Entropia

S. Roman : Introduction to coding & information theory, Springer 1997

Entropia jako miara "ilości informacji źródła".

Źródło : $S = (\{s_1, \dots, s_n\}, \{p_1, \dots, p_n\})$
(obiekty / symbole) rozkład prawdopodobieństwa $P(s_i) = p_i$; $\sum p_i = 1$, $p_i \geq 0$

Ilość informacji zdarzenia p_i : $I(p_i)$

Założenia $I : (0, 1] \longrightarrow (0, +\infty)$

1. $I(p) \geq 0$
2. I jest funkcją ciągłą
3. s_i, s_j są generowane niezależnie : $I(p_i p_j) = I(p_i) + I(p_j)$

Tw. Jedyną funkcją $I : (0, 1] \longrightarrow (0, +\infty)$ spełniającą warunki (2) i (3) jest $I(p) = C \log_2 \frac{1}{p}$ gdzie C jest stałą.

Dowód

a. Funkcja $I(p) = C \log \frac{1}{p}$ spełnia warunki (1)-(3).

b. Jeśli $I : (0,1] \to (0,+\infty)$ spełnia (2)-(3) to $I(p) = C \log_2 \frac{1}{p}$.

- $I(p^2) = I(p \cdot p) = I(p) + I(p) = 2I(p)$

...

$I(p^n) = n\, I(p)$ dla $n \geq 1$

- $I\left((p^{\frac{1}{n}})^n\right) = n\, I(p^{1/n}) \longrightarrow I(p^{\frac{1}{n}}) = \frac{1}{n} I(p)$

- $I(p^{\frac{n}{m}}) = \frac{1}{m} I(p^n) = \frac{n}{m} I(p)$ dla $n, m \geq 1$

Wobec tego dla każdej liczby wymiernej $q > 0$: $I(p^q) = qI(p)$

- Jeśli $r > 0$ jest liczbą niewymierną to istnieje ciąg $\{q_n\}$ liczb wymiernych zbieżny do r:

$q_n \xrightarrow[n\to\infty]{} r$. Wobec tego $p^{q_n} \xrightarrow[n\to\infty]{} p^r$.

Z ciągłości I : $I(p^r) = I\left(\lim_{n\to\infty} p^{q_n}\right) = \lim_{n\to\infty} I(p^{q_n}) = \lim_{n\to\infty} q_n I(p) = rI(p)$

Niech $p \in (0,1)$ będzie ustaloną liczbą. Dla każdego q ($q \in (0,1)$) mamy $q = p^{\log_p q}$.

Stąd $I(q) = I\left(p^{\log_p q}\right) = I(p) \cdot \log_p q =$

$= \left(\frac{I(p)}{\log_2 \frac{1}{p}}\right) \cdot \log_2 \frac{1}{q} = C \cdot \log_2 \frac{1}{q}$

C

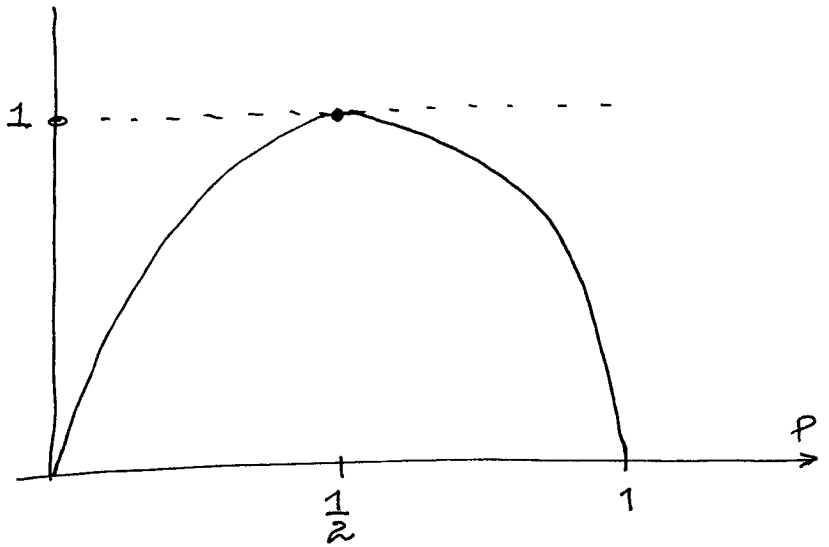
$\downarrow$

$\log_2 q = \log_p q \cdot \log_2 p$

$\log_p q = \frac{\log_2 q}{\log_2 p} = \frac{\log_2 \frac{1}{q}}{\log_2 \frac{1}{p}}$

