Elementos de Teoría de la Información y Codificación

Ultima actualización: 30/08/2002

INT. TEORIA DE LA INFORMACION

$I_{(E)} \cong f\left(\frac{1}{P_E}\right)$ Si $P_A > P_B$ $\Rightarrow$ $I_A < I_B$

Dados dos sucesos A y B simultáneos e independientes:

$I = I_A + I_B$ $\quad I = f\left(\frac{1}{P_A}\right) + f\left(\frac{1}{P_B}\right)$

$P_{A,B} = P_A \cdot P_B$

$\therefore \quad \boxed{I = \log \frac{1}{P}}$

$I = \log_2 \frac{1}{P}$ [bit]

FUENTE de Información de MEMORIA NULA

$S \rightarrow \begin{pmatrix} s_1 & s_2 & \dots & s_i & \dots & s_q \\ P_1 & P_2 & \dots & P_i & \dots & P_q \end{pmatrix}$

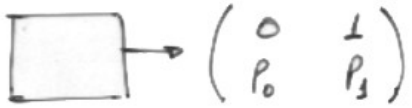
MEM. NULA $\Leftrightarrow$ S_i ESTADIST. INDEPENDIENTES.

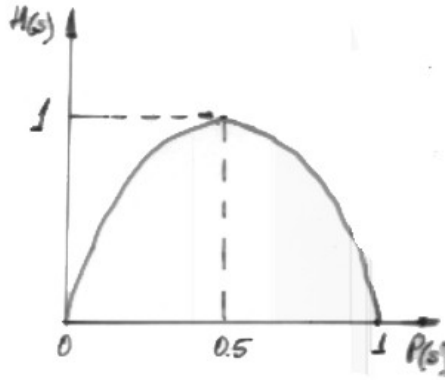
ENTROPIA $\quad H(S) \triangleq \sum_S P(S_i)\, I(S_i) = -\sum P(S_i) \log P(S_i)$

Cantidad media de información por símbolo

Ex $\quad S \rightarrow \begin{pmatrix} S_1 & S_2 & S_3 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}$ $\quad H(S) = \frac{1}{2}\log 2 + 2\left(\frac{1}{4}\log 4\right)$
$= 1.6$ bits/SIMB.

Fuente binaria





$$H(s) = \sum_{i=1}^{q} P_i \log \frac{1}{P_i}$$

$$H(s) \leq \log_2 q$$

$$H(s) = \log_2 q \iff P_i = \frac{1}{q} \ \forall i$$

$H(s)$ MAX si todos los símbolos son equiprobables

EXTENSIÓN DE UNA FUENTE DE MEMORIA NULA

Toda una fuente binaria podría agruparse de a dos bits, tres, etc. $\Rightarrow$ un nuevo alfabeto.

$$\{s_1\ s_2 \dots s_i \dots s_q\} \longrightarrow \{\sigma_1, \sigma_2, \sigma_3 \dots \sigma_i \dots \sigma_{q^n}\}$$

La nueva secuencia:

$$\sigma_i = \{s_{i_1}, s_{i_2} \dots s_{i_n}\} \qquad P(\sigma_i) = \prod_{\langle n \rangle} P_i$$

La extensión de orden "n", S^n

$$\boxed{H(S^n) = n\, H(S)}$$

Ej.

$H(S) = 1.5 \frac{\text{bits}}{\text{SIMB.}}$

$S \rightarrow s = \begin{pmatrix} s_1 & s_2 & s_3 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}$ Extensión de 2º orden

$\Longrightarrow \quad \Sigma \rightarrow \sigma = \begin{pmatrix} s_1 s_1 & s_1 s_2 & \cdots & s_3 s_3 \\ P_1 P_1 & P_1 P_2 & \cdots & P_3 P_3 \end{pmatrix}$

$H(S^2) = H(\Sigma) = 2\,H(S) = 3 \frac{\text{bits}}{\text{SIMB.}}$

$H(S^2) = \sum_{\langle S^2 \rangle} P(\sigma_i) \log \frac{1}{P(\sigma_i)}$

Fuente de Información de Markov

Si depende de "m" símbolos precedentes.

"Fuente de Markov de orden m".

$P(s_i / s_{j1}, s_{j2} \cdots s_{jm}) \qquad i = 1, 2 \cdots q$

$\exists$ "q" símbolos $\neq$

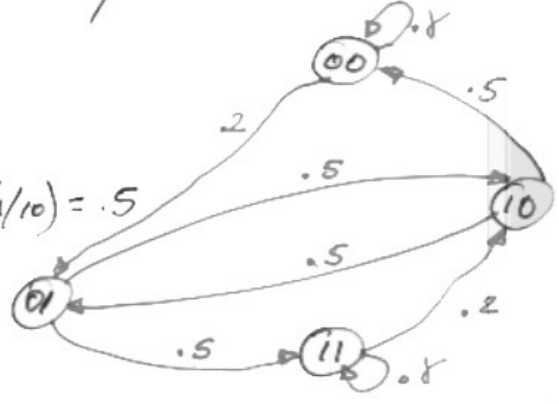
$\Rightarrow \exists$ "q^m" estados.

