Kodai ir dekodavimo taisyklės

Kodai

$S = \{s_1, ..., s_m\}$ - šaltinio abėcėlė;
$A = \{a_1, ..., a_n\}$ - kodo abėcėlė.
S^k - iš k raidžių sudarytų žodžių aibė. Pastebėsime, kad $|S^k| = m^k$.
Visų žodžių aibė

$$S^* = \bigcup_k S^k .$$

- Kodavimo funkcija $\varphi : S^* \to A^*$.
  Kodas vienareikšmiškai dekoduojamas, jei φ - injekcija ir egzistuoja dekodavimo algoritmas.
- Jei $s_i \to w_i,\ w_i \in A^*$, tai kodu vadinsime kodo žodžių aibę $C = \{w_1, ..., w_m\}$. $s_i s_j \to w_i w_j$. Tokiems visada egzistuoja dekodavimo algoritmas. Paprasčiausi pvz.: skaitymas iš kairės arba dešinės. p- kodai.

Krafto nelygybė (1949)

Teorema. *Tegul* $A = \{a_1, ..., a_n\}$ *- kodo abėcėlė,* $l_1, ..., l_m$ *- natūralieji skaičiai. p - kodas*

$$C = \{w_1, ..., w_m\}, \quad w_i \in A^{l_i},$$

egzistuoja tada ir tik tada, kai

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

Krafto nelygybė (1949)

Tegul $A(w,l) = \{wu \mid u \in A^{l-|w|}\}$ ir $l_1 \le l_2 \le \ldots \le l_m$.

(1) $w_i \in A^{l_i},\;\; i = 1,2,\ldots,k < m$. Parenkame $w_{k+1} \in A^{l_{k+1}} \setminus \cup_{i=1}^{k} A(w_i, l_{k+1}) \neq \emptyset$, nes

$$\left|\bigcup_{i=1}^{k} A(w_i, l_{k+1})\right| = \sum_{i=1}^{k} n^{l_{k+1}-|w_i|} < n^{l_{k+1}} = |A^{l_{k+1}}|$$

(2) $w_i \in A^{l_i},\;\; i = 1,2,\ldots,m$ p-kodas. Tada

$$1 \le \left|A^{l_m} \setminus \bigcup_{i=1}^{m-1} A(w_i, l_m)\right| = n^{l_m} - \sum_{i=1}^{m-1} n^{l_m-|w_i|} .$$

Todėl

$$n^{-l_m} \le 1 - \sum_{i=1}^{m-1} n^{-l_i} .$$

□

McMillan'o nelygybė (1956)

Teorema. *Tegul* $A = \{a_1, ..., a_n\}$ *- kodo abėcėlė,* $l_1, ..., l_m$ *- natūralieji skaičiai. Jei kodas*

$$C = \{w_1, ..., w_m\}, \quad w_i \in A^{l_i},$$

yra vienareikšmiškai dekoduojamas, tai

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

McMillan'o nelygybė (1956)

Tegul $K = \sum_{i=1}^{m} n^{-l_i}$ ir $l_1 \leq l_2 \leq \ldots \leq l_m$.

$$K^N = \sum_{i_1=1}^{m} \ldots \sum_{i_N=1}^{m} n^{-(l_{i_1}+\ldots+l_{i_N})} = \sum_{r=N}^{N \cdot l_m} \frac{h(r)}{n^r}\,,$$

$$h(r) = \#\{(l_{i_1}, \ldots, l_{i_N}) \mid l_{i_1} + \ldots + l_{i_N} = r\} \leq |A^r| = n^r\,.$$

Todėl

$$K^N \leq \sum_{r=N}^{N \cdot l_m} 1 < N \cdot l_m\,, \quad \forall N \in \mathbb{N}\,.$$

Vadinasi $K \leq 1$.

$\square$

Optimalus kodas

$A = \{a_1, ..., a_n\}$ - kodo abėcėlė;
$S = \{s_1, ..., s_m\}$ - šaltinio abėcėlė;
$p_i = P(s_i), \ \ i = 1, ..., m$ - šaltinio tikimybės.

