Informační entropie

Andrew Kozlík

KA MFF UK

Házení mincí

Experiment: k-krát po sobě hodíme mincí a výsledky hodů zaznamenáme jako posloupnost hodnot 0 (rub) a 1 (líc).

$$1001101\ldots01$$

- Dostaneme posloupnost z množiny $\{0,1\}^k$.
- Každá z možných posloupností má pravděpodobnost 2^{-k}.
- Zřejmě jde o nejkompaktnější možný zápis výsledků mezi všemi zápisy, které využívají symboly 0 a 1.
- Můžeme říct:
  - „Informační hodnota zápisu je k bitů.“
  - „Náhodná veličina s rovnoměrným rozdělením na množině $\{0,1\}^k$ má míru nejistoty k bitů.“

Házení dvěma mincemi

Experiment: k-krát po sobě hodíme dvěma *nerozlišitelnými* mincemi současně.

Zaznamenáme výsledky hodů:

Výsledek	Pravděpodobnost	Kód
rub a rub	$1/4$	00
rub a líc	$1/2$	01
líc a líc	$1/4$	11

Například:

$$01\ 00\ 11\ 01\ 01 = 0100110101$$

- Záznam bude mít délku $2k$.
- Nevyužíváme plně kapacitu zápisu, protože kód 10 se nepoužije.
- Například posloupnost 011001 je neplatná.

Lepší záznam házení dvěma mincemi

Výsledek (rub a líc) má dvakrát vyšší pravděpodobnost výskytu než ostatní výsledky. Měl by mít kratší zápis:

Výsledek	Pravděpodobnost	Kód A	Kód B
rub a rub	$1/4$	10	00
rub a líc	$1/2$	1	1
líc a líc	$1/4$	01	01

Kód A má problém: $10\ 1 = 101 = 1\ 01$
Kód B funguje dobře: $00\ 1 = 001 = 00\ 1$
U kódování B lze poznat, kde každý kód končí:

- Kód začínající symbolem 0 má vždy délku dvou symbolů.
- Kód začínající symbolem 1 má délku jednoho symbolu.
- Například $011001 = 01\ 1\ 00\ 1$.

Vlastnosti kódování B

- Z každé posloupnosti lze jednoznačně rekonstruovat výsledky hodů. Říkáme, že toto kódování je *prosté*.
- Délka výsledné posloupnosti nebude jednoznačně určena počtem hodů k.
- Víme, že délka záznamu bude ležet v intervalu $[k, 2k]$.
- Průměrná délka záznamu bude $\frac{1}{4}k \cdot 2 + \frac{1}{2}k \cdot 1 + \frac{1}{4}k \cdot 2 = \frac{3}{2}k$.
- Dá se ukázat, že toto kódování je z hlediska délky optimální.
- Průměrné množství informace získané z jednoho hodu dvěma nerozlišitelnými mincemi je tedy 1,5 bitů.

Informační entropie

Definice

Nechť $X : \Omega \to A$ je diskrétní náhodná veličina.
Potom *entropie* náhodné veličiny X je definována jako

$$\mathrm{H}(X) := -\sum_{a \in A} \Pr(X{=}a) \cdot \log_2 \Pr(X{=}a),$$

kde předepisujeme $0 \cdot \log_2 0 := 0$.

Ospravedlnění předpisu pro 0:

- Limitním chováním: $\lim_{z \to 0_+} z \log_2 z = 0$.
- Selským rozumem: Je-li $X{=}a$ nemožný jev, pak by neměl zvyšovat míru nejistoty.

Shannonova věta o kódování zdroje

- Entropie diskrétní náhodné veličiny je dolní mez na průměrnou délku kteréhokoliv jejího prostého kódování.

- Vždy existuje prosté kódování, kterým se lze přiblížit k této dolní mezi libovolně blízko.
  (Je však třeba kódovat více výsledků najednou.)