Ej. Una fuente de Markov 2º orden con $S = \{0, 1\}$ siendo las prob. condicionales

$P(0/00) = P(1/11) = .8$

$P(1/00) = P(0/11) = .2$

$P(0/01) = P(0/10) = P(1/01) = P(1/10) = .5$



– transiciones.

"ESTRUCTURA DEL LENGUAJE"

Tomando INGLES 26 LETRAS + 1 ESPACIO

- Aproximación cero:

  Fuente de memoria nula y todos los símb. equiprobables.

$$H(S) = \log 27 = 4.75 \frac{\text{bits}}{\text{símb.}}$$

- Primera Aproximación:

a) $P(A) = .0642$ PEZA 1961

$P(z) = 0.0005$

$P(\text{ESPACIO}) = 0.1859$

$$H(S) = -\sum_{27} P_i \log P_i = 4.03 \frac{\text{bits}}{\text{símb}}$$

b) Usando fuente de Markov de 1er orden con símb. de prob. condic. bien elegidos. (Prob 42)

$$H(S) = -\sum_{S^2} P(i,j) \log P(i,j) = 3.32 \frac{\text{bits}}{\text{símb.}}$$

- Segunda Aproximación

a) Shannon (51), observando sucesivos libros de TEXTO.

Orden n $H(S) = 3,1 \frac{\text{bits}}{\text{símb.}}$

Por otros métodos Shannon $0,6 < H(S) < 1,3$

Mejores aproximaciones hay que tomar palabras como símbolos.

Codigos

Bloque

Le le asigna secuencia fija.

No singulares

Todas sus palabras ≠

s_1	0
s_2	00
s_3	11
s_4	01

Pero si llega 0011 es singular.

Unívoco

Se dice unívocamente decodificable ⇔ su extensión de orden n es "no singular" para cualquier valor finito de n

Instantáneo

Puedo decodif. las palabras de la secuencia sin precisar los símbolos que la preceden.

Ej

	códigos		
	A	B	C
s_1	00	0	0
s_2	01	10	01
s_3	10	110	011
s_4	11	1110	0111

código coma

S → 10|110|0|110 (con C no distinguo si es 0 o 01

A y B son unívocamente.

Prefijos: la Ex 0111, los pref. son:

0, 01, 011, 0111

"Condición necesaria y suficiente para que un código sea unívocamente es que ninguna palabra del código coincida con el prefijo de otra

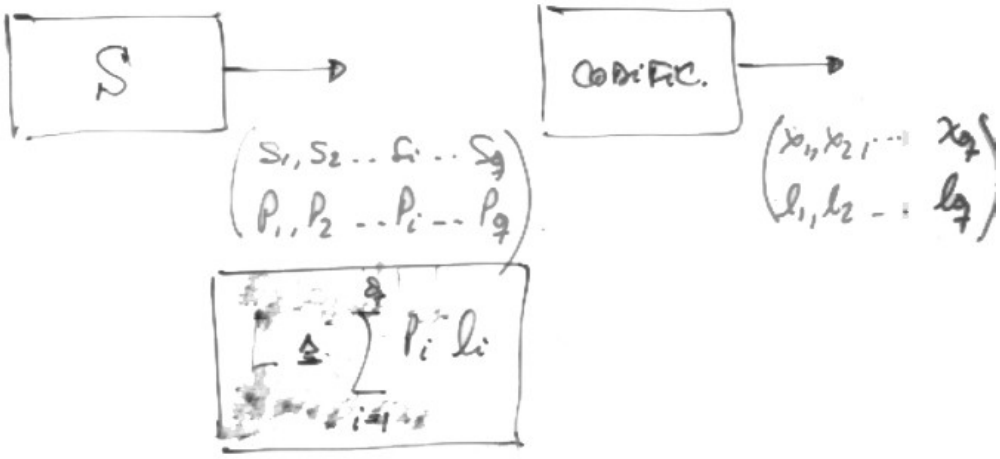
Inecuación de KRAFT

En un alfabeto binario

$$\sum_{i=1}^{4} 2^{-l_i} \leq 1$$ código inst.

Longitud Media de un código

Sea un código bloque



Def: Dado un código unívoco, este código será compacto (respecto a S) si su L es $\leq$ q' la L de todos los códigos unívocos q' puedan aplicarse a la misma fuente y el mismo alfabeto

En base a la inecuación de Kraft.

$$\boxed{H(s) \leq L}$$

Muestra la relación que $\exists$ entre la definición de información y una cantidad (L) que no depende de la definición.

