

Entropia: nieporządek czy fantazja?

Rafał SZTENCEL*

Termin „entropia” występuje w tak wielu dziedzinach nauki, że nie mogło zabraknąć go i w rachunku prawdopodobieństwa. Na ogół dużą entropię kojarzymy słusznie z nieporządkiem, wręcz bałaganem (który ma jedną miłą cechę – jest stanem stabilnym, co nie powinno dziwić, bo w pobliżu maksimum entropii żadne wysiłki już jej znacząco nie zwiększą).

W teorii prawdopodobieństwa może lepiej mówić o niepewności lub różnorodności, związanej z rozkładem. Dla prostoty zajmijmy się rozkładami dyskretnymi: niech liczby p_1, p_2, \dots, p_n wyznaczają rozkład prawdopodobieństwa (zatem $p_1 + p_2 + \dots + p_n = 1$, $p_i > 0$ dla $i = 1, 2, \dots, n$). Wartości liczbowe skojarzone z prawdopodobieństwami p_i są zupełnie nieistotne; zresztą nie muszą to być liczby, a mogą to być, na przykład, imiona pań, które pojawiły się na pewnym protokole egzaminacyjnym. Imion tych było $n = 45$.

Gdybyśmy chcieli odgadnąć, jak ma na imię losowo wybrana dziewczyna, znając możliwe imiona i zadając pytania, na które można odpowiedzieć tak/nie, to 5 pytań mogłoby nie wystarczyć ($2^5 = 32$), ale 6 pytań wystarczy na pewno, bo wszystkich możliwych ciągów odpowiedzi jest $2^6 = 64$. Czytelnik na pewno wie, jak rozsądnie zadawać pytania, żeby ich średnia liczba mieściła się pomiędzy 5 i 6.

Kluczową rolę w tym rozumowaniu odgrywała nierówność

$$32 < 45 < 64,$$

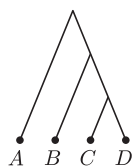
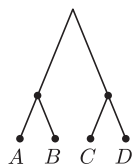
równoważna z

$$\log_2 32 < \log_2 45 < \log_2 64,$$

czyli

$$5 < \log_2 45 = 5,49 < 6.$$

Można zatem podejrzewać, że logarytm dwójkowy z liczby możliwości n faktycznie jakoś mierzy różnorodność. Ale przecież nie wzięliśmy pod uwagę tego, że imiona występują z różnymi częstościami. W naszym protokole cztery najczęstsze imiona to Anna – 14%, Joanna, Katarzyna i Magdalena – po 8%. Jak można z tego skorzystać przy zadawaniu pytań?



Rozważmy prostszą sytuację. Jeśli są tylko cztery imiona: Agnieszka, Barbara, Celina i Dorota, pojawiające się jednakowo często, to średnio (i zawsze) potrzebne są 2 pytania. Niech teraz A. ma częstość 50%, B. – 25%, C. i D. – po 12,5%. Każdy widzi, że z prawdopodobieństwem $\frac{1}{2}$ wystarczy jedno pytanie, $\frac{1}{4}$ – dwa, $\frac{1}{4}$ – trzy. Średnio jest $1\frac{3}{4}$ pytania.

Zapiszmy tę średnią tak:

$$q = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 =$$

$$= -\frac{1}{2} \cdot \log_2 \frac{1}{2} - \frac{1}{4} \cdot \log_2 \frac{1}{4} - \frac{1}{8} \cdot \log_2 \frac{1}{8} - \frac{1}{8} \cdot \log_2 \frac{1}{8} = 1\frac{3}{4}.$$

Średnia liczba pytań okazała się równa

$$H = -\sum_{i=1}^n p_i \log_2 p_i,$$

i to jest właśnie entropia rozkładu prawdopodobieństwa. Można udowodnić, że

$$(*) \quad H \leq q < H + 1,$$

zatem entropia H jest dolnym ograniczeniem średniej liczby pytań q .

Entropia jest największa, gdy wszystkie p_i są równe. Wynika to z wklęsłości funkcji $f(x) = -x \log_2 x$ i nierówności Jensena:

$$t_1 f(x_1) + t_2 f(x_2) + \dots + t_n f(x_n) \leq f(t_1 x_1 + t_2 x_2 + \dots + t_n x_n),$$

gdzie liczby t_i są dodatnie i dają w sumie 1.

W takim razie

$$\begin{aligned} H &= f(p_1) + f(p_2) + \dots + f(p_n) = \\ &= n \left(\frac{1}{n} f(p_1) + \frac{1}{n} f(p_2) + \dots + \frac{1}{n} f(p_n) \right) \leq \\ &\leq n f \left(\frac{p_1 + p_2 + \dots + p_n}{n} \right) = n f \left(\frac{1}{n} \right) = \log_2 n. \end{aligned}$$

Jasne jest, że $H = \log_2 n$, gdy $p_i = \frac{1}{n}$, $n = 1, 2, \dots$

W przypadku 45. imion z protokołu faktyczna entropia jest równa 4,76, podczas gdy maksymalna możliwa to $\log_2 45 = 5,49$; niewielka różnica obu liczb świadczy o dużej fantazji rodziców przy nadawaniu imion dziewczynom.

Pozostaje pytanie, czy jest prosty sposób na zadawanie pytań tak, by spełniona była nierówność (*). O tym za miesiąc – doprowadzi nas to do tak zwanego kodu Huffmana (dajemy słowo honoru, że Leonardo da Vinci nie mógł go znać).

*Instytut Matematyki, Uniwersytet Warszawski