Komprese dat aneb
Zdrojové kódování

Typy kódů s různou délkou

Notace: A ... konečná množina

A^* ... množina konečných posloupností prvků z A

tj.:

$$A^* = A \cup (A \times A) \cup \dots \cup (\underbrace{A \times \dots A}_{k\times}) \cup \dots$$

$$= A \cup A^2 \cup A^3 \times \dots$$

X - náhodná veličina s konečně mnoha hodnotami x

$$p(x) = P[X = x]$$

$\mathcal{D} = \{0, 1, \dots, D-1\}$ D-ární abeceda

Definice: Zdrojový kód C pro X je zobrazení z $\mathcal{X}$ do $\mathcal{D}^*$.

Značení: $c(x)$ - kódové slovo pro $x \in \mathcal{X}$

$l(x)$ - délka kódu $c(x)$

Příklad $X = \{\text{red}, \text{blue}\}$

$\mathcal{D} = \{0, 1\}$

$c(\text{red}) = 00 \qquad c(\text{blue}) = 11$

Definice: Střední délka, $l(C)$, kódu C je

$$L(C) = \sum_{x \in X} p(x)\, l(x)$$

Př: $\mathcal{D} = \{0,1\}$

$p[X=1] = 1/2$ $\quad c(1) = 0$

$p[X=2] = 1/4$ $\quad c(2) = 10$

$p[X=3] = 1/8$ $\quad c(3) = 110$

$p[X=4] = 1/8$ $\quad c(4) = 111$

$$L(C) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 = 1{,}75$$

$H(X) = 1{,}75.$

Každá posloupnost kódových slov (bez mezer) se dá dekódovat

0|110|111|10|0|110

1 3 4 2 1 3

Je-li $|\mathcal{D}| > |\mathcal{X}|$
můžeme jednoznačně přiřadit kódová slova všechny délky 1. Pak je $L(C)$ nejmenší možné

$$L(C) = \sum_{x \in \mathcal{X}} p(x) \cdot 1 = 1$$

Většinou ale máme malou abecedu

Definice **Extense** C^* kódu C je zobrazení $C^*: \mathcal{X}^* \to \mathcal{D}^*$ definované

