CODIFICA DI SORGENTE

Sorgente S
(d.d.p. $(p_1, p_2, \dots, p_K)$)

→ Codice di Shannon-Fano (asintoticamente ottimo; non necessar. ottimo per n finito)
→ Codice di Fano (" " ; " " " " ")
→ Codice di Huffman (ottimo per qualunque n finito)

Codici basati sulla descrizione statistica della sorgente (la d.d.p. $(p_1, \dots, p_K)$)

$S \longrightarrow$ $x_{i_1}\, x_{i_2} \dots x_{i_n}$ . . . .
$(\cancel{p_1, \dots,} p_K)$ $\quad x$

non abbiamo accesso a p

Codifica universale: asintoticamente ottima ma indipendente dalla d.d.p. della sorgente

→ Codice multinomiale
→ Codice Ziv-Lempel (alla base del formato di compressione .zip)
($\underline{LZ77}$, LZ78)

<u>Avere</u> codici efficienti senza conoscere la d.d.p. della sorgente S

$S \rightarrow$ $\mathcal{A} = \{x_1, \dots, x_K\}$ (x_i)
$\quad\downarrow n_1 \quad \downarrow n_K$

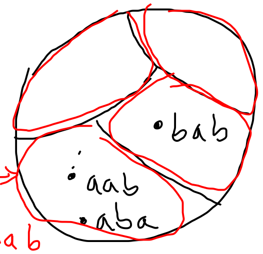
Una stringa $x = (x_{i_1}, x_{i_2}, \dots, x_{i_n})$ ha un tipo esatto

$T_x = (n_1, n_2, \dots, n_K)$ dove n_i è il numero di occorrenze del simbolo x_i nella stringa x

Es. $\mathcal{A} = \{a, b\}$, $x = aba \rightarrow T_x = (2, 1)$ $\quad x = aab \rightarrow T_x = (2,1)$

$(0,3) \to 0$ $\Gamma_n = 4$
$(3,0) \to 1$
$(2,1) \to 2$
$(1,2) \to 3$
$n=3$

A^n



$T_{aba} = T_{aab}$

A^n è partizionato in classi di equivalenza date dai tipi delle sequenze.

- Il numero di classi di equivalenza è indicato con Γ_n

Osservazione : $\Gamma_n \leq (n+1)^K$ ← numero di K-ple di interi con valore tra 0 e n

$(\cdot , \cdot , \cdot , \cdot , \cdot)$

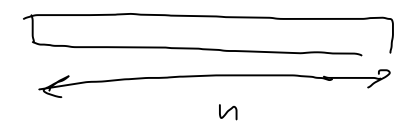
$T_x = (n_1, n_2, \dots, n_K)$

$0 \leq n_1 \leq n$ $\quad$ $n+1$ modi
$0 \leq n_i \leq n$

Inoltre sia T_x l'insieme delle sequenze in A^n che hanno lo stesso tipo di x

$$|T_x| = \binom{n}{n_1, n_2, \dots, n_K} = \frac{n!}{n_1! \, n_2! \dots n_K!}$$

(Coefficiente multinomiale)



$(\sum_i n_i = n)$

Codifica multinomiale : per codificare $x \in A^n$, usa la parola a lunghezza variabile

$$\varphi(x) = \varphi_P(T_x) * \varphi_S(x)$$

prefisso che dipende solo dal tipo di x

suffisso (dipende da x)

- $\varphi_p(T_x)$ è il numero d'ordine di T_x in un elenco lessicografico di tutti i tipi (identifica il tipo).
- $\varphi_s(x)$ è il numero d'ordine di x in un elenco lessicografico dell'insieme T_x di tutte le sequenze con lo stesso tipo di x (identifica x se già conosco T_x)

1	$(0,3)$
2	$(1,2)$
3	$(2,1)$
4	$(3,0)$

Parola di codice:

$$w = \underbrace{\varphi_p(T_x)}_{\text{lunghezza costante}} * \underbrace{\varphi_s(x)}_{\text{lungh. variabile}}$$

$$A^n \xrightarrow{\varphi} B^+$$

$$A^n \xrightarrow{\varphi_p} B^{\ell_p}$$

$|A| = K$ $\qquad$ $|B| = D$

$|B^{\ell_p}| = D^{\ell_p}$

Sia $\ell_p = |\varphi_p(T_x)|$ la lungh. del prefisso (costante)

$\ell_s(x) = |\varphi_s(x)|$ la lungh. del suffisso (variabile)