Apibrėžimas. *Kodo $c : S \to A^*$ vidutinis kodo žodžio ilgis yra*

$$L(c) = \sum_{i=1}^{m} |c(s_i)| p_i \, .$$

Apibrėžimas. *p-kodą $\bar{c} : S \to A^*$ vadinsime optimaliu, jei*

$$L(\bar{c}) = \min_c L(c) \, ;$$

čia minimumas imamas pagal visus p-kodus $c : S \to A^*$.

Optimalus kodas

$A = \{a_1, ..., a_n\}$ - kodo abėcėlė;
$S = \{s_1, ..., s_m\}$ - šaltinio abėcėlė;
$p_i = P(s_i),\ \ i = 1, ..., m$ - šaltinio tikimybės.

Apibrėžimas. *Kodo $c : S \to A^*$ vidutinis kodo žodžio ilgis yra*

$$L(c) = \sum_{i=1}^{m} |c(s_i)| p_i \,.$$

Apibrėžimas. *p-kodą $\bar{c} : S \to A^*$ vadinsime optimaliu, jei*

$$L(\bar{c}) = \min_c L(c)\,;$$

čia minimumas imamas pagal visus p-kodus $c : S \to A^*$.

Optimalus kodas

$A = \{a_1, ..., a_n\}$ - kodo abėcėlė;
$S = \{s_1, ..., s_m\}$ - šaltinio abėcėlė;
$p_i = P(s_i), \ \ i = 1, ..., m$ - šaltinio tikimybės.

Apibrėžimas. *Kodo $c : S \to A^*$ vidutinis kodo žodžio ilgis yra*

$$L(c) = \sum_{i=1}^{m} |c(s_i)| p_i \,.$$

Apibrėžimas. *p-kodą $\bar{c} : S \to A^*$ vadinsime optimaliu, jei*

$$L(\bar{c}) = \min_{c} L(c)\,;$$

čia minimumas imamas pagal visus p-kodus $c : S \to A^*$.

Hafmano kodai

$A = \{0, 1\}$ - kodo abėcėlė;
$S = \{s_1, ..., s_m\}$ - šaltinio abėcėlė;
$p_1 \geq p_2 \geq ... \geq p_m$ - šaltinio tikimybės.

1. Abėcėlių redukcija.

$$S \to S_1 = \{s_1, ..., s_{m-2}, \sigma\},$$

   čia $\sigma =< s_{m-1}, s_m >, \;\; P(\sigma) = p_{m-1} + p_m.$

2. Kodo konstrukcija. Tegul $S' \to S'', \;\; \sigma_1, \sigma_2 \in S'$, $\sigma =< \sigma_1, \sigma_2 >\in S''$. Jei $c''(\sigma) = w$, tai $c'(\sigma_1) = w0, \;\; c'(\sigma_2) = w1$.

Jei $A = \{a_1, ..., a_n\}$, tai pirmame žingsnyje jungiamų simbolių skaičius $2 \leq u \leq n$ turi tenkinti sąlygą

$$u \equiv m \mod (n-1)\,.$$

Hafmano kodų optimalumas

Lema. *Jei $\sigma_1, \sigma_2 \in S$ turi mažiausias nenulines tikimybes, tai $\exists$ toks optimalus binarinis kodas c, kad*

- $|c(\sigma_1)| = |c(\sigma_2)| = \max_{s \in S} |c(s)|$;
- *kodo žodžiai $c(\sigma_1)$ ir $c(\sigma_2)$ skiriasi tik paskutiniuoju simboliu.*

$\triangleright$

(1) Jei yra tik vienas max ilgio kodo žodis, jį sutrumpinę, gautume geresnį kodą.
(2) Jei bet kurie du ilgiausi žodžiai skiriasi ne tik paskutiniu simboliu, juos galėtume trumpinti. Taigi $\exists$ ilgiausi žodžiai $c(a_1) = w0,\ c(a_2) = w1$.
(3) Jei $\{a_1, a_2\} \neq \{\sigma_1, \sigma_2\}$, tai sukeitę jų kodo žodžius, gausime kitą optimalų kodą c', tenkinantį lemos sąlygas, nes

$$L(c') = L(c) - \sum_{i=1,2} (|c(a_i)| - |c(\sigma_i)|)(p(a_i) - p(\sigma_i)) \leq L(c).$$

