Informacijska entropija - uvod

Mario Cagalj

University of Split

Uvod

Informacijska **entropija** mjera je povezana sa slučajnom varijablom i interpretira se kao prosječan **sadržaj informacije** ili prosječna **razina iznaneđenja** povezana s ishodom te slučajne varijable.

Example

Koja od sljedećih izjava je informativnija (više iznenađuje)?

1. *Elon Musk upisuje računarstvo na FESB-u.*
2. *Učenik MIOC-a Marko Matić upisuje računarstvo na FESB-u.*

Intuitivno, oni ishodi slučajne varijable koji su manje vjerojatni imaju veći sadržaj informacije (faktor iznaneđenja).

Diskretna slučajna varijabla

Razmatramo diskretnu slučajnu varijablu K čiji ishodi su definirani konačnim skupom $\mathcal{K} = \{k_1, k_2, \ldots k_n\}$, tj. $K \in \mathcal{K}$. Učestalost pojedinih ishoda slučajne varijable K definirana je **distribucijom vjerojatnosti** $P(K = k_i) = p_i, i = 1, \ldots, n$. Budući da se K mora realizirati, vrijedi $\sum_{i=1}^{n} p_i = 1$.

Primjer (*bacanje novčića*)

$K \in \{\text{pismo}, \text{glava}\}, P(K = \text{pismo}) = \frac{1}{2}, P(K = \text{glava}) = \frac{1}{2}$

Primjer (*bacanje pristranog novčića*)

$K \in \{\text{pismo}, \text{glava}\}, P(K = \text{pismo}) = \frac{2}{3}, P(K = \text{glava}) = \frac{1}{3}$

Primjer (*slučajni ℓ-bitni enkripcijski ključ*)

$K \in \{0, 1, \ldots, 2^\ell - 1\}, P(K = k) = \frac{1}{2^\ell}, k = 0, 1, \ldots, 2^\ell - 1$

Entropija diskretne slučajne varijable

Za diskretnu slučajnu varijablu $K \in \mathcal{K}$, gdje je $\mathcal{K} = \{k_1, k_2, \ldots k_n\}$, entropija $H(K)$ dana je sljedećim izrazom:

$$\boxed{H(K) = -\sum_{k \in \mathcal{K}} P(K = k) \log_2 P(K = k)}$$

- Entropija $H(K)$ izražava se u **bitovima** (za bazu logaritma 2)
- Zadovoljava, $0 \leq H(K) \leq log_2|\mathcal{K}| = log_2 n$
- Ostvaruje maksimalnu vrijednost samo ako svi ishodi slučajne varijable K imaju jednaku vjerojatnost pojavljivanja, tj. $P(K = k_i) = \frac{1}{|\mathcal{K}|} = \frac{1}{n}, i = 1, 2, \ldots, n$

Primjeri

Primjer (*bacanje novčića*)

$$H(K) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \text{ bit}$$

Primjer (*bacanje pristranog novčića*)

$$H(K) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.92 \text{ bit}$$

Primjer (*slučajni ℓ-bitni enkripcijski ključ*)

$$H(K) = -\sum_{k=0}^{2^\ell-1}\frac{1}{2^\ell}\log_2\frac{1}{2^\ell} = \log_2 2^\ell = \ell \text{ bits}$$

Primjer - entropija zaporke

Problem

Korisnik bira zaporku iz rječnika $\mathcal{D}$:

$$\mathcal{D} = \{\text{password}, \text{iloveyou}, \ldots, \text{test4321}\}.$$

Veličina rječnika je $|\mathcal{D}| = 1073741824 = 2^{30}$. Prosječna zaporka iz rječnika duga je 7 znakova, odnosno $7 \cdot 8 = 56$ bita.

Kolika je entropija nepoznate zaporke, odnosno kolika je vaša neizvjesnost vezana uz nepoznatu zaproku?

Primjer - entropija zaporke

Rješenje

Točnu entropiju ne možemo izračunati jer ne znamo s kojom vjerojatnošću korisnik bira pojedine zaporke. Međutim, možemo konzervativno procjeniti maksimalnu entropiju odnosno maksimalnu neizvjesnost o korisnikovoj zaporci.

Konzervativnom pretpostavkom da se zaporke iz $\mathcal{D}$ biraju s jednakom vjerojatnošću $1/|\mathcal{D}| = 2^{-30}$, dobijemo:

$$H(Z) \leq -\sum_{z \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \log_2 \frac{1}{|\mathcal{D}|} = \log_2 |\mathcal{D}| = 30 \text{ bit}$$

Važno: Naša neizvjesnost vezana uz nepoznatu zaporku je 30 bita (značajno manja od prosječne dužine zaporke - 56 bita).

Primjer - entropija zaporke

Razmatramo rječnik zaporki $\mathcal{D}$ veličine $|\mathcal{D}| = 2^{30}$:

- Korisnici preferiraju jednu zaporku iz $\mathcal{D}$, entropija zaporke:

$$H(Z) = 0$$

- Korisnici biraju zaporke samo iz polovice rječnika, entropija zaporke:

$$H(Z) \leq \log_2 \frac{|\mathcal{D}|}{2} = \log_2 |\mathcal{D}| - 1 = 30 - 1 = 29 \text{ bit}$$

- Korisnici biraju zaporke iz 2 puta većeg rječnika $\mathcal{D}'$, entropija zaporke:

$$H(Z) \leq \log_2 |\mathcal{D}'| = \log_2 2^{31} = 31 \text{ bit}$$

Primjetite: $\log_2 |\mathcal{D}'| = \log_2 |\mathcal{D}| + 1$