Przykład $P = \{p, 1-p\}$

$$H(P) = p \log_2 p + (1-p) \log_2 \frac{1}{1-p}$$



$$H_r(S) \stackrel{def}{=} \sum_{i=1}^{n} p_i \log_r \frac{1}{p_i} = \frac{H(S)}{\log_2 r}$$

Entropia źródła S — średnia wartość informacji

$$H(S) = \sum_{i=1}^{n} p_i \, I(p_i) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Uwaga: przyjmujemy dla $p_i = 0$ $\quad p_i \log \frac{1}{p_i} = 0$

Inny zapis: $H(S) = H(s_1, \dots, s_n) = H(p_1, \dots, p_n)$

Lemat 1. $P = \{p_1, \dots, p_n\}$, $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$
$R = \{r_1, \dots, r_n\}$, $r_i \geq 0$, $\sum_{i=1}^{n} r_i \leq 1$
$\Bigg\} \longrightarrow \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \leq \sum_{i=1}^{n} p_i \log_2 \frac{1}{r_i}$

Dowód

$$\sum_{i=1}^{n} p_i \log_2 \frac{r_i}{p_i} \leq \frac{1}{\ln 2} \sum_{i=1}^{n} p_i \left(\frac{r_i}{p_i} - 1 \right) = \frac{1}{\ln 2} \left(\sum_{i=1}^{n} r_i - \underbrace{\sum_{i=1}^{n} p_i}_{1} \right) \leq 0$$

$\uparrow$

$$\log_2 x \leq \frac{x-1}{\ln 2} \quad \text{dla } x > 0$$

$\downarrow$ równość: $x = 1$

Tw. $0 \leq H(S) \leq \log_2 n$

Dowód. Stosujemy lemat dla $R = \{\frac{1}{n}, \dots, \frac{1}{n}\}$: $H(S) \leq \sum_{i=1}^{n} p_i \log_2 n = \log_2 n$

Entropia (przykłady)

System informacyjny $\mathcal{A} = (U, A)$; $a : U \to V_a$, dla $a \in A$.

	a_1 ...	a_i	a_n
x_1			
⋮			
x_j		$a_i(x_j)$	
⋮			
x_n			

$U = \{x_1, \ldots, x_n\}$

→ relacja nierozróżnialności $IND(A) =$

$\{(x_i, x_j) \in U \times U : \forall a \in A\,(a(x_i) = a(x_j))\}$

wyznacza podział U:

$U/IND(A) = \{C_1, \ldots, C_k\}$

$$p_i = \frac{|C_i|}{U} \quad (i = 1, \ldots, k)$$

$$E(A) = -\sum_{i=1}^{n} p_i \log p_i$$

1. Entropia $E(A) = 0$ jeśli podział $U/IND(A)$ redukuje się do jednej klasy nierozróżnialności
2. $E(A)$ ma wartość największą gdy $IND(A) = \{(x, x) : x \in U\}$.

Entropia warunkowa

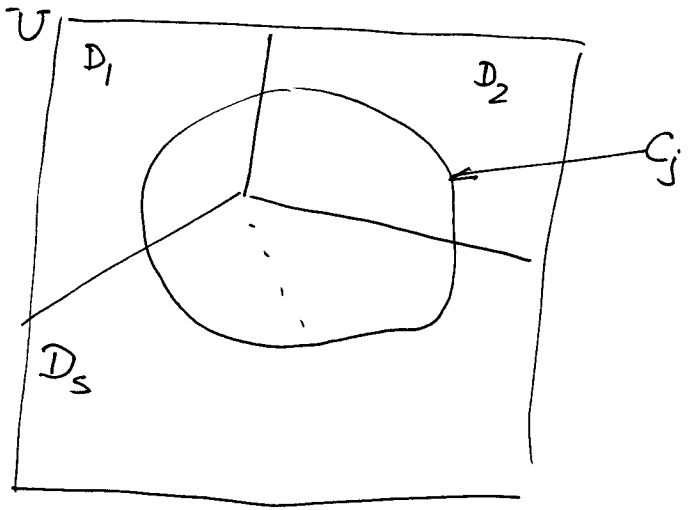
$DT = (U, A, d)$ → decyzja wyznacza podział U zdefiniowany przez $IND(d)$;

