

Unidad 2: Teoría de la Información y Codificación

Información

Definición

Sea E un suceso que puede presentarse con probabilidad $P(E)$. Cuando E tiene lugar, decimos que hemos recibido

$$I(E) = \log \frac{1}{P(E)}$$

unidades de información.

- La unidad de medición depende de la base del logaritmo. Cuando usamos base 2 la unidad es el **bit**.
- Si $P = 1/2$, tendremos $I(E) = 1 \text{ bit}$. Es decir 1 bit es la información obtenida al especificar una de dos alternativas igualmente probables.

Fuente de memoria nula

- Una fuente de información es cualquier proceso que genera eventos.
- Una fuente de información es de **memoria nula** si la probabilidad de cada evento es independiente de los eventos anteriores.
- Es decir, los eventos son variables aleatorias independientes e idénticamente distribuidas.
- Una fuente de memoria nula puede describirse completamente con el alfabeto que representa los eventos (S) y la probabilidad de cada evento (P_i).

Información en una fuente de memoria nula

La presencia de un símbolo da una información igual a

$$I(s_i) = \log \frac{1}{P(s_i)} \text{ bits}$$

Dado que no son equiprobables, la cantidad media de información por símbolo de la fuente es

$$\sum_S P(s_i) I(s_i) \text{ bits}$$

Esta magnitud recibe el nombre de **entropía de la fuente de información**, o $H(S)$.

$$H(S) = \sum_S P(s_i) I(s_i) = - \sum_S P(s_i) \log_2 P(s_i)$$

Propiedades:

- $H(S)$ es una medida de la incertidumbre media de una fuente de información.
- Se puede interpretar también como una idea de la información media que provee cada símbolo.
- Una fuente que produce siempre el mismo símbolo no provee información, y por ende tendrá $H(S) = 0$

La entropía es máxima cuando todos los valores son equiprobables. En este caso:

- $P(s_i) = 1/n$
- ...
- $H(S) = \log_2 n$

Extensión de una fuente de memoria nula

Consideramos S^n , la extensión de orden n de la fuente de memoria nula S . Tenemos:

$$H(S^n) = n H(S)$$

Fuente de Markov

- Una **fuente de Markov de orden m** es una fuente de información en donde la probabilidad de un símbolo viene determinada por los m símbolos que lo preceden.
- En un determinado momento queda definido el estado de la fuente (la probabilidad asociada a cada símbolo) por los m estados anteriores. Puesto que existen q símbolos distintos, habrá q^m estados posibles, y cada nuevo evento generado modificará el estado.

Codificación

Codificar es establecer una correspondencia entre los símbolos de una fuente y los símbolos de un alfabeto de representación (*código*).

- Sólo cuando todos los símbolos de la fuente son equiprobables cada símbolo provee $\log_2 n$ bits.
- En otros casos, se puede utilizar el conocimiento de la fuente para lograr representaciones más eficientes eliminando redundancia. (ver *Eficiencia de códigos* más adelante)

Definición: Un código se dice **no singular** si todas las palabras del mismo son distintas.

Definición: Un código se dice **unívocamente decodificable** si, para cualquier n , la extensión de orden n es no singular. Es decir, una sucesión de palabras del código sólo puede representar una única sucesión de símbolos del alfabeto original.

Definición: Un código se dice **instantáneo** si es libre de prefijos. Es decir, puedo decodificar cualquier secuencia sin conocer los símbolos que la siguen.

Inecuación de Kraft

Consideremos

- $S = s_1, s_2, \dots, s_q$ (alfabeto fuente)
- $X = x_1, x_2, \dots, x_r$ (alfabeto código)
- X_1, X_2, \dots, X_q las palabras del código (una por cada símbolo de S)

La **inecuación de Kraft** es una condición necesaria y suficiente para que exista un código instantáneo de longitudes l_1, l_2, \dots, l_q :

$$\sum_{i=1}^q r^{-l_i} \leq 1$$

En el caso de un alfabeto de código binario, considerando un bloque de b símbolos a codificar:

$$\sum_{i=1}^b 2^{-l_i} \leq 1$$

Eficiencia de códigos

Se buscará asignar palabras más cortas a los símbolos más probables. Es decir, símbolos que brindan poca información deberían tener representaciones que ocupen poco espacio (y viceversa).

Consideremos:

- l_i longitud de la palabra que codifica el mensaje m_i
- r número de símbolos del alfabeto de código

Entonces, el número medio de símbolos utilizado para codificar S será:

$$L = \sum_i p_i l_i$$

Utilizando la inecuación de Kraft se puede deducir:

$$\frac{H(S)}{\log r} \leq L$$

Si expresamos la entropía en unidades r -arias, tenemos:

$$H_r(S) \leq L$$

Esto significa que el tamaño mínimo que podrá utilizar un código para un símbolo s_i será la cantidad de información que provee la aparición de ese símbolo. Un **codificador óptimo** es aquel que para codificar un mensaje utiliza la menor cantidad de bits posible.

Ejemplo: codificación de Huffman

La codificación de Huffman es una forma de definir códigos óptimos asumiendo que se conoce la probabilidad de ocurrencia de los símbolos, que la codificación es símbolo por símbolo y la probabilidad de ocurrencia de cada símbolo es independiente.

- Se utiliza tomando un texto a partir del cuál se extrae la frecuencia de cada símbolo
- Se organizan los símbolos en un árbol dependiendo de la frecuencia de cada uno.
- Este árbol deja más cerca de la raíz a los símbolos más frecuentes.
- El código de un símbolo será entonces el camino de la raíz al nodo donde está ubicado (utilizando ceros cuando se toma la rama izquierda y 1 cuando se toma la rama derecha).
- Así, los símbolos más frecuentes tendrán códigos más cortos.

Medios de transmisión y perturbaciones

Consideramos un modelo donde se envía un mensaje desde una fuente a un destino mediante un transmisor y un receptor. En el canal de transmisión puede haber distorsiones que hagan diferir la señal enviada de la recibida.

Posibles causas de las distorsiones:

- ruido: señales extrañas a la transmisión
- atenuación: pérdida de potencia de la señal por causa de la distancia
- distorsión de retardo: en medios guiados, los distintos armónicos no viajan a la misma velocidad (necesidad de ecualizar)

Ancho de banda de Nyquist

Es la máxima teórica velocidad de transmisión C (bits por segundo) de un canal. Depende de la velocidad de modulación B del canal expresada en baudios (símbolos por segundo) y la cantidad de niveles M :

$$C = B \log_2 M$$

Capacidad de Shannon

Es la máxima teórica velocidad de transferencia libre de errores de un canal, en la presencia de ruido.

$$C = B \log_2 \left(1 + \frac{S}{N} \right)$$

donde

- C es la capacidad (bits por segundo)
- B es el ancho de banda del canal (hertz)
- S es la potencia de la señal (watts)
- N es la potencia media del ruido (watts)

Inicialmente se pensaría que aumentando el ancho de banda arbitrariamente aumentará la capacidad de transmisión. Sin embargo, cuando aumenta el ancho de banda también aumenta el ruido, que introduce errores y por lo tanto baja la capacidad de transmisión.