3.5. CODIFICACIÓN Y ENTROPÍA

19/07/2021

Codificar un conjunto de caracteres que aparecen en un "texto" es asignar a cada uno de ellos un código, generalmente binario (0 y 1). El objetivo es minimizar el número de bits necesarios para codificar el texto.

"Texto" = X ; los símbolos de X se escriben $S = \{x_1, \dots, x_N\}$

Ejemplo 1. X = página de un libro, S = letras del alfabeto, incluidos los signos de puntuación (,), ., ;) y los espacios.

Ejemplo 2. X = fotografía digitalizada en (R, G, B)
$S = \{0, 1, \dots, 255\}$ en cada canal

Ejercicio 3.5.1. $S = \{A, O, E, S\}$, $X = AASAEEAO$

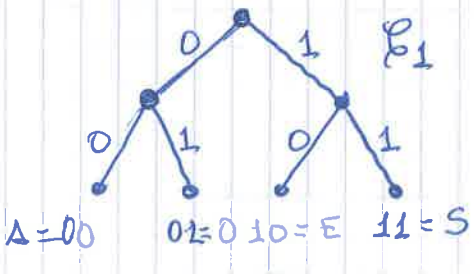
Ejemplo de Código

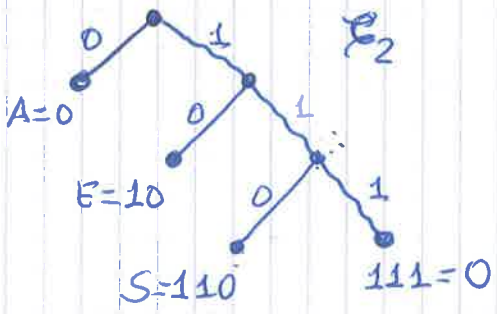
$\mathcal{C}_1 = \{A=00, O=01, E=10, S=11\}$

AASAEEAO → 0000110010100001 (16 bits)

$\mathcal{C}_2 = \{A=0, E=10, S=110, O=111\}$

AASAEEAO → 0011001010011 1 (14 bits)





ℓ_k = longitud del código binario que representa a x_k

$p_k = Pr(x_k \in X)$ la probabilidad de que x_k esté en X

La media de bits necesarios para codificar cada símbolo en un código $\mathcal{C}$ es $M_X(\mathcal{C}) = \sum_{k=1}^{N} \ell_k p_k$.

El objetivo es encontrar un código $\mathcal{C}$ para el texto X que haga mínima $M_X(\mathcal{C})$

Ejercicio 3.5.2 $X = AASAEEAO$, $S = \{A, O, E, S\}$

$Pr(A \in X) = \frac{4}{8} = \frac{1}{2}$; $Pr(O \in X) = \frac{1}{8}$, $Pr(E \in X) = \frac{1}{4}$, $Pr(S \in X) = \frac{1}{8}$

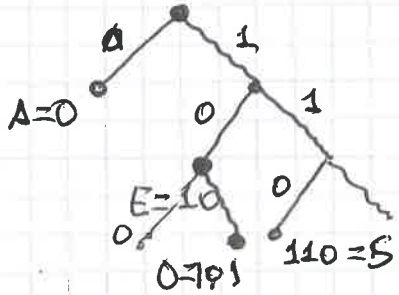
$M_X(\mathcal{C}_1) = 2\times\frac{1}{2} + 2\times\frac{1}{8} + 2\times\frac{1}{4} + 2\times\frac{1}{8} = 2$

$M_X(\mathcal{C}_2) = 1\times\frac{1}{2} + 3\times\frac{1}{8} + 2\times\frac{1}{4} + 3\times\frac{1}{8} = 1 + \frac{6}{8} = 1 + \frac{3}{4} = 1'75.$

Los códigos con palabras binarias de longitud variable, NO siempre son buenos para decodificar. Por ejemplo,