System informacyjny określa podział U wyznaczony przez relację nierozróżnialności $IND(A)$

Elementy podziału nazywane są klasami decyzyjnymi

$\mathcal{D} = U/IND(d) = \{D_1, .., D_s\}$

$\mathcal{C} = U/IND(A) = \{C_1, .., C_k\}$



$$E(\mathcal{D}|C_j) \stackrel{df}{=} -\sum_{i=1}^{s} p(D_i|C_j) \log p(D_i|C_j)$$

$$E(\mathcal{D}|\mathcal{C}) \stackrel{df}{=} \sum_{j=1}^{k} p(C_j) E(\mathcal{D}|C_j) =$$

$$= -\sum_{j=1}^{k} \sum_{i=1}^{s} \cancel{p(C_j)} \frac{p(D_i \cap C_j)}{\cancel{p(C_j)}} \log \frac{p(D_i \cap C_j)}{p(C_j)} =$$

$$= -\sum_{j=1}^{k} \sum_{i=1}^{s} p(D_i \cap C_j) \log (D_i \cap C_j) + \sum_{j=1}^{k} \left(\sum_{i=1}^{s} p(D_i \cap C_j) \right) p(C_j) =$$

$$= E(\mathcal{D} \vee \mathcal{C}) - E(\mathcal{C}).$$

gdzie $\mathcal{D} \vee \mathcal{C} = \{D_i \cap C_j : D_i \in \mathcal{D} \,\&\, C_j \in \mathcal{C} \,\&\, D_i \cap C_j \neq \emptyset\}$.

$\mathcal{D}$ jest rozdrobnieniem $\mathcal{C}$: $\mathcal{C} \leq \mathcal{D} \overset{df}{\longrightarrow} \forall D \in \mathcal{D}\ \exists C \in \mathcal{C}(D \subseteq C)$

$\mathcal{C} \leq \mathcal{D} \longrightarrow H(\mathcal{C}) \leq H(\mathcal{D})$.

Schemat kodowania dla źródła $S = (\{s_1, .., s_n\}, \{p_1, .., p_n\})$ nad alfabetem A

$$f : \{s_1, \dots, s_n\} \to A^*$$

s_1	. . . .	s_n
c_1		c_n

$$c_i = f(s_i) \in A^*$$

$$l_i = \text{długość}(c_i)$$

$$r = |A|$$

$$\text{AveCodeLen}_f(S) \stackrel{\text{def}}{=} \sum_{i=1}^{n} l_i p_i$$

inne oznaczenie: $\text{AveCodeLen}_C(S)$ gdzie $C = \{c_1, .., c_n\}$

Tw (Shannon, 1948)

$$H_r(S) \leq \text{Min AveCodeLen}_f(S) < H_r(S) + 1$$

↓

minimum po wszystkich schematach kodowania dla S nad alfabetem A mającym r liter przy założeniu że schematy są jednoznaczne wstępnie

jeśli $w \in A^*$ daje się przedstawić w postaci $w = w_1 \dots w_k$ gdzie $w_i \in \{c_1, .., c_n\}$ dla $i = 1, .., k$ to wtedy ten fakt jednoznaczny

Lemat 2 Niech A będzie alfabetem r literowym oraz niech $C = \{c_1, \dots, c_n\}$ będzie schematem kodującym dla $S = (\{s_1, \dots, s_n\}, \{p_1, \dots, p_n\})$ skończonego rozkładu.

Wtedy $$\sum_{i=1}^{n} \frac{1}{r^{l_i}} \leq 1$$

Dowód (indukcja po ~~prawej~~ lewej strona co tw.)

Niech C będzie schematem kodującym jak w lemacie. Przyjmijmy $v_i = \frac{1}{r^{l_i}}$.

Mamy $\sum_{i=1}^{n} v_i \leq 1$ (z poprzedniego lematu).

Wobec tego $$H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \underset{\text{nierówność poprzednie (Lemat 1)}}{\leq} \sum_{i=1}^{n} p_i \log \frac{1}{v_i} = \sum_{i=1}^{n} p_i l_i \log_2 r$$

$$= (\log_2 r) \, AveCodeLen_C(S)$$

Czyli $H_r(S) \leq AveCodeLen_C(S)$ dla każdego jednoznacznie rozkładalnego schematu kodującego C dla S nad alfabetem r literowym.

