Huffmanovo kódování

Huffman nalezl v r. 1951 optimální kód C_H.

Příklad: $A = \{a, b\}$

$p(a) = 0{,}9999$

$p(b) = 0{,}0001$

Huffman: $C_H(a) = 0$

$C_H(b) = 1$

$L(C_H)$ = optimální délka = 1

Shannon: délky 1, 14

$L(C_S) = 0{,}9999 + 0{,}0014$) ↓

---

algoritmus Huffmanova kódování:

D - různé prvky

- Seřaďme symboly sestupně dle pravděpodobnosti
- Stanovme dva nejméně pravděpodobné symboly do jídelníčku a zastupme součtem pravděpodobností
- Pokračujme až dostaneme jazyk D znaků a ty "okódujme". $C_1 \dots D-1$.
- Zpětně kódujme větvemi!

Příklad: binární kód pro
$\vec{p} = (0,25; 0,25; 0,20; 0,15; 0,15)$

x	p(x)	x	p(x)	x	p(x)	x	p(x)	C(x)
1	0,25	(4,5)	0,30	(2,3)	0,45	(1,4,5) 0	0,55	01
2	0,25	1	0,25	(4,5) 00	0,30	(2,3) 1	0,45	10
3	0,20	2 10	0,25	1 01	0,25			11
4	0,15 000	3 11	0,20					000
5	0,15 001							001

kód

$L(C_H) = 2 \cdot 0{,}25 + 2 \cdot 0{,}25 + 2 \cdot 0{,}2$
$+ 3 \cdot 0{,}15 + 3 \cdot 0{,}15 = 2{,}3$

$H(X) =$

dělíme obálek: jsem v $\langle 1, 4, 5 \rangle$ nebo v $\langle 2, 3 \rangle$

↙

jsem 1 nebo $\langle 4, 5 \rangle$

↙

jsem 4 nebo jsem 5

Entropie: dolní hodnota pro průměrný počet otázek vedoucích k identifikaci algoritmu (znaku)

Příklad: $D = 3$

x	$p(x)$	$c(x)$
1	0,25	1
2	0,25	2
3	0,2	00
4	0,15	01
5	0,15	02

x	$\langle 3, 4, 5 \rangle$	1	2
$p(x)$	0,5	0,25	0,25

$L(C_H) = 1{,}5$

Příklad: $A = \{a, b, c, d\}$

$\vec{p} = (1/3, 1/3, 1/4, 1/12)$

Huffmanův kód dá délky
(2, 2, 2, 2) nebo (1, 2, 3, 3)

$L(C_H) =$

Příklad: zpráva obsahuje 10^4 znaků = 10^4 bytů
z abecedy $\{a, b, c, d, e\}$
s pravděpodobnostmi

$\vec{p} = (0,35;\ 0,17;\ 0,17;\ 0,16;\ 0,15)$

$L(C_H) = 2,3$

zkomprimovaná zpráva: 2 300 bitů

$10^4 \cdot L(C_H) = 2875$ bytů

$H(X) = 2,23284$ ... rozdíl asi 3 %

↓

lze zlepšit blokovým kódováním.

Kompresní poměr: 28,75 %.

výhody: optimalita
: snadná implementace

: není patentová ochrana
: je instantní

nevýhody: vyžaduje načtení celého souboru dat kvůli určení četnosti (ale existuje adaptivní kódování)

: nutno připojit kódovací tabulku

: nejsou explicitně dané délky kódových slov

. není jednoznačný

Poznámka: U D-nárnho kódu pokračujeme na konci D-znaků. Po k-krocích máme

$|X| - kD$ .. znaků.

dbychom dostali D znaků musíme někdy na počátek přidat symboly "Dummy" s nulovými pravděpodobnostmi.

Úvahy o optimálním kódu

$p_1, p_2, \dots, p_n$ pravděpodobnosti $p_1 \geq p_2 \geq \dots \geq p_n$
$l_1, l_2, \dots, l_n$ příslušné délky kódových slov

Tvrzení (1) Každý optimální kód splňuje vlastnost monotonie:

$$p_j > p_k \Rightarrow l_j \leq l_k$$

12) Pro instantní optimální kód platí, že každý kód maximální délky má sousedy.