Vlastnosti entropie

$$\mathrm{H}(X) = -\sum_{a \in A} \Pr(X{=}a) \cdot \log_2 \Pr(X{=}a)$$

Pozorování

- Má-li X rovnoměrné rozdělení, pak

$$\mathrm{H}(X) = -\sum_{a \in A} \frac{1}{|A|} \log_2 \frac{1}{|A|} = \log_2 |A|.$$

  V případě $|A| = 2^k$ tak dostáváme $\mathrm{H}(X) = k$.

- Existuje-li $a_0 \in A$ takové, že $\Pr(X{=}a_0) = 1$, pak $H(X) = -1 \cdot \log_2 1 = 0$.
  Jinými slovy, je-li výsledek jistý, pak je entropie nulová.

Huffmanovo kódování náhodné veličiny

Máme náhodnou veličinu $X : \Omega \to A$.

- Prvkům z A přiřazujeme kódy z $\{0,1\}^*$.
- Vlastnosti Huffmanova kódování:
  - Je prosté.
  - Průměrná délka zakódovaného prvku leží v intervalu $[\mathrm{H}(X), \mathrm{H}(X)+1)$.

Příklad

Mějme $A = \{1,2,3,4,5,6\}$ a rozdělení pravděpodobnosti:

a	1	2	3	4	5	6
$\Pr(X{=}a)$	0,05	0,10	0,15	0,12	0,13	0,45

Huffmanovo kódování

- Sestrojíme binární strom.
- Listy budou prvky z A.
- Váha listu $x \in A$ bude $\Pr(X{=}a)$.
- Váha vnitřního vrcholu bude součet vah jeho podvrcholů.
- Na začátku máme množinu vrcholů A bez hran.
  (Každý z nich je kořen jednoprvkového stromu.)
- Zvolíme dva nejlehčí kořeny a připojíme je pod nový kořenový vrchol. Opakujeme dokud nezůstane jediný kořen.
- Pro každý vrchol označíme jednu vycházející hranu symbolem 0 a druhou symbolem 1.
- Huffmanův kód $h(a)$ určíme tak, že přečteme symboly na cestě od kořene k listu a.

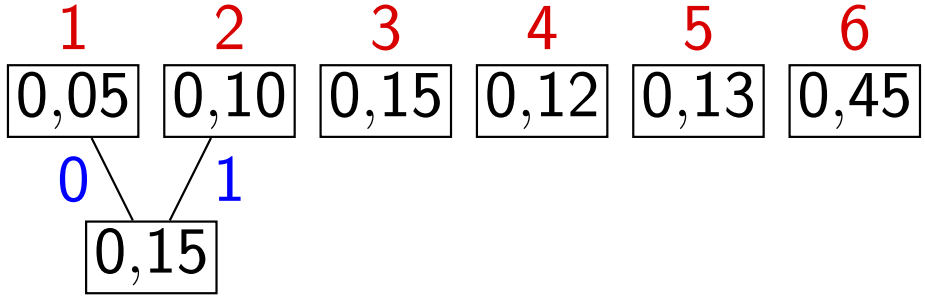
Huffmanův algoritmus.

1	2	3	4	5	6
0,05	0,10	0,15	0,12	0,13	0,45

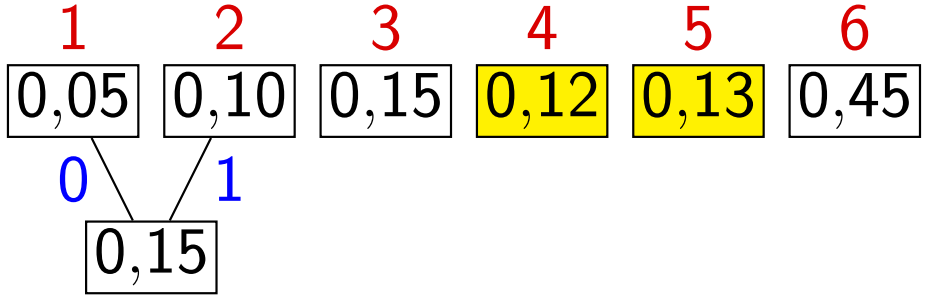
Huffmanův algoritmus.