$$C^*(x_1, x_2 \dots x_n) = \underbrace{C(x_1)\, C(x_2) \dots C(x_n)}$$

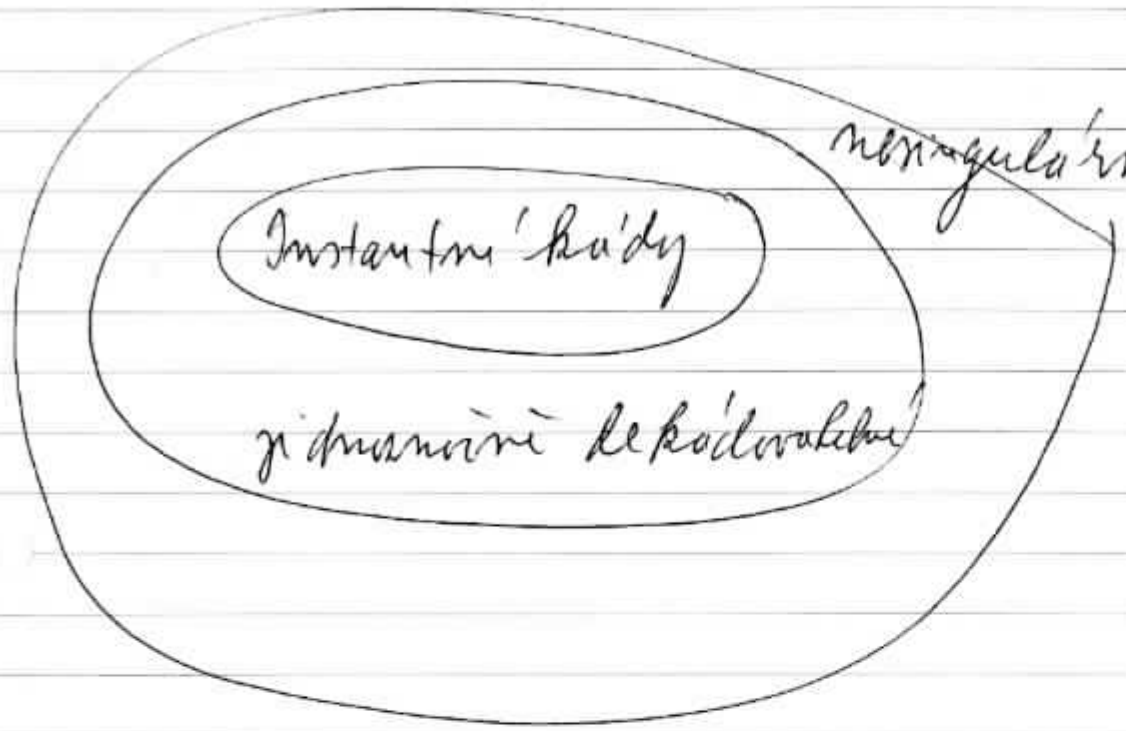
spojení kódových slov

Definice (1) Kód C je **nesingulární** je-li C prosté zobrazení z $\mathcal{X}$ do $\mathcal{D}^*$
tj. $x \neq x' \Rightarrow C(x) \neq C(x')$

(2) Kód C je **jednoznačně dekódovatelný** jestliže C^* je prosté zobrazení.

(3) Kód C je **instantní** jestliže žádné kódové slovo není předponou jiného kódového slova.

instantní = prefix code

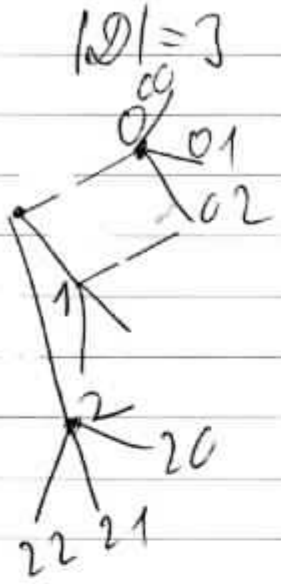


Příklad:

x			
1	0	0	10
2	0	010	00
3	0	01	11
4	0	10	110
	↓	↓	↓
	singulární	nesingulární, není instantní, není jedn. dekódovatelný	jednoznačně dekódovatelný, ale není instantní

010 — 2
— 14
\ 31

Stromová implementace kódu



užel = kód (mimo počátek)
délka = počet předků

Věta (Kraftova nerovnost)

Pro instantní kód C nad D-ární abecedou s délkami kódových slov

$l_1, l_2, \dots, l_m$

platí $\sum_{i=1}^{m} D^{-l_i} \leq 1$.

Naopak, k daným $l_1, \dots, l_m$ splňujícím nerovnost (*) existuje instantní kód s těmito délkami kódových slov.

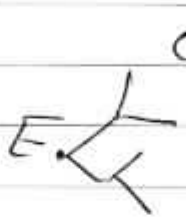
Důkaz č. 1

D - náš strom

kód = uzel

délka kódu = úroveň uzlu

E - uzel



$Ch(E)$ ... děti uzlu E

$Ch^{l}(E)$ ... děti uzlu E na úrovni l

Mějme $E_1, E_2, \dots, E_m$ uzly s délkami

$l_1, l_2, \dots, l_m$

Lze předpokládat, že $l_1 \leq l_2 \leq \dots \leq l_m = l_{max}$

- Kód je instantní $\Leftrightarrow E_i \notin Ch(E_j)\ \forall i \neq j$

$\Leftrightarrow Ch(E_i) \cap Ch(E_j) = \emptyset \quad i \neq j$

k bodu z předchozí strany zpětně z dvou různých úrovní dané úrovně $\Rightarrow$ dva různé $l_{max} - l_i$ - prvky E_j z dětí E_j

- $|Ch^{l_{max}}(E_i)| = D^{l_{max} - l_i} \quad ; \ i = 1, \dots, m$

- je-li kód instantní, pak

$\sum_{i=1}^{m} |Ch^{l_{max}}(E_i)| \leq D^{l_{max}}$ (z disjunktnosti)

tj. $\sum_{i=1}^{m} D^{l_{max} - l_i} \leq D^{l_{max}} \Rightarrow \sum_{i=1}^{m} D^{-l_i} \leq 1.$

Mějme naopak délky $l_1 \leq l_2 \leq \dots \leq l_m$ splňující $\sum_{i=1}^{m} D^{-l_i} \leq 1$. $\Leftarrow$ $\sum_{i=1}^{m} D^{l_{max}-l_i} \leq D^{l_{max}}$

Sestrojíme inst. kód s těmito délkami.