(3) Existuje optimální kód tak, že kódová slova odpovídající dvou nejmenším pravděpodobnostem se liší v posledním bitu.

Dk: (1) Kdyby ne tak prohozením příslušných kódových slov zkrátíme délku
Stačí pro dva

$$p_1 > p_2 \Rightarrow \underbrace{p_1 l_1 + p_2 l}_{\text{optimální}} \le \underbrace{p_1 l_2 + p_2 l_2}_{\text{prohozené délky}}$$

$$\Rightarrow (p_1 - p_2)(l_2 - l_1) \ge 0 \Rightarrow l_2 \ge l_1.$$

("swap")

(2) Ať existuje kód s maximální délkou bez sourozence

pak zkrácením o větev bez sourozence kód vylepším (opět bude instantní) – spor.

(3) Každý kód max. délky má sourozence. ∴ Prohodíme [illegible] dvojicí dvěma nejmenším pravděpodobnostem

Tato výměna neovlivní střední délku.
(příspěvek: l_{max} (součet pro posl. úrovně))

Def: Kanonický kód je optimální instantní kód splňující (1), (2), (3) výše.

Příklad: Které z následujících kódů jsou Huffmanovy kódy

a) {0, 10, 11} ✓ pro $\vec{p} = (1/2, 1/4, 1/4)$

b) {00, 01, 10, 110}
↑ není saturace, musí být zkrácen

c) {01, 10} - není optimální.

Důkaz k optimalitě Huffmanova kódu

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

$p_1 \geq p_2 \geq \dots \geq p_m$ $\quad L^*_m$ – optimální délka

$\vec{q} = (p_1, p_2, \dots, p_{m-1} + p_m)$ $\quad L^*_{m-1}$ – optimální délka

Expanze kódu: $C^*_{m-1}(\vec{q})$ – kanonický kód pro $\vec{q}$

Přirozeně nyní
kód pro $\vec{p}$ následujícím způsobem

$C^*_{m-1}(\vec{q})$	Kódy	délky		$C_m(\vec{p})$	
	↓				
p_1	w_1	l_1	→	p_1	w_1
p_2	w_2	l_2	→	p_2	w_2
			→		
p_{m-2}	w_{m-2}	l_{m-2}	→	p_{m-1}	$w_{m-1}0$
$p_{m-1}+p_m$	w_{m-1}	l_{m-1}		p_m	$w_{m-1}1$

Srovnáme si délky

(*) $L(C_m(\vec{p}')) = L^*_{m-1} + p_{m-1} + p_m$

Naopak, mějme kanonický kód pro $\vec{p}$. Sestrojme kód pro $\vec{q}$ tak, že o jeden bit zkrátíme dva slova s nejmenšími pravděpodobnostmi a přiřadíme ke pravděpodobnosti $p_m + p_{m-1}$ nový kód, $C_{m-1}(\vec{q})$ má délku ostatní

(**) $L(C_{m-1}(\vec{q}')) = L^*_m - p_{m-1} - p_m$

Spojíme (*) a (**):

$$L(C_m(\vec{p}')) + L(C_{m-1}(\vec{q}')) = L^*_{m-1} + L^*_m$$

$$\Rightarrow \underbrace{\left(L^*_{m-1} - L(C_{m-1}(\vec{q}'))\right)}_{\geq 0} + \underbrace{\left(L^*_m - L(C_m(\vec{p}'))\right)}_{\geq 0} = 0$$

$$\Rightarrow \left.\begin{array}{l} L(C_{m-1}(\vec{q}')) = L^*_{m-1} \\ L(C_m(\vec{p}')) = L^*_m \end{array}\right\} \Rightarrow \begin{array}{l} C_m(\vec{p}') \text{ a} \\ C_{m-1}(\vec{q}') \\ \text{jsou optimální!} \end{array}$$

Tedy v Huffmanově kódování se optimalita zachová v obou směnech. Tedy máme:

Věta: Huffmanův kód je optimální.