1	2	3	4	5	6
0,05	0,10	0,15	0,12	0,13	0,45

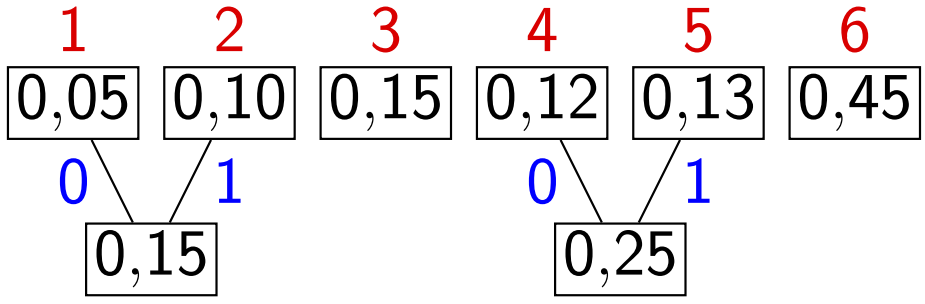
Huffmanův algoritmus.



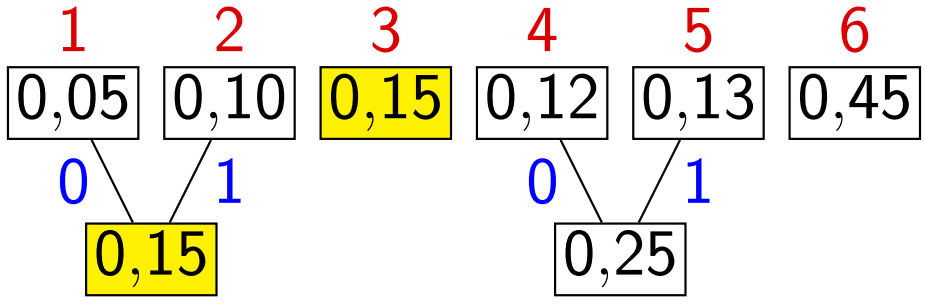
Huffmanův algoritmus.



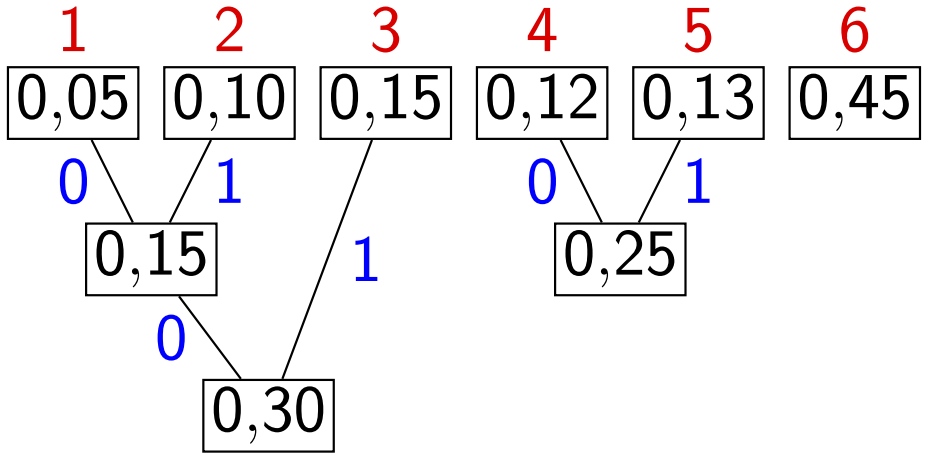
Huffmanův algoritmus.