- Vyberme jakýkoliv uzel E_1 s délkou l_1.
  $D^{l_{max}-l_1} < D^{l_{max}}$ a tedy existuje uzel na úrovni l_{max}, který není v $Ch(E_1)$. Vezměme jeho otce E_2 na úrovni l_2. Musí platit $Ch(E_2) \cap Ch(E_1) = \emptyset$. Jelikož

$$|Ch^{l_{max}}(E_1) \cup Ch^{l_{max}}(E_2)| = D^{l_{max}-l_1} + D^{l_{max}-l_2} < D^{l_{max}}$$

existuje uzel E_3 na úrovni l_3 tak, že $Ch(E_3)$, $Ch(E_2)$, $Ch(E_1)$ jsou disjunktní atd. -

poslední krok: $\sum_{i=1}^{m-1} D^{l_{max}-l_i} < D^{l_{max}}$

a najdeme E_m.

□

Diskrétní č. 2

D-nární rozvoj

$y = y_1 \dots y_\ell \longrightarrow$ číslo $v \in \langle 0,1)$

$$y = \sum_{j=1}^{\ell} y_j D^{-j} =$$

kód: $y_i \in \{0, \dots, D-1\}$

$$= \frac{y_1}{D} + \frac{y_2}{D^2} + \dots + \frac{y_\ell}{D^\ell}$$

Všechna pokračování tohoto rozvoje jsou menší než

$$0, y_1 y_2 \dots y_\ell + \underbrace{\sum_{j=\ell+1}^{\infty} (D-1) D^{-j}}$$

$$= (D-1) \frac{1}{D^{\ell+1}} \cdot \frac{1}{1 - 1/D} = \frac{1}{D^\ell}$$

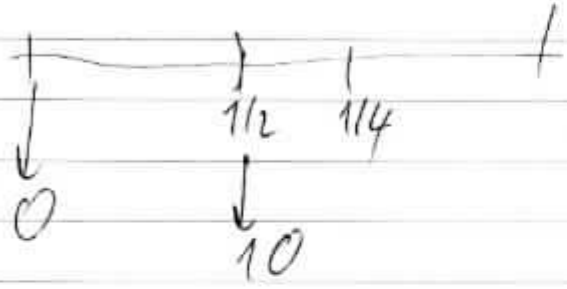
pokračování $\equiv \langle y, y + \frac{1}{D^\ell})$

Délka tohoto intervalu je $1/D^\ell$.

Mějme nyní instantní kód s délkami $\ell_1, \ell_2, \dots, \ell_n$.

Jim odpovídají disjunktní intervaly v $\langle 0,1)$ délky $D^{-\ell_1}, D^{-\ell_2}, \dots, D^{-\ell_n}$. Součet těchto délek je $\leq$ menší než 1.

Naopak, máme-li délky
$l_1 \leq l_2 \leq \dots \leq l_n$ splňující Kraftovu
nerovnost, jak jsme mohli přiřadit
intervaly od dalšího konce délky D^{-l_i}
např. binární kód $l_1 = 1, l_2 = 2$



0, 10 - je instantní kód.

□

Příklad Kolik prstů má Montáč?
Máme kódová slova instantního
kódu délek

1, 1, 2, 3, 2, 3

jak velká je abeceda?

Řešení: $f(D) = D^{-1} + D^{-1} + D^{-2} + D^{-3} + D^{-2} + D^{-3}$

$= 2D^{-1} + 2D^{-2} + 2D^{-3} \leq 1$

$f(2) > 1$

$f(3) = 2\left(\frac{1}{3} + \frac{1}{9} + \frac{1}{27}\right) = \frac{26}{27} \leq 1$

musíme vzít alespoň 3 znaky.
Pozorování: $l_1, \dots, l_m$ – délky instantního kódu pro D
$\Rightarrow$ jsou dobré i pro $D+1, D+2, \dots$

Překvapivý výsledek

Věta (McMillan, 1956)
Kódové délky $l_1, l_2, \dots, l_m$ jednoznačně dekódovatelného kódu s D-ární abecedou splňují nerovnost

$$\sum_{i=1}^{m} D^{-l_i} \leq 1.$$

Dk: C^+ … uzávěr kódu C : $C^+ \cdot C^+ \to D^*$

[illegible]

Kolik je nejvýše prvků z X^*, které C^+ zobrazí do D^m (kód délky m)
Jelikož je C^+ prosté je takových prvků nejvýše $|D^m| = D^m$.