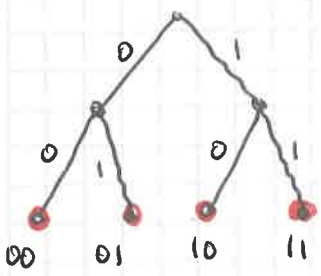
$$\mathcal{C}_3 = \{A=0,\ E=10,\ S=110,\ O=101\}$$

no es bueno porque 1010 puede ser EE ó OA

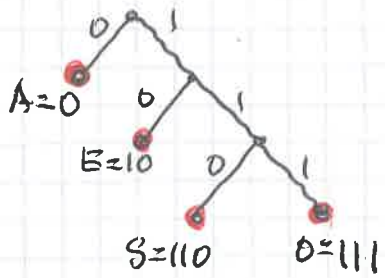


Def 3.5.1. Un código $\mathcal{C}$ cumple la condición del prefijo si no hay códigos de $\mathcal{C}$ que sean el comienzo de otro código de $\mathcal{C}$.

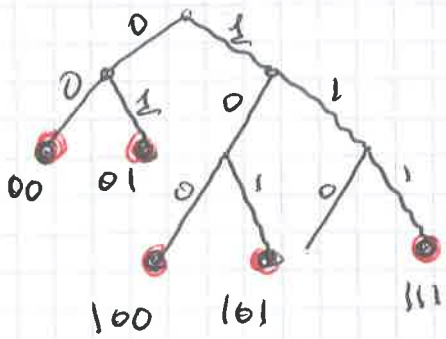
$\mathcal{C}_1$



$\mathcal{C}_2$



Otro ejemplo



Ejemplo de códigos que cumplen la condición del prefijo

3.5.1. LA ENTROPÍA DE SHANNON

Def 3.5.2. Dado un texto X en el que cada símbolo X_k tiene probabilidad $p_k = Pr(X_k \in X)$, la entropía de X es

$$\mathcal{E}(X) = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k} \geq 0$$

- Supongamos que $p_1 = Pr(X_1 \in X) = 1$ y $p_k = Pr(X_k \in X) = 0$ si $k = 2, \dots, N$. Entonces $\mathcal{E}(X) = 1 \cdot \log_2 \frac{1}{1} + \sum_{k=2}^{N} 0 \cdot \log_2 \frac{1}{0} = 1$ pq. $\lim_{x \to 0^+} x \log_2 \frac{1}{x} \overset{\text{L'Hôpital}}{=} 0$

- Supongamos que $p_k = Pr(X_k \in X) = \frac{1}{N}$, $k = 1, \dots, N$.
  Entonces $\mathcal{E}(X) = \sum_{k=1}^{N} \frac{1}{N} \log_2 \frac{1}{1/N} = (\log_2 N) \frac{1}{N} \sum_{k=1}^{N} 1 = \log_2 N$

Ejercicio 3.5.3 $X = \{AASAEEAO\}$, $\mathcal{S} = \{A, O, E, S\}$

Calcular $\mathcal{E}(X)$

S/ $Pr(A \in X) = \frac{1}{2}$, $Pr(O \in X) = \frac{1}{8}$, $Pr(E \in X) = \frac{1}{4}$, $Pr(S \in X) = \frac{1}{8}$

$$\mathcal{E}(X) = \frac{1}{2} \log_2 2 + \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 + \frac{1}{8} \log_2 8$$

$$= \frac{1}{2} 1 + \frac{1}{8} \cdot 3 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 = 1'75$$

Proposición 3.5.3 Para cualquier texto X escrito con N símbolos, $0 \leq \mathcal{E}(X) \leq \log_2 N$.

