Információs rendszerek elméleti alapjai

Információelmélet

A forrás kódolása csatorna jelekké

Forrás : $\xi_1,\ldots,\xi_N$ változó sorozat

Kódolás eredménye :

$\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}\cdots$kódsorozat

Ha a kódolás $f\left(x_{i_1},\ldots,x_{i_N}\right)=\left(y_{i_1},\ldots,y_{i_{M\left(x_{i_1},\ldots,x_{i_N}\right)}}\right)$függvény,

akkor

$$H\left(\xi_1,\ldots,\xi_N\right)\geq H\left(f\left(\xi_1,\ldots,\xi_N\right)\right)=H\left(\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}\right)$$

Ha nincs veszteség $\rightarrow$ minden visszaállítható, azaz

$\exists f^{-1}$ inverz,

$$H\left(\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}\right)\geq H\left(f^{-1}\left(\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}\right)\right)=H\left(\xi_1,\ldots,\xi_N\right)$$

A forrás kódolása csatorna jelekké

Ebből veszteségmentes esetben

$H(\xi_1,\ldots,\xi_N) = H\left(\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}\right)$

Ha minden T ideig megy, akkor a csatorna kimenetén

$N(T) = 2^{T(C\pm\delta)}$ jelsorozat jelenik meg. Ezen a jelsorozaton az egyenletes eloszlás entrópiája $T(C\pm\delta)$

$C = \lim_{T\to+\infty}\left(\frac{1}{T}\log_2 N(T)\right) \Rightarrow$ egyenletes eloszlás esetén

a maximális : $\log_2 N(T) = T(C\pm\delta), \delta > 0, \delta \to 0$

Az automata állapota ilyenkor belső nem ismert érték :

$f\left(x_{i_1},\ldots,x_{i_N}\right) = \left(y_{i_1},\ldots,y_{i_{M(x_{i_1},\ldots,x_{i_N})}}, \delta\left(x_{i_1},\ldots,x_{i_N}\right)\right)$

$H(\xi_1,\ldots,\xi_N) \geq H(f(\xi_1,\ldots,\xi_N)) = H\left(\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}, \delta(\xi_1,\ldots,\xi_N)\right) \geq$

$\geq H\left(\eta_1,\eta_2,\ldots,\eta_{M(\xi_1,\ldots,\xi_N)}\right)$

A csatorna kimenetén ebben az esetben is TC entrópia a maximális.

Ha véges detzerminisztikus automatával (VDA)kódolunk, akkor a kódoló automatának van még egy plussz állapota, δ

A csatorna kimenetén ez nem észlelhető, ezért az entrópia kisebb.

T ideig működő forrás entrópiája nem haladhatja meg a csatorna maximális entrópiáját.

A zajmentes csatorna alaptételének bizonyítása

Az alaptétel bizonyítása:

Forrás: H bit/szimbólum entrópia, V szimbólum/sec sebesség

Csatorna: C bit/sec

Tétel: a) Nem lehet a forrást C/H-nál nagyobb sebességgel működtetni, hogy a csatornán minden veszteségmentesen átvihető legyen.

b) Tetszőleges $\delta>0$-hoz létezik kódolás, hogy C/H-δ működtetve a forrást minden veszteségmentesen átvihető legyen.

A zajmentes csatorna alaptételének bizonyítása

Legyen a sebesség V

T idő alatt a forrás entrópia :

$\left[TV\left(H\text{-}\rho\right), TV\left(H+\rho\right)\right]$, közé esik.

A csatorna jelek száma : $2^{T(C-\gamma)} < N(T) < 2^{T(C+\gamma)}$

A csatorna kimenetén az entrópia legfeljebb $T(C+\gamma)$ lehet.

Ha $\langle TV\left(H\text{-}\rho\right) > T(C+\gamma) \Rightarrow$ biztos veszteséges$\rangle$

Tehát $TV\left(H\text{-}\rho\right) \le T(C+\gamma)$, amiből

$$V \le \frac{C+\gamma}{H\text{-}\rho} \to \frac{C}{H}, \gamma \to 0, \delta \to 0$$

A zajmentes csatorna alaptételének bizonyítása

b) A N hosszú üzenethez mennyi az átlagos átviteli idő

Kétféle hosszúságú kód lesz: tipikus sorozatokra rövidebb, a lényegtelenre hosszabb (Független, azonos eloszlás!)