$u \in X \dots l(u) \dots$ délka kódu

Fixujeme přirozené K

$$\left(\sum_{u \in X} D^{-l(u)} \right)^K =$$

$$\sum_{u_1 \in X} \sum_{u_2 \in X} \cdots \sum_{u_K \in X} D^{-l(u_1)} \cdot D^{-l(u_2)} \cdots D^{-l(u_K)}$$

$$= \sum_{(u_1, \dots, u_K) \in X^K} D^{-l(u_1 \cdots u_K)} = \sum_{m=1}^{K l_{max}} a(m) D^{-m} \leq$$

↓ počet posl. $u_1, \dots, u_K \in X^K$, které mají rozšířený kód délky m

Víme $|a(m)| \leq D^m$

$$\hookrightarrow \leq \sum_{m=1}^{K l_{max}} D^m \cdot D^{-m} = K l_{max}$$

$$\Rightarrow \sum_{u \in X} D^{-l(u)} \leq (K l_{max})^{1/K}$$

$\downarrow K \to \infty$

$$1$$

$$\Rightarrow \sum_{u \in X} D^{-l(u)} \leq 1.$$

Optimalita kódu

$x \in \mathbb{R}$

$\lceil x \rceil$ = nejmenší celé číslo $\geq x$.

Tedy $x \leq \lceil x \rceil < x+1$

Optimální kód je instantní kód (jednoznačně dekódovatelný) s nejmenší střední délkou.

Existence: $\vec{p} = (p_1, p_2, \dots, p_m)$

$p_1 \geq p_2 \geq \dots \geq p_m > 0$.

Ať L je délka konkrétního instantního kódu.

Pro jakýkoliv instantní kód jehož alespoň jedno kódové slovo má délku větší než L / p_m je jeho délka větší než L.

Lze se tedy omezit na kódy s $l_{max} \leq \lceil L/p_m \rceil$ a těch je konečně mnoho.

Def: Instantní kód C^* je optimální jestliže

$$L(C^*) \leq L(C)$$

pro každý instantní kód daného zdroje L^* optimální délka

---

máme-li dostatečně znaků (velké D) můžeme dosáhnout $L^* = 1$.

Čísluje další věta pod kterou nemůžeme jít:

Věta (Shannonova věta o zdrojovém kódování)

Střední délka L libovolného instantního D-márného kódu pro náhodnou veličinu X je větší nebo rovna

$$H_D(X):$$

$$L \geq H_D(X).$$

Rovnost nastává $\Leftrightarrow$ $p_i = D^{-l_i}$

Důkaz: Máme předpoklad $p_i > 0 \ \forall i$.

$l_1, l_2, \dots, l_m$ - délky kódových slov

$$L - H_D(X) = \sum_{i=1}^{m} p_i l_i - \sum_{i=1}^{m} p_i \log_D \frac{1}{p_i} =$$

$$= -\sum_{i=1}^{m} p_i \log_D D^{-l_i} + \sum_{i=1}^{m} p_i \log_D p_i$$

$$= \sum_{i=1}^{m} p_i \log_D \frac{p_i}{D^{-l_i}}$$

dle $c = \sum_{i=1}^{m} D^{-l_i} \leq 1$ - Kraft!

$r_i = \frac{D^{-l_i}}{c} \Rightarrow \vec{r} = (r_1, \dots, r_m)$

je pravděpodobnostní vektor.

$$L - H_D(X) = \sum_{i=1}^{m} p_i \log_D \frac{p_i}{r_i c} =$$

$$= \sum_{i=1}^{m} p_i \log_D \frac{p_i}{r_i} - \log_D c$$

$$= D(\vec{p} \| \vec{r}) + \log_D \frac{1}{c} \geq 0$$

$D(\vec{p}\|\vec{r}) \geq 0$.

$c \leq 1 \Rightarrow \log_D \frac{1}{c} \geq 0$

$\Rightarrow L \geq H_D(X)$.

Rovnost nastane $\Leftrightarrow$ $\vec{p} = \vec{r}$ a $c = 1$

□

- Optimální délka je nová entropie $\Leftrightarrow$ rozdělení je D-adické a Kraftova nerovnost je satisfovaná $p_i = D^{-l_i}$

- Jak blízko k entropii můžeme jít?

Shannonův kód:

$$l_i = \lceil \log_D \frac{1}{p_i} \rceil$$

Tyto délky splňují Kraftovu nerovnost:

$$\sum_{i=1}^{m} D^{-\lceil \log_D \frac{1}{p_i} \rceil} \leq \sum_{i=1}^{m} D^{-\log_D \frac{1}{p_i}} = \sum p_i = 1$$

Výběr délek vyhovuje nerovnosti

$$\log_D \frac{1}{p_i} \leq l_i < \log_D \frac{1}{p_i} + 1$$

$$\Downarrow$$

$$\sum_{i=1}^{m} p_i \log_D \frac{1}{p_i} \leq \sum_{i=1}^{m} l_i p_i < \sum_{i=1}^{m} p_i \log_D \frac{1}{p_i} + 1$$