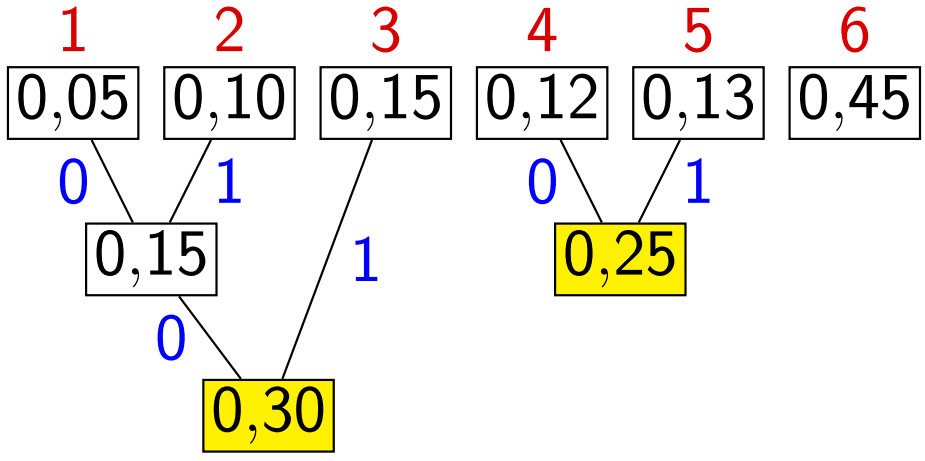
Huffmanův algoritmus.



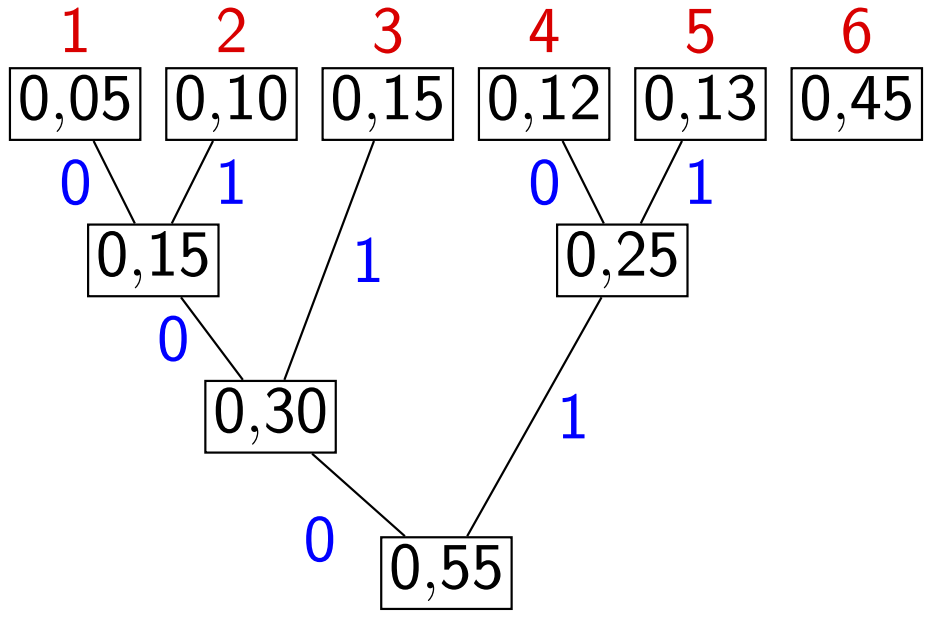
Huffmanův algoritmus.