<A kis görög betűk tetszőlegesen kicsi mennyiségeket jelentenek, N, T elég nagy számokat>

A zajmentes csatorna alaptételének bizonyítása

A N hosszú sorozatok jellemzése szerint

(i) A tipikus halmaz $1-\varepsilon$ -nál nagyobb valószínűségű elemeinek a száma $\leq 2^{N(H+\rho)}$

(ii) Lényegtelen halmaz : legfeljebb ε valószínűségű, elemeinek a száma $\leq n^N = 2^{N \log_2 n}$

Kódhossz időben

(i)-hez : T_1 idő (kódhossz) a csatornán kódolással az átvihető jelek száma

$$2^{T_1(C-\gamma)} > 2^{N(H+\rho)}$$

$$T_1 \geq N \frac{H+\rho}{C-\gamma} = N \cdot \left(\frac{H}{C} + \alpha \right), \alpha \to 0,$$

A zajmentes csatorna alaptételének bizonyítása

(ii) - höz, egy speciális T_1 hosszú kód, majd T_2 idő

(hosszú jellel kódoljuk a nem tipikus üzeneteket)

$2^{T_2(C-\gamma)} > 2^{N \log_2 n}$, amiből

$$T_2 > N\left(\frac{\log_2 n}{C-\gamma}\right) = N\left(\frac{\log_2 n}{C} + \gamma_2\right)$$

Az átviteli idő várható értéke

$$(1-\varepsilon)T_1 + \varepsilon(T_1 + T_2) =$$

$$= N\left(\frac{H}{C} + \gamma_1\right) + \varepsilon N\left(\frac{\log_2 n}{C} + \gamma_2\right) \quad / \quad : N$$

egy szimbólumra jutó idő:

$$\left(\frac{H}{C} + \gamma_1\right) + \varepsilon\left(\frac{\log_2 n}{C} + \gamma_2\right)$$

A másodpercenkénti szimbólumok száma, ennek a reciproka

$$V = \left(\frac{H}{C} + \gamma_1\right)^{-1} = \frac{C}{H} - \delta. \text{ Q.E.D.}$$

Nevezetes kódolások

Kódolási feladat:

Adott a $P(\xi = x_i) = p_i, \quad i = 1,\ldots,n$ eloszlás.

Kódoljuk az x_i - t véges bináris szavakkal $\langle (0,1) abc \rangle$

x_i kódja ω_i

Invertálható kódolás

$x_{e_1},\ldots,x_{i_n}$ az $\omega_{i_1},\ldots,\omega_{i_n}$ kódból visszaállítható legyen.

Elégséges feltétel: prefix - mentes kódolás (kódfa!)

ω_i nem kezdőszelete (prefixe) ω_j - nek, ha $i \neq j, \quad l(\omega_i)$ az ω_i hossza

A kódolás jóságát a kódhossz várható értékével jellemezzük.

$$\sum_{i=1}^{n} l(\omega_i) p_i .$$

A csatorna - alaptétel következménye, invertálható kódolásra

$$H(p_1,\ldots,p_n) \leq \sum_{i=1}^{n} l(\omega_i) p_i .$$

Nevezetes kódolások

Hibajavító kódok – Hamming

Titkosítás RSA, elliptikus

Huffman kód –tömörítési jelleg

Előre meg kell határozni, hogy mit mivel kódolunk,

A jövőben véletlen eloszlás szerint érkeznek a szimbólumok

Nevezetes kódolások

$x_1, \ldots, x_n$, – lehetséges szimbólumok

Bináris kód : $x_i \to \omega_i$ bináris kód.

$l(\omega_i)$ - kód hossza.

Feladat visszafejthetőség

$x_1, \ldots, x_n \to \underbrace{\omega_1, \ldots, \omega_n}$

ebből az eredeti sorozat visszaállítható legyen

Prefix – mentes kód garantálja a visszafejthetőséget

$\forall i, j, \quad \omega_i$ nem kezdőszelete ω_j - nek.

$\sum_{i=1}^{n} 2^{-l(\omega_i)} = 1$ véges teljes kódrendszer

$\sum_{i=1}^{\infty} 2^{-l(\omega_i)} \le 1$ (Kraft egyenlőtlenség : végtelen kódrendszer)

szimbólumok eloszlása

$x_1, \ldots, x_n$ - $p_1, \ldots, p_n$ valószínűségekkel

Kódhossz várható értéke : $\sum_{i=1} p_i l(\omega_i)$

Nevezetes kódolások

Tétel: $H(p_1,\ldots,p_n) \le \sum_{i=1}^{n} l(\omega_i)p_i.$

$q_i = 2^{-l(\omega_i)}$ eloszlásra a logaritmikus szummációs lemma miatt

$$\sum_{i=1}^{n} a_i \log_2 \frac{b_i}{a_i} \le a \log_2 \frac{b}{a}$$

$$\sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} q_i = 1 \; \forall i - re, p_i, q_i \ge 0$$