Dk: Předchozí úvahy + kód pro 2-prvkovou abecedu je optimální.

Příklad: Špatné víno

Máme 6 lahví vína, jedno je ztracené to s pravděp.

$$\left(\frac{8}{23}, \frac{6}{23}, \frac{4}{23}, \frac{2}{23}, \frac{2}{23}, \frac{1}{23}\right)$$

a) Složitost lahev po lahvi. Jaký je průměrný počet testování?

Bereme to od největší pro k nejmenší

$$\sum p_i l_i = 1 \cdot \frac{8}{23} + 2\frac{6}{23} + 3\frac{4}{23} + 4\frac{2}{23}$$

$$+ 5 \cdot \frac{2}{23} + 6 \cdot \frac{1}{23} = \frac{55}{23} = 2{,}39$$

b) Lepší strategie. Máme skutečné láhev do ní můžeme nidat vzorky

Udiláme Huffmannův kód: Dvě délky
l_c (2, 2, 2, 3, 4, 4)

$$\sum_{i=1} p_i l_i = 2{,}35 \text{ - optimální}$$

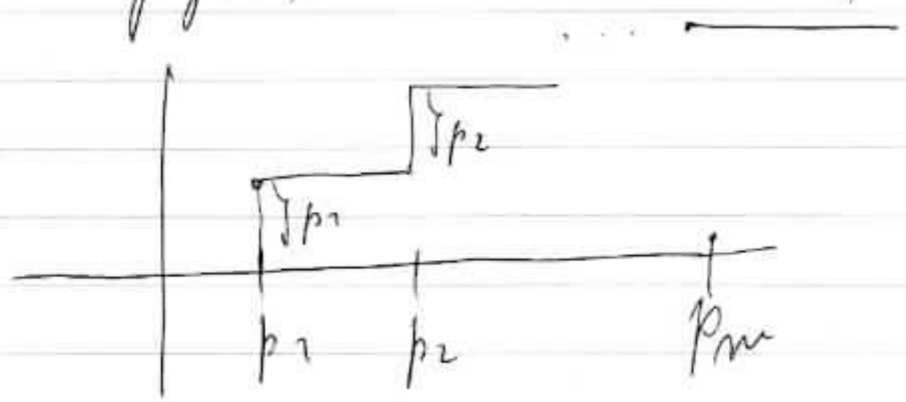
Shannon-Fano-Elias kód

$\mathcal{X} = \{1, 2, \dots, m\}$

$p(x) > 0 \quad \forall x$

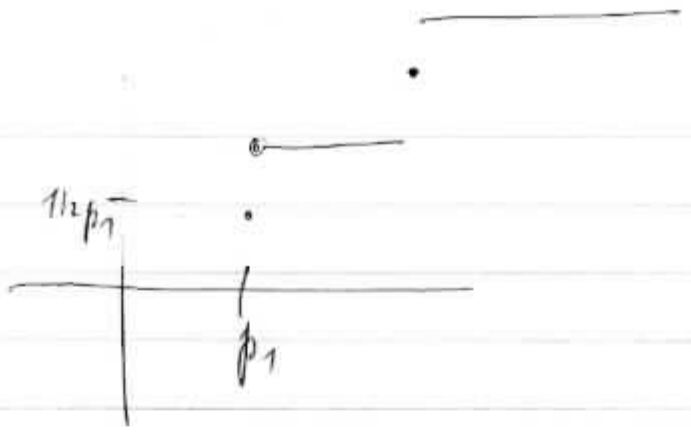
Distribuční funkce $F(x) = \sum_{a \leq x} p(a)$

(zprava spojitá)



Modifikace

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x)$$



$\bar{F}(x)$ rozvineme dyadicky a vezmeme za kódové slovo $l(x)$ bitů.