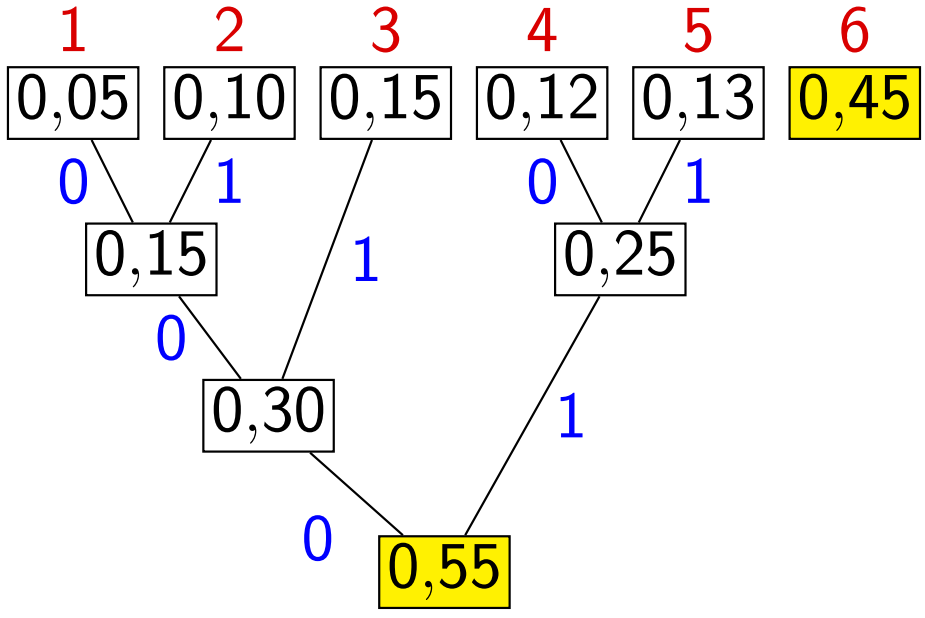
Huffmanův algoritmus.



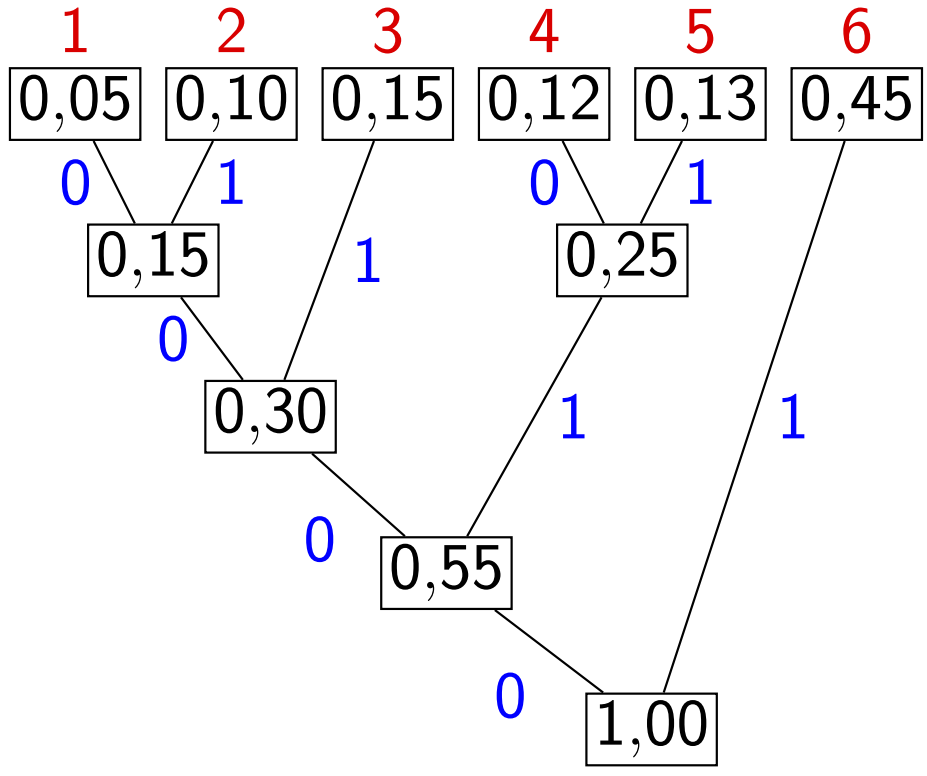
Huffmanův algoritmus.