D/ Hallar el máximo de $\mathcal{E}(X) = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k} = -\sum_{k=1}^{N} p_k \log_2 p_k$ con las condiciones $0 \leq p_k \leq 1$ y $\sum_{k=1}^{N} p_k = 1$. Lagrange

$$F(p_1, \dots, p_N, \lambda) = -\sum_{k=1}^{N} p_k \log_2 p_k + \lambda \left(\sum_{k=1}^{N} p_k - 1 \right)$$

$$\left.\begin{aligned}\frac{\partial F}{\partial p_k} &= -\log_2 p_k - \frac{1}{\ln 2} + \lambda = 0 \\ k &= 1, 2, \dots, N \\ \frac{\partial F}{\partial \lambda} &= \sum_{k=1}^{N} p_k - 1 = 0\end{aligned}\right\} \Rightarrow \log_2 p_k = \lambda - \frac{1}{\ln 2}$$

$$p_k = 2^{\lambda} \cdot 2^{-\frac{1}{\ln 2}} = 2^{\lambda} 2^{-\log_2 e}$$

$$\Rightarrow p_k = \frac{2^{\lambda}}{e}$$

$$1 = \sum_{k=1}^{N} p_k = \sum_{k=1}^{N} \frac{2^{\lambda}}{e} = \frac{2^{\lambda}}{e} N \Rightarrow 2^{\lambda} = \frac{e}{N} \Rightarrow \lambda = \log_2 \frac{e}{N}$$

$$p_k = \frac{2^{\lambda}}{e} = \frac{2^{\log_2 e/N}}{e} = \frac{e/N}{e} = \frac{1}{N} \text{ . Entonces}$$

$$\mathcal{E}(X) = \sum_{k=1}^{N} \frac{1}{N} \log_2 N = \log_2 N .$$

Teorema 3.5.4 (C. Shannon)

X = texto con símbolos $S = \{x_1, \dots, x_N\}$; $p_k = Pr(x_k \in X)$, $k = 1, \dots, N$; $\mathcal{C}$ código binario con la condición del prefijo.

Entonces,

$$M_X(\mathcal{C}) \geq \mathcal{E}(X) = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k}$$

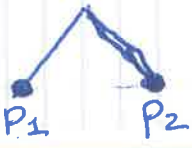
Además, existe un código binario $\mathcal{C}'$ con la condición del prefijo tal que

$$M_X(\mathcal{C}) \leq \mathcal{E}(X) + 1 .$$

3.5.2. EL CÓDIGO DE HUFFMAN

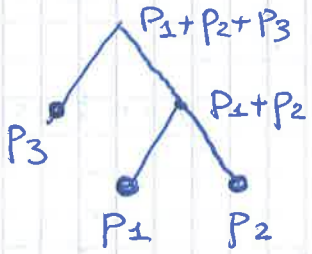
Este algoritmo permite hacer un código binario con la condición del prefijo que minimiza $M_X(\mathcal{C})$

1. $P(x) = Pr(x \in X)$. Ordenar de menor a mayor las probabilidades $P(x)$. Supongamos $p_1 \leq p_2 \leq \dots \leq p_N$
2. Tomar los símbolos con las **dos** probabilidades menores y asignarlos a dos hojas de un árbol binario

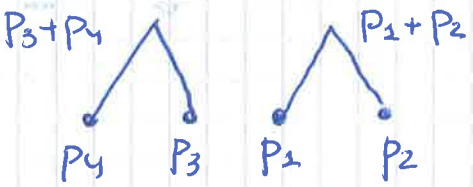


3. Calcular p_1+p_2 y ordenar la lista $p_1+p_2, p_3, p_4, \dots, p_N$ de menor a mayor

4'. Si tenemos $p_1+p_2 \leq p_3 \leq p_4 \leq \dots p_N$ o $p_3 \leq p_1+p_2 \leq p_4 \leq \dots \leq p_N$ hacemos el árbol



4''. Si tenemos $p_3 \leq p_4 \leq \dots \leq p_N$ hacemos el árbol



5. Ordenar la nueva lista de probabilidades según cada caso y continuar hasta asignar la última probabilidad a una hoja del árbol.