$\square$

Hafmano kodų optimalumas

Teorema. *Hafmano kodai yra optimalūs.*

$\triangleright$

(1) Abėcėlių redukcija ir kodų konstravimas:
$S = S_1 \to S_2 \to ... \to S_{k-1} \to S_k \to ... \to S_{m-1}\,, \quad |S_{m-1}| = 2.$
$c_1 \leftarrow c_2 \leftarrow ... \leftarrow c_{k-1} \leftarrow c_k \leftarrow ... \leftarrow c_{m-1}.$

(2) Indukcija: $c_{m-1}, ..., c_{k+1},\ c_k$ - optimalūs. Įrodome, kad c_{k-1} optimalus.
Tarkime priešingai: jei $\sigma_1, \sigma_2 \in S_{k-1}$ ir $< \sigma_1, \sigma_2 >= \sigma \in S_k$, tai $\exists$ optimalus S_{k-1} kodas d_{k-1}, tenkinantis sąlygas:
$d_{k-1}(\sigma_1) = w0,\ d_{k-1}(\sigma_2) = w1,\ \ L(d_{k-1}) < L(c_{k-1})$.

(3) Sudarome naują p-kodą d_k abėcėlei S_k:
$d_k(s) = d_{k-1}(s)$, jei $s \in S_{k-1} \bigcap S_k$ ir $d_k(\sigma) = w$. Tada
$L(c_k) = L(c_{k-1}) - (p(\sigma_1) + p(\sigma_2)),$
$L(d_k) = L(d_{k-1}) - (p(\sigma_1) + p(\sigma_2)),\ \Rightarrow L(d_k) < L(c_k)$ (priešt.)

$\square$

Duomenų suspaudimo koeficientas

$A = \{a_1, ..., a_n\}$ - kodo abėcėlė.
$\mathcal{F}_1 \to \mathcal{F}_2$, čia $\mathcal{F}_1, \mathcal{F}_2 \in A^*$.
Jei, žinant $\mathcal{F}_2$, galima atstatyti $\mathcal{F}_1$, tai suspaudimas be nuostolių.

Suspaudimo koeficientas

$$\lambda = \frac{|\mathcal{F}_1|}{|\mathcal{F}_2|} \, .$$

Gali būti aktualu ne tik λ, bet ir greitis bei paslėptos sąnaudos.

$S = \{s_1, ..., s_m\} \subset A^*$ ir $W = \{w_1, ..., w_m\} \subset A^*$. Jei $\mathcal{F}_1 = s_{i_1}...s_{i_t}$ ir $s_i \to w_i$, tai $\mathcal{F}_2 = w_{i_1}...w_{i_t}$. Tada

$$\lambda = \frac{\sum_{i=1}^m m_i|s_i|}{\sum_{i=1}^m m_i|w_i|} = \frac{t \cdot L(S)}{t \cdot L(W)} = \frac{L(S)}{L(W)}.$$

Apibrėžimas. *S vadinamas vienareikšmiškai skaidančiu (kitaip: pilnuoju) kodu, jei* $\forall w \in A^*$ *vienareikšmiškai išreiškiamas*

$$w = s_{i_1}...s_{i_t}\nu\ ,$$

čia joks s_i *nėra* ν *priešdėlis ir* $|\nu| < \max|s_i|$.

Pavyzdys.
$S_1 = \{00, 11, 10\}, \qquad S_2 = \{0, 10, 110, 1110, 1111\}.$

Pilnieji kodai

Teorema. *Kodas $S = \{s_1, ..., s_m\} \subset A^*$ yra pilnas $\Leftrightarrow$ S yra p-kodas ir*

$$\sum_{j=1}^{m} n^{-|s_j|} = 1,$$

čia $n = |A|$.

Šenono kodas

$A = \{0, 1\}, S = \{s_1, ..., s_m\}, \quad p_i = P(s_i),$
$p_1 \geq p_2 \geq ... \geq p_m > 0$. Tegul

$$l_i = \left\lceil \log_2 \frac{1}{p_i} \right\rceil, \quad F_1 = 0, \quad F_k = \sum_{i=1}^{k-1} p_i, \;\; k \geq 2.$$

Pastebėsime, kad $l_i < -\log_2 p_i + 1$.
Šenono kodo

$$W_{Sh} = \{w_1, w_2, ..., w_m\}$$

žodis $w_i \in A^{l_i}$ sudarytas iš l_i pirmųjų bitų F_i dvejetainėje išraiškoje.