Značení $\lfloor \bar{F}(x) \rfloor_{l(x)}$

Kde $l(x) = \lceil \log 1/p(x) \rceil + 1$

Pak $\log \frac{1}{p(x)} + 1 \leq l(x) < \log \frac{1}{p(x)} + 2$

$$2 \cdot \frac{1}{p(x)} \leq 2^{l(x)} < 4 \cdot \frac{1}{p(x)}$$

$$\Rightarrow \quad 2^{-l(x)} \leq \frac{p(x)}{2}$$

Dále $\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_{l(x)} <$

$$< \frac{1}{2^{l(x)}} \leq \frac{p(x)}{2}$$

Tedy $\lfloor \bar{F}(x) \rfloor_{l(x)}$ zůstane uvnitř skoku odpovídajícímu x.

Tedy $l(x)$ bitů stačí k popisu x.
Navíc dostaneme instantní kód
neboť všechna zobrazování kódů
$z_1 \dots z_\ell$ jsou v intervalu

$$\left[0, z_1 \dots z_\ell ; 0 z_1 \dots z_\ell + \frac{1}{2^\ell}\right)$$

Tyto intervaly jsou disjunktní
neboť mají délku $\frac{1}{2^{l(x)}} < \frac{p(x)}{2}$ a $\left[\bar{F}(x)\right]_{l(x)} \le \bar{F}(x)$
Délka kódu: ale končí ve spodní půlce

$$L = \sum_{x \in X} p(x) l(x) = \sum p(x)\left(\left\lceil \log \frac{1}{p(x)} \right\rceil + 1\right)$$

$$< H(X) + 2$$

Příklad:

x	$p(x)$	$\bar{F}(x)$	binárně	$l(x)$	kód
1	0,25	0,125	0,001	3	001
2	0,5	0,5	0,10	2	10
3	0,125	0,8125	0,1101	4	1101
4	0,125	0,9375	0,1111	4	1111

0,25 | 0,8
0,25 | 0,5
0,125 = 2^{-3}

Střední délka je jediné z kritérií jakosti kódu. Ale důležitá je i relativní úspěšnost

---

Věta: Nechť $l(x)$ je kódová délka asociovaná s Shannonovým kódem. Nechť $l'(x)$ je kódová délka asociovaná s libovolným jiným jednoznačně dekódovatelným kódem.

Pak

$$P[l(X) \geq l'(X) + c] \leq \frac{1}{2^{c-1}}$$

---

např. pravděp., že $l(x)$ je horší o 5 je $\frac{1}{2^4} = \frac{1}{16}$.

horší o 3 s prav. $\leq 1/2$

hra se dvěma hráči a ten bude mít lepší kód.

Důkaz

$$P[l(X) \geq l'(X) + c] \leq$$

$$\leq P\left[\log \frac{1}{p(X)} \geq l'(X) + c - 1\right)$$

$$= P[p(X) \leq 2^{-l'(X) - c + 1}]$$

$$= \sum_{x \mid p(x) \leq 2^{-l'(x)-(c-1)}} p(x)$$

$$\leq \sum_{x \mid p(x) \leq 2^{-l'(x)-c+1}} 2^{-l'(x)-(c-1)}$$

$$\leq \sum_{x} 2^{-l'(x)-(c-1)} \leq 2^{-(c-1)}$$

Kraftova nerovnost.

Dalším rozborem lze ukázat, že Shannonův kód vykazuje všechna konkurenční.

Entropie stochastických procesů

Časová řada, diskrétní stochastický proces je posloupnost náhodných veličin

$X_1, X_2, \dots$

na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$.