Lemat 3. Istnieje jednoznaczne optymalny schemat kodujący C_0 dla S nad alfabetem o r symbolach ($r \geq 2$) taki, że

$$-\log_r p_i \leq l_i < -\log_r p_i + 1$$

[ten kod nazywa się kodem Shannona–Fano]

Dowód (mieszany po pracy domowej twierdzenie)

$$\text{Ave Code Len}_{C_0}(S) = \sum_{i=1}^{n} p_i l_i < \sum_{i=1}^{n} p_i \left(\log_r \frac{1}{p_i} + 1 \right) = H_r(S) + 1$$

Wobec tego $\text{Min Ave Code Len}_f(S) < H_r(S) + 1$.

Lemat

Niech $C = (c_1, \dots, c_n)$ będzie schematem kodowym nad alfabetem r-literowym dla zbioru $S = (\{s_1, \dots, s_n\}, \{p_1, \dots, p_n\})$.

Jeśli C jest jednoznacznie odkodowalny to $\sum_{k=1}^{n} \frac{1}{r^k} \leq 1$.

Dowód. α_j - liczba słów kodowych o dł. j; $m = \max_i(l_i)$, $l_i = \text{długość}(c_i)$

$$\sum_{i=1}^{n} \frac{1}{r^k} = \sum_{j=1}^{m} \frac{\alpha_j}{r^j}$$

u - k. naturalne > 0

$$\left(\sum_{j=1}^{m} \frac{\alpha_j}{r^j}\right)^u = \left(\frac{\alpha_1}{r} + \frac{\alpha_2}{r^2} + \dots + \frac{\alpha_m}{r^m}\right)^u = \sum_{(i_1,\dots,i_u):\ 1 \leq i_j \leq m} \frac{\alpha_{i_1}}{r^{i_1}} \cdot \dots \cdot \frac{\alpha_{i_u}}{r^{i_u}} = \sum_{\substack{(i_1,\dots,i_u):\\ 1 \leq i_j \leq m}} \frac{\alpha_{i_1} \cdots \alpha_{i_u}}{r^{i_1+\dots+i_u}}$$

$1 \leq i_j \leq m \rightarrow \quad u \leq i_1 + \dots + i_u \leq u \cdot m$

$$= \sum_{k=u}^{u \cdot m} \left(\sum_{(i_1,\dots,i_u):\ i_1+\dots+i_u=k} \alpha_{i_1} \cdots \alpha_{i_u} \right) \frac{1}{r^k}$$

Dla ustalonego ciągu $(i_1, \ldots, i_n)$ liczbę $\alpha_{i_1} \ldots \alpha_{i_n}$ jest liczbą słów o dł. $i_1 + \ldots + i_n$ rozkładalnych na słowa kodowe $W_1, \ldots, W_n$ o dł. $i_1, \ldots, i_n$, odpowiednio.

a suma $\sum_{(i_1, \ldots, i_n) : i_1 + \ldots + i_n = k} \alpha_{i_1} \ldots \alpha_{i_n}$

jest liczbą słów o dł k rozkładalnych na n słów kodowych. Oznaczmy tę liczbę przez N_k

Nie każde słowo o dł k jest rozkładalne ale póki jest to

jednoznacznie. Stąd

$$\sum_{(i_1, \ldots, i_n) : i_1 + \ldots + i_n = k}{}' \alpha_{i_1} \ldots \alpha_{i_n} \leq \nu^k$$

gdzie ν^k jest liczbą słów o dł k nad alfabetem r literowym ($r \geq 2$).

Wobec tego

$$\left(\sum_{k=1}^{m} \frac{\alpha_k}{r^k}\right)^u \leq \sum_{k=u}^{u\cdot m} \frac{N_k}{r^k} \leq \sum_{k=u}^{u\cdot m} 1 \leq u\cdot m$$

$$\sum_{k=1}^{m} \frac{\alpha_k}{r^k} \leq u^{\frac{1}{u}} m^{\frac{1}{u}} \quad \text{dla kazdego } u>0 \text{ naturalnego}$$