W_{Sh} yra p-kodas.

Šenono teorema

$A = \{0, 1\}, S = \{s_1, ..., s_m\}, \quad p_i = P(s_i),$
$H(S) = -\sum_{i=1}^{m} p_i \log_2 p_i.$

Teorema. *Šaltinio S Hafmano kodo W_H ir Šenono kodo W_{Sh} vidutiniai kodo žodžio ilgiai tenkina nelygybes*

$$H(S) \leq L(W_H) \leq L(W_{Sh}) < H(S) + 1\,.$$

Be to, $H(S) = L(W_H)$ tada ir tik tada, kai $p_i = 2^{-l_i}$.

Įrodymas.
$L(W_{Sh}) = \sum_{i=1}^{m} p_i l_i < \sum_{i=1}^{m} p_i(-\log_2 p_i + 1) = H(S) + 1.$

$$H(S) - L(W_H) = \sum_{i=1}^{m} p_i(-\log_2 p_i - l_i) \leq$$

$$\log_2 \mathrm{e} \sum_{i=1}^{m} p_i \left(\frac{2^{-l_i}}{p_i} - 1\right) = \log_2 \mathrm{e} \left(\sum_{i=1}^{m} 2^{-l_i} - 1\right) \leq 0.$$

Šenono rėžis

$$\lambda = \frac{L(S)}{L(W)} \le \frac{L(S)}{L(W_H)} \le \frac{L(S)}{H(S)}.$$

Galima nagrinėti šaltinį S^N. Tada

$$H(S^N) \le L(W_H(N)) < H(S^N) + 1\,.$$

Todėl

$$\frac{N \cdot L(S)}{H(S^N) + 1} < \lambda = \frac{N \cdot L(S)}{L(W_H(N))} \le \frac{N \cdot L(S)}{H(S^N)}.$$

Kodo abėcėlės tikimybės

$A = \{a_1, ..., a_n\}$ - kodo abėcėlė;
$S = \{s_1, ..., s_m\}$ - šaltinio abėcėlė;
$\alpha_j = P(s_j), \;\; j = 1, ..., m$ - šaltinio tikimybės.

$S \xrightarrow{c} A^* - - -$ $\boxed{\text{Kanalas su matrica } Q = (q_{ik})_{n\times r}}$ $- - - - - >$

Žinodami α_j ir $w_j = c(s_j) \in A^*$, rasime $p_i = P(a_i), \;\; i = 1, ..., n$.

Teorema. *Jei a_i įeina į w_j lygiai u_{ij} kartų, tai*

$$p_i = \frac{1}{L(c)} \sum_{j=1}^{m} u_{ij}\alpha_j, \;\; \text{čia} \;\; L(c) = \sum_{j=1}^{m} |w_j|\alpha_j.$$

Kodo abėcėlės tikimybės

Įrodymas. Teksto iš N simbolių kode w_j sutinkamas m_j kartų. Aišku, kad $m_1 + m_2 + \ldots + m_m = N$. Tada

$$p_i(N) = \frac{\sum_{j=1}^m u_{ij} m_j}{\sum_{j=1}^m |w_j| m_j} = \frac{\sum_{j=1}^m u_{ij} \frac{m_j}{N}}{\sum_{j=1}^m |w_j| \frac{m_j}{N}}.$$

Todėl

$$p_i = \lim_{N\to\infty} p_i(N) = \frac{\sum_{j=1}^m u_{ij} \alpha_j}{\sum_{j=1}^m |w_j| \alpha_j},$$

nes pagal DSD

$$\lim_{N\to\infty} \frac{m_j}{N} = \alpha_j.$$

$\square$

Dekodavimo taisyklės

$A = \{a_1, ..., a_n\}$ - kodo abėcėlė;
$B = \{b_1, ..., b_r\}$ - gavėjo abėcėlė;
Kanalo matrica

$$Q = (q_{ik})_{n \times r}, \;\; q_{ik} = P(b_k | a_i).$$

Tolygus(blokinis) kodas

$$\{w_1, ..., w_m\}, \;\; w_j \in A^l, \; m \leq n^l.$$

Kodo žodis w_j siunčiamas su tikimybe $\alpha_j = P(w_j)$.