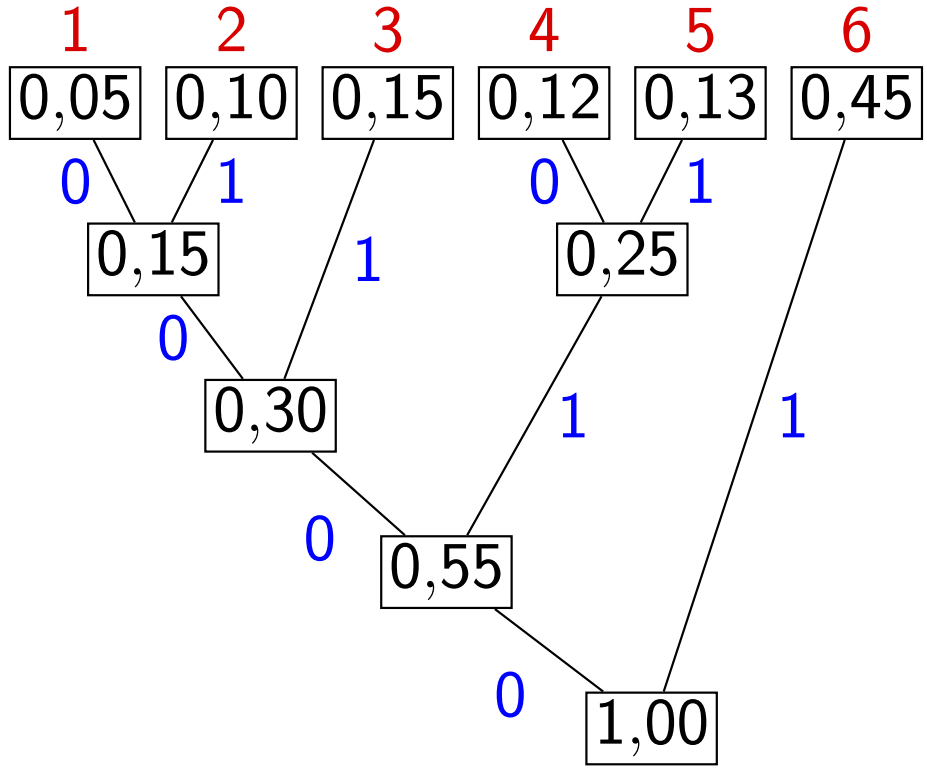
Huffmanův algoritmus.



Huffmanův algoritmus.



Huffmanův algoritmus.



a	$h(a)$
1	0000
2	0001
3	001
4	010
5	011
6	1

Entropie a střední délka kódu

a	$\Pr(X{=}a)$	$h(a)$	$\log_2 \Pr(X{=}a)$
1	0,05	0000	−4,32
2	0,10	0001	−3,32
3	0,15	001	−2,74
4	0,12	010	−3,06
5	0,13	011	−2,94
6	0,45	1	−1,15

- Délka kódu $\ell(h(a))$ je přibližně rovna $-\log_2 \Pr(X{=}a)$.
- Srovnání entropie a střední délky kódu:

$$\mathrm{H}(X) = -\sum_{a \in A} \Pr(X{=}a) \cdot \log_2 \Pr(X{=}a) \approx 2{,}23$$

$$\mathrm{E}\,\ell(h(X)) = \sum_{a \in A} \Pr(X{=}a) \cdot \ell(h(a)) = 2{,}25.$$

Věta (Jensenova nerovnost)

Nechť

- $f : I \to \mathbb{R}$ *je ryze konkávní funkce na intervalu* I,
- $x_1, \ldots, x_n \in I$,
- $\lambda_1, \ldots, \lambda_n \in (0, \infty)$,
- $\sum_{i=1}^{n} \lambda_i = 1$.