NOTA:

L alcanza su valor mínimo $\Leftrightarrow$ puede elegirse la long de las palabras l_i, $l_i = \log_2 \frac{1}{P_i}$

$$P_i = \left(\frac{1}{2}\right)^{\alpha_i} \qquad \alpha_i: \text{entero.}$$

Basta tomar $l_i = \alpha_i$ para diseñar un código de L MÍN.

Ej:

$$S \rightarrow \begin{pmatrix} s_1 & s_2 & s_3 & s_4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$H(S) = -\sum_{i=1}^{4} P_i \log P_i = 2 \frac{\text{bits}}{\text{símb.}}$$

Es imposible codificar estos símbolos mediante un código binario unívoco de $L < 2 \frac{\text{bits}}{\text{símb}}$

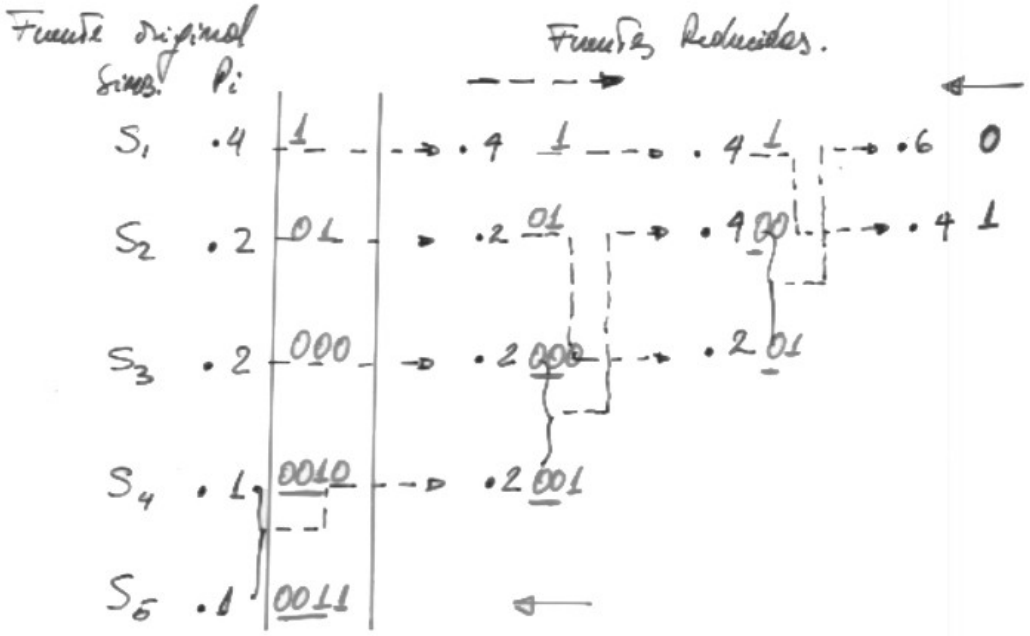
Un código compacto $P_i = \frac{1}{4}$ $P_i = \left(\frac{1}{2}\right)^2$

$l_i = \alpha_i = 2$

s_1 00
s_2 01
s_3 10
s_4 11

$$L = 2 \frac{\text{bits}}{\text{símb.}}$$

*



* Construcción de Códigos Compactos - Código Huffman.

TECNICAS DE COMPRESION

Se pueden hacer
- Nº finito del conjunto de símbolos
- frecuencia relativa
- Contexto donde aparecen.

CODIFICACIÓN INDEPENDIENTE de la FRECUENCIA

- HUFFMAN (52)

Pero no alcanza en muchos casos el límite teórico (H(S)) Prob.: algo de 2.5 bits con 3 bits

Huffman lo que hace c/símb. codifica en forma indep. y después lo transmite.

- CODIFICACION ARITMETICA

Von Prob. 2.5 bits con 3 bits

No codifica a los símbolos en forma independiente.

CODIFICACION dependiente CONTEXTO

RLL

Compresión datos { Esquemas basados en "diccionario"
Métodos estadísticos.

En sistemas pequeños basado en diccionarios es el más popular.
Pero la combinación de codific. aritmética y técnicas de "modelado" dan una buena performance.

Sistema basado en diccionario para reemplazar un grupo de símbolos del TEXT input con código de long FIXED. (El más conocido LZW: Lempel-Ziv-Welch)

- MODELADO BASADO CONTEXTO FINITO

Las probabilidades se calculan en función del "contexto" en que los símbolos aparecen.
Modelo orden 0 se calcula las prob. en forma "independiente" de cada símbolo previo.

En orden -0 una sola tabla de frecuencias.
orden -1 256 tablas ≠.
orden -2 65.536

Modelo Adaptativo

Usa solamente 5% pero si el contexto previo "g" puede llegar a 95%.
Como el modelo crece linealmente, pero la memoria necesaria exponencial. !!
Es por ello que la solución modelo "adaptativo"

Modelo Semi Adaptativo:

Usa un modelo ≠ para cada texto.
Se envía el decodificador el modelo.