Gautasis žodis $w \in B^l$. Kaip sužinoti kuris iš kodo žodžių w_j buvo pasiųstas?

Idealaus stebėtojo taisyklė

Gautasis žodis $w \in B^l$ dekoduojamas kodo žodžiu w_j, kuriam tikimybė

$$P(w_j|w) = \frac{P(w_j, w)}{P(w)}$$

yra didžiausia. Jei

$$w_j = a_{j_1} \ldots a_{j_l}\ , \quad w = b_{t_1} \ldots b_{t_l}\ ,$$

tai reikės rasti j su didžiausia tikimybe

$$P(w_j, w) = P(w_j)P(w|w_j) = \alpha_j q_{j_1 t_1} \ldots q_{j_l t_l}\ .$$

Didžiausio tikėtinumo taisyklė

Gautasis žodis $w \in B^l$ dekoduojamas kodo žodžiu w_j, kurį pasiuntus, yra didžiausia tikimybė gauti w.
Jei

$$w_j = a_{j_1} \dots a_{j_l}\,, \quad w = b_{t_1} \dots b_{t_l}\,,$$

tai reikės rasti j su didžiausia tikimybe

$$P(w|w_j) = q_{j_1 t_1} \dots q_{j_l t_l}\,.$$

Pastebėsime, kad šiuo atveju nereikia žinoti šaltinio tikimybių α_j.

Mažiausio atstumo taisyklė

Tegul $A \subset B$. Jei

$$w_j = a_{j_1} \dots a_{j_l}\,, \quad w = b_{t_1} \dots b_{t_l}\,,$$

tai Hamingo atstumas tarp w_j ir w yra

$$d_H(w_j, w) = \sum_{\substack{i=1 \\ a_{j_i} \neq b_{t_i}}}^{l} 1\,.$$

Gautasis žodis $w \in B^l$ dekoduojamas kodo žodžiu w_j, kuriam atstumas $d_H(w_j, w)$ yra mažiausias.
Šiuo atveju reikia žinoti tik kodo žodžius.

Dekodavimo taisyklių ekvivalentumas

Teorema.

1) *Jei* $\alpha_1 = \ldots = \alpha_m = 1/m$, *tai IST* $\sim$ *DTT.*
2) *Jei* $A = B,\ \ q_{ii} = q > 1/n$ *ir* $q_{ik} = q'$, $i \neq k$, *tai DTT* $\sim$ *MAT.*

Įrodymas.
1) $P(w_j, w) = P(w_j)P(w|w_j) = \frac{1}{m}P(w|w_j)$.
2)

$$P(w|w_j) = q^{l-d_H(w_j,w)}(q')^{d_H(w_j,w)} = q^l \left(\frac{q'}{q}\right)^{d_H(w_j,w)} .$$

Bet

$$\frac{q'}{q} = \frac{1-q}{q(n-1)} < 1$$

$\square$

Dekodavimo taisyklių patikimumas

Dekodavimo taisyklės patikimumas R nusako teisingo dekodavimo tikimybę. Tegul dekodavimo taisyklė nusakoma aibėmis $W_j = \{w \in B^l \mid w$ dekoduojamas kodo žodžiu $w_j\}$, $j = 1, \ldots, m$.
Tada

$$\begin{aligned} R &= \sum_{j=1}^{m} P(w_j, w \in W_j) = \sum_{j=1}^{m} P(w_j) P(w \in W_j | w_j) \\ &= \sum_{j=1}^{m} \alpha_j \sum_{w \in W_j} P(w|w_j) = \sum_{j=1}^{m} \alpha_j \sum_{w \in W_j} q_{j_1 t_1} \ldots q_{j_l t_l} . \end{aligned}$$

Klaidos tikimybė

$$E = 1 - R .$$

Maksimali klaidos tikimybė

$$\hat{E} = \max_{1 \le j \le m} P(w \notin W_j | w_j) = 1 - \min_{1 \le j \le m} P(w \in W_j | w_j) .$$