$$\sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \le 1 \cdot \log_2 \frac{1}{1} = 0; \underbrace{-\left(\sum_{i=1}^{n} p_i \log_2 p_i\right)}_{H(p_1,\ldots,p_n)} - \left(-\sum_{i=1}^{n} p_i \log_2 q_i\right) \le 0$$

$$H(p_1,\ldots,p_n) \le -\sum_{i=1}^{n} p_i \log_2 q_i$$

Állítás: ∃ prefix - mentes kódolás, amin

$$\sum_{i=1}^{n} l(\omega_i)p_i. < H(p_1,\ldots,p_n) + 1;$$

1 bit veszteséggel tudjuk megközelíteni az entrópiát

Nevezetes kódolások

Nevezetes kódok :

Shannon - Fano kód :

Feltesszük : $p_1 \geq p_2 \geq \ldots \geq p_n$.

Legyen $Q_1 = 0, \quad Q_i = \sum_{j=1}^{i-1} p_j, \quad i = 2, \ldots, n$

ω_i megadása : fejtsük ki binárisan a $Q_i - t$ olyan l_i hosszban, amire

$2^{-l_i} \leq p_i < 2^{-l_i+1}$ legyen

Ebből $l_i \geq -log_2\, p_i > l_i - 1$

Nevezetes kódolások

A Kódhossz várható értéke :

$$\sum_{i=1}^{n} p_i l_i < \sum_{i=1}^{n} p_i\left(-log_2\ p_i + 1\right) = \underbrace{\sum_{i=1}^{n} - p_i\ log_2\ p_i}_{H(p_1,\ldots,p_n)} + \underbrace{\sum_{i=1}^{n} p_i}_{1} = H\left(p_1,\ldots,p_n\right) + 1$$

Bizonyítandó a prefix mentesség :

A kódszavak mint bináris törtek foghatók fel

- monoton növekvő értékek
- hosszuk is monoton növekszik

- Egy kódszó csak a nagyobb indexű prefixe lehet, de

akkor már a rákövetkezőnek is prefixe,

hiszen az kevesebbel nőtt értékben, mint a többi rákövetkező

Nevezetes kódolások

Elég ω_i és ω_{i+1} - re bizonyítani, hogy ω_i nem prefixe ω_{i+1} - nek.

ω_i hossza l_i és olyan, hogy

p_i első értékes jegye vagyis 1 bitje éppen az $l_i - dik$

Tehát $Q_i + p_i = Q_{i+1}$

ω_i – hez utolsó értékes jegyénél nagyobb értéket adunk

ezért az első l_i bitben is lesz változás.

$(\omega_i$ véges $(l_i$ hosszú$)$ diadikus tört, a végéről a ∞ sok 0 - t elhagyjuk$)$