Tedy $H_D(X) \leq L < H_D(X) + 1.$

Příklad

$\vec{p} = (1/2, 1/4, 1/8, 1/16, 1/16)$

Jaké jsou délky Shannonova kódu:

1, 2, 3, 4, 4.

střední délka

$$1/2 + 1/2 + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} = \frac{16+6+8}{16} = \frac{30}{16} = \frac{15}{8}$$

$$= 1,875$$

$H(\vec{p}) = 1,875$

Optimální délka může být jen lepší

$$H_D(X) \leq L^* \leq H_D(X) + 1$$

Věta: Pro optimální délku kódového slova L^* platí

$$H_D(X) \leq L^* < H_D(X) + 1.$$

- neexistuje formule pro výpočet L^*
- binární kód ≡ série ano a ne otázek
Shannon: průměrný počet otázek k identifikaci objektu daného rozdělení je roven entropii a šikovně lze zařídit jsou to prvé entropie + 1 bit

---

Příklad: Hráč A volí objekt ve vesmíru a hráč B identifikuje objekt sérií ano-ne otázek. V průměru přitom potřebuje 38,5 otázek k uhádnutí objektu.
Volí optimální strategii.
Jak velký je soubor objektů?

Řešení: $L^* = 38{,}5$

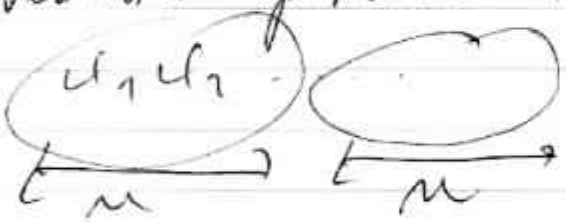
$\log |\mathcal{X}| \geq H(X) \geq L^* - 1$

$= |\mathcal{X}| \geq 2^{37{,}5} = 1{,}94 \cdot 10^{11}$

Blokové kódování

- lze dosáhnout ~~entropie~~ délky libovolně blízko entropii

Myšlenka: nekódujeme X ale $(X_1, X_2, \dots X_n)$ nezávislá opakování



L_n^* ... střední délka optimálního kódu pro $(X_1, X_2, \dots, X_n)$

$$H(X_1, X_2, \dots, X_n) \leq L_n^* \leq H(X_1, \dots, X_n) + 1$$

$$n H(X) \leq L_n^* \leq n H(X) + 1$$

$$H(X) \leq \frac{L_n^*}{n} \leq H(X) + \frac{1}{n}$$

$\frac{L_n^*}{n}$ ... střední délka na 1 symbol

Souvislost s rychlostí entropie – pokrač.

Cena za špatný kód:

X ... $x - p(x)$ rozdělení

Y ... $Y = x - q(x)$ rozdělení

Vyrobíme Shannonův kód dle $q(x)$

$$l(x) = \lceil \log \frac{1}{q(x)} \rceil$$

Věta (špatný kód) Střední délka, L, kódu s délkami

$$l(x) = \lceil \log \frac{1}{q(x)} \rceil$$

splňuje nerovnost

$$H(X) + D(X||Y) \leq L < H(X) + D(X||Y) + 1$$

Důkaz $L = \sum_{x \in X} p(x) l(x) =$

$$= \sum_{x \in X} p(x) \lceil \log \frac{1}{q(x)} \rceil \leq$$

$$\sum_{x \in X} p(x) \left[\log \frac{1}{q(x)} + 1 \right] =$$

$$= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \cdot \frac{1}{p(x)} + 1$$

$$= D(X \| Y) + H(X) + 1$$

Druhá nerovnost se dokáže stejně.

Příklad: binární kódy

Symboly	$p(x)$	$q(x)$	$c_1(x)$	$c_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

c_1 je optimální kód pro p:
$H(\vec{p}) = L(c_1) = 1,875$

c_2 je optimální kód pro q:
$H(\vec{q}) = 2 = L(c_2)$

Používáme kód c_2 pro distribuci $\vec{p}$
Kód c_2 je Shannonův pro $\vec{q}$
Kód c_1 je Shannonův pro $\vec{p}$

dostaneme délku
$$1/2 + 3/4 + 3/8 + \frac{6}{16} = \frac{8+12+6+6}{16} =$$
$$= \frac{32}{16} = 2$$

Rozdíl od $H(\vec{p}) = L(c_1)$ je $0,125$
Což je $D(p||q)$.