Devo avere $D^{\ell_p} \geq \Gamma_n$ $\quad \rightarrow \quad \ell_p = \lceil \log_D \Gamma_n \rceil \leq \lceil \log_D (n+1)^K \rceil \leq \lceil K \log_D (n+1) \rceil$

$$D^{l_s(x)} \geq |T_x| \quad \rightarrow \quad l_s(x) = \lceil \log_D |T_x| \rceil$$

Esempio. $\mathcal{A} = \{a, b\}$. $\mathcal{B} = \{0, 1\}$ $\quad D = 2 \quad K = 2 \quad n = 3$

$\Gamma_n = 4 \;:\; (0,3), (1,2), (2,1), (3,0) \;;\; l_p = \lceil \log_2 4 \rceil = 2$

	Tipo	n-pla	$\varphi_p(T_x)$	$l_s(x)$	$\varphi_s(x)$	$\varphi(x) = \varphi_p(T_x) * \varphi_s(x)$
0	$(0,3)$	bbb	00	$\lceil \log_2 1 \rceil = 0$	—	00
1	$(1,2)$	abb 0	01	$\lceil \log_2 3 \rceil = 2$	00	0100
		bab 1	01	" 2	01	0101
		bba 2	01	" 2	10	0110
2	$(2,1)$	baa	10	$\lceil \log_2 3 \rceil = 2$	00	1000
		aba	10	"	01	1001
		aab	10	"	10	1010
3	$(3,0)$	aaa	11	$\lceil \log_2 1 \rceil = 0$	—	11

aaa
aab
aba
baa
abb
bab
bba
bbb

Cosa succede per $n \to \infty$?

$$E[L^{(n)}] = \sum_{x \in A^n} p(x)\,|\varphi(x)| = \sum_{x \in A^n} p(x)\left(\ell_p + \ell_s(x)\right) =$$

lung. della parole di codice

$$= \ell_p \overbrace{\sum_{x \in A^n} p(x)}^{=1} + \sum_{x \in A^n} p(x)\,\ell_s(x)$$

non dipende da x

$$= \lceil \log_D \Gamma_n \rceil + \sum_{j=1}^{\Gamma_n} \underbrace{\sum_{x \in \tau_j} p(x)} \lceil \log_D |\tau_j| \rceil$$

Prob. complessiva delle sequenze di tipo τ_j

$$= \lceil \log_D \Gamma_n \rceil + \sum_{j=1}^{\Gamma_n} \lceil \log_D |\tau_j| \rceil \, Pr(\tau_j)$$

$(Pr[X \in \tau_j])$ $\shortparallel$ $Pr(\tau_j)$

$(\lceil x \rceil < 1 + x)$

$$< 1 + \log_D \Gamma_n + \sum_{j=1}^{\Gamma_n} (1 + \log_D |\tau_j|)\, Pr(\tau_j)$$

$$= 1 + \log_D \Gamma_n + 1 + \boxed{\sum_{j=1}^{\Gamma_n} Pr(\tau_j) \log |\tau_j|}$$

Consideriamo la v.a. Z (intera) che rappresenta il numero d'ordine del tipo τ_j nella lista dei tipi

Consideriamo la seguente entropia condizionata :

$$H(X|Z) = \sum_{j=1}^{\Gamma_n} \Pr(Z=j)\, H(X|Z=j) = \sum_{j=1}^{\Gamma_n} \overbrace{\Pr(\tau_j)}^{=\Pr(Z=j)}\, H(X|Z=j)$$

Tutte le sequenze di tipo τ_j sono equiprobabili, quindi :

$$H_D(X|Z=j) = \log_D |\tau_j| \qquad \Longrightarrow \qquad H(X|Z) = \boxed{\sum_{j=1}^{\Gamma_n} \Pr(\tau_j) \log_D |\tau_j|}$$

Riprendendo i conti, $E[L^{(n)}] < 2 + \log_D \Gamma_n + H(X|Z)$

$$\leq 2 + \log_D \Gamma_n + H(X)$$

$$\leq 2 + K \log_D (n+1) + H(X) .$$

Quindi il tasso della codifica è :

$$R = \frac{E[L^{(n)}]}{n} \leq \frac{2}{n} + K \frac{\log_D (n+1)}{n} + \frac{\overbrace{H(X)}^{= n H(S)}}{n} = \underbrace{\frac{2}{n}}_{\downarrow\, 0} + \underbrace{K \frac{\log_D (n+1)}{n}}_{\downarrow\, 0} + \underbrace{H(S)}_{\downarrow\, H(S)}$$

Per $n \to \infty$

$\Longrightarrow$ La codifica è asintoticamente ottima.

Codifica Ziv-Lempel: codifica "strutturale"; cerca di identificare ripetizioni