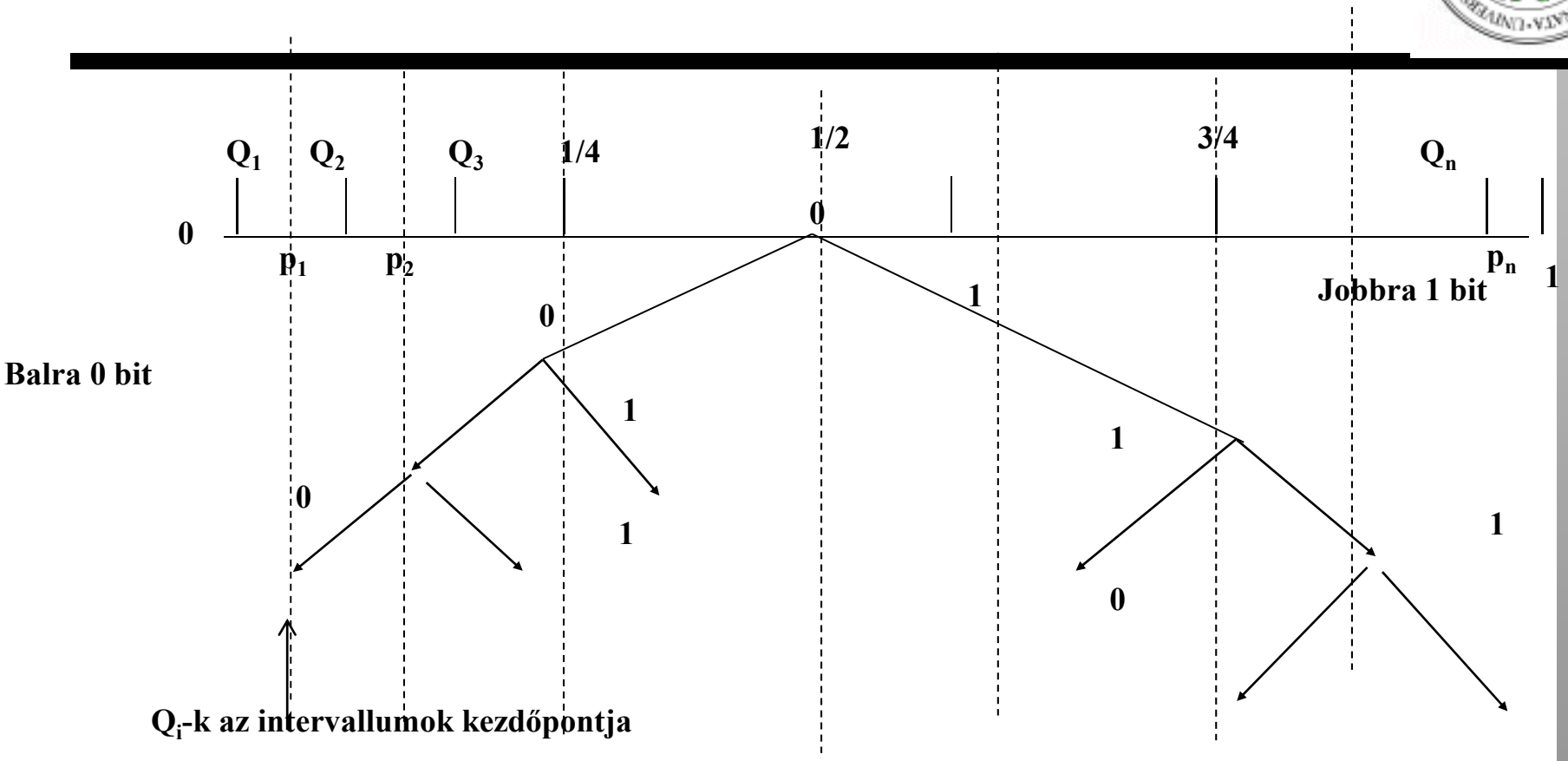
Nevezetes kódolások

Szimbólumok gyakoriság szerint csökkenő sorrendbe való rendezése

Lista két részre osztása úgy, hogy a két részben a szimbólumok gyakoriságának összege közel egyenlő legyen (0-1 kód hozzárendelése)

A felezés addig tart, amíg csak egy szimbólum marad az intervallumban

szemléltetés



Q_i-k az intervallumok kezdőpontja

Addig felezünk, amíg az intervallumban csak egy Q_i marad

Hátránya, rendezni kell az eloszlásokat

Kódfa építhető, (prefix kód) –gyakrabban szereplő elemek kódja rövidebb

Shannon kód példa

Example 2.11

symbol	probability	P_i		length l_i		code ($r = 2$)
u_1	1/4	P_1	$= 0$	l_1	$= 2$	$= 00$
u_2	1/4	P_2	$= 1/4$	l_2	$= 2$	$= 01$
u_3	1/8	P_3	$= 1/2$	l_3	$= 3$	$= 100$
u_4	1/8	P_4	$= 5/8$	l_4	$= 3$	$= 101$
u_5	1/16	P_5	$= 3/4$	l_5	$= 4$	$= 1100$
u_6	1/16	P_6	$= 13/16$	l_6	$= 4$	$= 1101$
u_7	1/32	P_7	$= 7/8$	l_7	$= 5$	$= 11100$
u_8	1/32	P_8	$= 29/32$	l_8	$= 5$	$= 11101$
u_9	1/32	P_9	$= 15/16$	l_9	$= 5$	$= 11110$
u_{10}	1/32	P_{10}	$= 31/32$	l_{10}	$= 5$	$= 11111$

△

Matematikai kitérő – A valószínűségszámítás alapfogalmairól

Diadikus törtek:

$\frac{k}{2^m}$ alakú racionális számok. $m > 0, k \in \mathbb{N}$

Ha $1 \le k < 2^m$, akkor $k = \sum_{j=0}^{m-1} a_j 2^j$

ahol $a_0, \ldots, a_{m-1} \in \{0,1\}$, tehát

$\frac{k}{2^m} = \sum_{j=0}^{m-1} a_j 2^{j-m} = \sum_{k=1}^{m} a_{m-k} 2^{-k} = (.a_{m-1} \ldots a_0)_2$, tehát

minden diadikus tört felírható 0 és 1 között felírható egy véges" bináris" reprezentá cióval, a végén végtelen sok zéróval kiegészítv e. Megfordítv a minden 0 és 1 közötti szám véges bináris ábrázolásss al diadikus tört.