$u\to\infty \downarrow \quad \downarrow u\to\infty$

$1 \qquad 1$

czyli $\sum_{k=1}^{m} \frac{\alpha_k}{r^k} \leq 1$

Kod Shannon'a-Fano

$p_i > 0$; $\sum p_i = 1$

$p_1 \geq p_2 \geq \ldots \geq p_n$

$$E : \{1, \ldots, n\} \longrightarrow \{0,1\}^r \quad ; \quad A = \{0,1\}$$

Procedura obliczania $E(i)$ dla $1 \leq i \leq n$

1. $P_1 = 0$

$$P_i = \sum_{j=1}^{i-1} p_i \quad \text{dla } 1 < i \leq n$$

2. Wyznaczamy kolejne wzmocnienia P_i

3. Odnajdujemy wyrazy rozwinięcia powyżej funkcji $\ell(E(i))$ (gdzie $\ell(E(i))$ oznacza długość kodu $E(i)$) ~~gdzie~~ takiej, że

$$-\log p_i \leq \ell(E(i)) < 1 - \log p_i$$

Przykład $E = \{1, 2, 3, 4\}$

$p_1 = \frac{1}{2}$	$P_1 = 0$
$p_2 = \frac{1}{4}$	$P_2 = p_1 = \frac{1}{2}$
$p_3 = \frac{3}{16}$	$P_3 = p_1 + p_2 = \frac{3}{4}$
$p_4 = \frac{1}{16}$	$P_4 = p_1 + p_2 + p_3 = \frac{15}{16}$
$\Sigma = 1$	

Trzeba wyznaczyć kody $E(1), E(2), E(3), E(4) \in \{0,1\}^*$

$\boxed{E(1) = 0}$

Wyznaczamy rozwinięcie binarne P_1, czyli 0 i odmierzamy rozwinięcie długości $l(E(1))$

$$-\log \tfrac{1}{2} \leq l(E(1)) < 1 - \log \tfrac{1}{2}$$

$$1 \leq l(E(1)) < 1 + 1$$

$$\underline{l(E(1)) = 1}$$

$E(2)$

Wyznaczamy rozwinięcie binarne $P_2 = \frac{1}{2}$ i odmierzamy rozwinięcie długości $l(E(2))$

$$-\log \tfrac{1}{4} \leq l(E(2)) < 1 - \log \tfrac{1}{4}$$

$$\underline{l(E(2)) = 2}$$

$P_2 = 0.10$, $\times$ $\boxed{E(2) = 10}$

$\boxed{E(3) = 110}$

$P_3 = \frac{3}{4}$ i odmierzamy rozwinięcie długości $l(E(3))$

$$-\log \tfrac{3}{16} \leq l(E(3)) < 1 - \log \tfrac{3}{16}$$

$$2.\tfrac{1}{>0} \leq l(E(3)) < 3.\tfrac{1}{>0}$$

$$l(E(3)) = 3$$

$E(4) = 1111$

Z nierówności

$$-\log p_i \leq l(E(i)) < 1 - \log p_i$$

$$-l(E(i)) \leq \log p_i$$

$$2^{-l(E(i))} \leq p_i < 2^{1-l(E(i))} \qquad (*)$$

Kod $E(i)$ różni się od $E(i+1), E(i+2), \ldots, E(n)$ na co najmniej jednej pozycji, bo jeśli $i+1 \leq j \leq n$ to

$$P_j = P_i + p_i + \cdots + p_{j-1} \geqslant$$

$$P_i + p_i \geqslant P_i + 2^{-l(E(i))} \qquad (**)$$

$\uparrow$ z $(*)$

Wobec tego $E(i), E(j)$ $(i < j)$ różnią się na pierwszych $l(E(i))$ pozycjach.

Stąd 1) E jest $1-1$

2) $E(i)$ nie jest przedrostkiem zadnego $E(j)$ dla $j > i$

- w p.p. kod $E(i)$ dla i byłby przedrostkiem $E(j)$ dla pewnego $j > i$, a stąd

$$P_j - P_i \leq 0.\underbrace{0\ldots0}_{i}\,\underbrace{1\ldots1}_{i+1\ \ j} =$$

$$\leq \frac{1}{2^{i+1}}\left(1 + \frac{1}{2} + \cdots\right) = \frac{1}{2^{i+1}}\,\frac{1}{1-\frac{1}{2}} = \frac{1}{2^i}$$

co przeczy $(**)$. Wobec tego kodowanie E jest jednoznacznie rozkodowalne.

Niech $H_1 = \sum_{i=1}^{n} p_i\, l(E(i))$ i $H(S) = -\sum_{i=1}^{n} p_i \log p_i$.

Pozostało pokazać, że $H(S) \leq H_1$. Mamy $H_1 \leq \sum_{i=1}^{n} (1 - \log p_i) p_i =$

$$= \sum_{i=1}^{n} p_i + H(S) = 1 + H(S).$$