Ejercicio 3.5.4 a) Construir el código de Huffmann $\mathcal{C}$ para los símbolos $\{a, b, c, d, e, f\}$ de los que conocemos que aparecen en un texto X con probabilidades

$p(a) = 0.2$, $p(b) = 0.05$, $p(c) = 0.15$, $p(d) = 0.4$

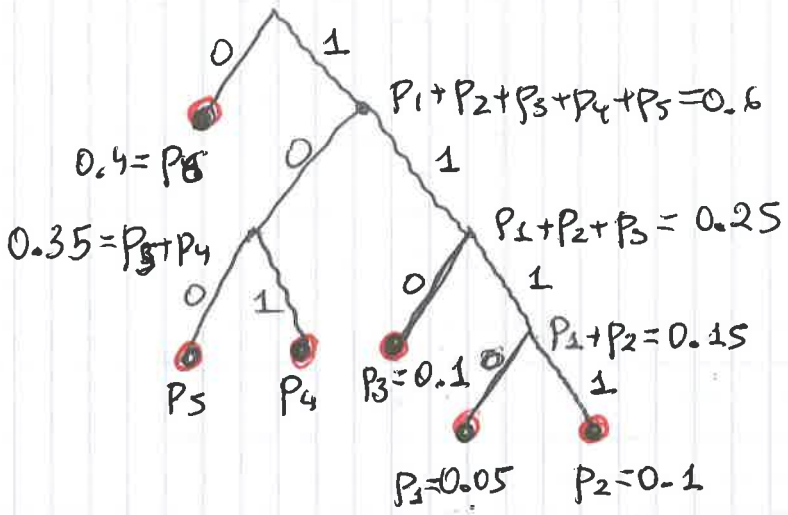
$p(e) = 0.1$, $p(f) = 0.1$

b) Calcula $\mathcal{C}(X)$ y $M_X(\mathcal{C})$

(a) $p_1 \le p_2 \le p_3 \le p_4 \le p_5 \le p_6$

$0.05 \quad 0.1 \quad 0.1 \quad 0.15 \quad 0.2 \quad 0.4$

$p(b) \quad p(e) \quad p(f) \quad p(c) \quad p(a) \quad p(d)$



$\{ \overline{p_3 \le p_1+p_2} \le p_4 \le p_5 \le p_6 \}$ Caso 4'

$0.1 \quad 0.15 \quad 0.15 \quad 0.2 \quad 0.4$

$\{ p_4 \le p_5 \le p_1+p_2+p_3 \le p_6$ Caso 4''

$0.15 \quad 0.2 \quad 0.25 \quad 0.4$

$p_1+p_2+p_3 = 0.25 \le p_5+p_4 = 0.35, \le p_6 = 0.4$ Caso 4'

$\{b \to 1110, \; e \to 111, \; f \to 110, \; c \to 101,$

$a \to 100, \; d \to 0 \}$

b) $\mathcal{E}(X) = \sum_{k=1}^{6} p_k \log_2 \frac{1}{p_k} = 0.05 \log_2 \frac{1}{0.05} + (0.1 \log_2 \frac{1}{0.1})2$

$+ 0.15 \log_2 \frac{1}{0.15} + 0.2 \log_2 \frac{1}{0.2} + 0.4 \log_2 \frac{1}{0.4} = 2.2841$

$M_X(\mathcal{E}) = (0.2) \times 3 + 4 \times (0.05) + 3 \times (0.15) + 1 \times (0.4) +$

$+ 4 \times (0.1) + 3 \times (0'1) = 2.35 \ge \mathcal{E}(X)$.

Ejercicio 3.5.5 (Para entregar el 24/07 (23/07), Sexta feira)

(a) Construir el código de Huffman $\mathcal{C}$ para el texto

$$X = 13524\,135\,13524\,135\,13524\,111$$

hecho con los símbolos de $S = \{1, 2, 3, 4, 5\}$.

b) Calcular $\mathcal{L}(X)$ y $M_X(\mathcal{C})$, comprobando que $M_X(\mathcal{C}) \geq \mathcal{L}(X)$