Předpoklad: $X_1 : \Omega \to \mathcal{X}$ – konečná množina

Definice. Stochastický proces $(X_i)_{i=1}^{\infty}$ se nazývá stacionární, jestliže

$$P[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] =$$

$$P[X_{1+\ell} = x_1, X_{2+\ell} = x_2, \dots, X_{n+\ell} = x_n]$$

$\forall n, \ell = 1, 2, \dots$ a $x_1, \dots, x_n \in \mathcal{X}$

- Každý nezávislý proces je stacionární
- Závislost pouze na předchozí hodnotě – Markovův řetězec

Def. Časová řada $X_1, X_2, \dots$ se nazývá Markovův řetězec, jestliže

$$P[X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \dots X_1 = x_1)$$

$$= P[X_{n+1} = x_{n+1} \mid X_n = x_n] \quad \forall n, x_1 \to x_{n+1} \in \mathcal{X}$$

Příklad: $(X_n)_n$ – Markovův řetězec

$$P[X_3 = a_3, X_2 = a_2, X_1 = a_1] =$$

$$= P[X_3 = a_3 / X_2 = a_2, X_1 = a_1] \cdot P[X_2 = a_2, X_1 = a_1]$$

$$= P[X_3 = a_3 / X_2 = a_2] \cdot P[X_2 = a_2 / X_1 = a_1] \cdot P[X_1 = a_1]$$

Vedeme k řetězovému pravidlu

$$p(a_1, \dots, a_n) = p(a_1) \cdot p(a_2/a_1)\, p(a_3/a_2) \cdots p(a_n / a_{n-1})$$

Def Markovův řetězec je časově invariantní jestliže

$$P[X_{n+1} = a / X_n = b] = P[X_2 = a / X_1 = a]$$

$\forall n, \forall a, b \in \mathcal{X}$.

Věta: Všechny Markovské řetězce budou časově invariantní.

Přechodová matice

$X = \{1, 2, \dots, n\}$

$P = (p_{ij})$

$p_{ij} = P[X_{n+1} = j \mid X_n = i] = P[X_2 = j \mid X_1 = i]$

$\vec{p}_n$ - pravd. vektor pro X_n

$$\vec{p}_{n+1} = \vec{p}_n \cdot P$$

$$\Rightarrow \vec{p}_{n+1} = \vec{p}_1 \cdot P^n$$

Def: Stacionárním rozdělením $\vec{p}$ Markovského řetězce je pravděp. vektor

$$\vec{p} = \vec{p} \cdot P$$

---

x-li $\vec{p}_1 = \vec{p} \Rightarrow \vec{p}_n = \vec{p}$

Stacionární rozdělení $\vec{p}$ je řešením soustavy rovnic

$$(P^T - E)\vec{p} = 0.$$

P^T ... transp. matice

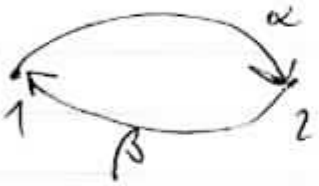
Nutnou podmínkou pro existenci netriviálního řešení je regularita $P^T - E$.

ani řádky nemusí být lineárně závislé - musí být součet $a_{i1} + \dots + a_{in} = 1$.

Příklad: Řešíme se dvěma stavy $X = \{1, 2\}$.

matice přechodu

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \quad 1 \geq \alpha, \beta \geq 0.$$



$$P^T - E = \begin{bmatrix} -\alpha & \beta \\ \alpha & -\beta \end{bmatrix}$$

tedy $-\alpha p_1 + \beta p_2 = 0$

$p_1 + p_2 = 1$

$$p_1 = \frac{\beta}{\alpha+\beta} \mid p_2 = \frac{\alpha}{\alpha+\beta}$$ stacionární stav.

Definice Entropie stochastického procesu

$\mathcal{X} = (X_1, X_2, \dots)$

je definována jako

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{H(X_1, X_2, \dots, X_n)}{n}$$

pokud tato limita existuje

Příklady: a) Posloupnost generuje náhodně znak z množiny n různých znaků se stejnou pravděpodobností nezávisle na sobě.

$$H(X_1, \dots, X_n) = \log n^n = n \cdot \log n$$

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{n \log n}{n} = \log n$$

---

b) zákonem i.i.d

$$H(\mathcal{X}) = \frac{1}{n} \lim_{n \to \infty} H(X_1, \dots, X_n) =$$

$$= \dots \frac{1}{n} n \cdot H(X_1) = H(X_1).$$