bij continue kansverdelingen
- maximale entropie bij normale verdeling
- voorwaardelijke entropie
- informatie
- afstanden tussen kansverdelingen

Opgaven

84. 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.

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

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

87. 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 $H(Z) < H(X, Y)$.

88. 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

$$H(X) = -n(p\ {}^2\log(p) + (1-p)\ {}^2\log(1-p)).$$

89. Laat zien dat de entropie $H(X)$ van een continue stochast X met een exponentiële verdeling met dichtheidsfunctie

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \text{ voor } x \geq 0$$

gegeven is door

$$H(X) = {}^2\log(\lambda\, e).$$

90. Bij een *best-of-five* tennis match is de speler de winnaar die als eerste drie sets gewonnen heeft. Stel dat de spelers A en B (ongeveer) even sterk zijn, zo dat een set met kans $\frac{1}{2}$ door A of B gewonnen wordt.

    Zij X de stochast die de mogelijke rijtjes van gewonnen sets beschrijft, dus bijvoorbeeld AAA, $ABBAA$ of $ABBB$. Verder zij Y de stochast die het aantal benodigde sets aangeeft (en dus een van de waarden 3, 4 of 5 heeft).

    Bepaal de entropieën $H(X)$ en $H(Y)$ en de voorwaardelijke entropieën $H(Y \mid X)$ en $H(X \mid Y)$.

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

92. 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 onthult?

93. 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 onthult de haarkleur van zijn vrouw?