Potom

$$\sum_{i=1}^{n} \lambda_i f(x_i) \leq f\left(\sum_{i=1}^{n} \lambda_i x_i \right),$$

přičemž rovnost nastává právě tehdy, když $x_1 = x_2 = \cdots = x_n$.

Důkaz.

- V případě $n = 2$ se jedná o definici ryze konkávní funkce: $\lambda_1 f(x_1) + \lambda_2 f(x_2) < f(\lambda_1 x_1 + \lambda_2 x_2)$, $\lambda_1 + \lambda_2 = 1$, $x_1 \neq x_2$.
- Dále postupujeme indukcí podle n.

Věta (o maximální entropii)

Nechť $X : \Omega \to A$ je diskrétní náhodná veličina.
Potom

$$\mathrm{H}(X) \leq \log_2 |A|,$$

přičemž rovnost nastává, právě když X má rovnoměrné rozdělení.

Důkaz.

Využijeme Jensenovu nerovnost s $f = \log_2$.

$$\mathrm{H}(X) = \sum_{a \in A} \underbrace{\Pr(X{=}a)}_{\text{„}\lambda_i\text{“}} \cdot \log_2 \frac{1}{\underbrace{\Pr(X{=}a)}_{\text{„}x_i\text{“}}} \leq \log_2 \sum_{a \in A} 1.$$

Ověření předpokladů:

- Funkce $\log_2$ je ryze konkávní na $(0, \infty)$.
- $\sum_{a \in A} \Pr(X{=}a) = 1$.

Sdružená entropie

Definice

Nechť $X : \Omega \to A$ a $Y : \Omega \to B$ jsou diskrétní náhodné veličiny.

- *Kartézský součin* veličin X a Y definujeme jako

$$\begin{aligned}(X, Y) \; : \; & \Omega \to A \times B \\ & \omega \mapsto (X(\omega), Y(\omega)).\end{aligned}$$

- *Sdruženou entropii* veličin X a Y definujeme jako entropii jejich kartézského součinu a značíme ji $\mathrm{H}(X, Y)$.

Věta (o sdružené entropii)

Nechť $X : \Omega \to A$ a $Y : \Omega \to B$ jsou diskrétní náhodné veličiny. Potom

$$\mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y),$$

přičemž rovnost nastává, právě když X a Y jsou nezávislé.

Podmíněná entropie

Definice

Nechť

- $X : \Omega \to A$ a $Y : \Omega \to B$ jsou diskrétní náhodné veličiny a
- $E \subseteq \Omega$ je jev takový, že $\Pr(E) \neq 0$.

Entropii veličiny X podmíněnou jevem E definujeme jako

$$\mathrm{H}(X \mid E) := -\sum_{a \in A} \Pr(X{=}a \mid E) \log_2 \Pr(X{=}a \mid E).$$

Podmíněnou entropii $\mathrm{H}(X \mid Y)$ definujeme jako

$$\mathrm{H}(X \mid Y) := \sum_{b \in B} \Pr(Y{=}b)\, \mathrm{H}(X \mid Y{=}b).$$

Tvrzení

Nechť $X : \Omega \to A$ a $Y : \Omega \to B$ jsou diskrétní náhodné veličiny.
Potom $\mathrm{H}(X \mid Y) = \mathrm{H}(X, Y) - \mathrm{H}(Y)$.

Pozorování

Podle tohoto tvrzení a věty o sdružené entropii platí:

$$0 \;\leq\; \mathrm{H}(X \mid Y) \;=\; \mathrm{H}(X, Y) - \mathrm{H}(Y) \;\leq\; \mathrm{H}(X),$$

Rovnost nastává, právě když X a Y jsou nezávislé.

Důsledek

Nechť $X : \Omega \to A$ a $Y : \Omega \to B$ jsou diskrétní náhodné veličiny.
Potom:

1. $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$,
   přičemž rovnost nastává, právě když X a Y jsou nezávislé.
2. $\mathrm{H}(Y) \leq \mathrm{H}(X, Y)$.