Teoría de la Información y de la Transmisión

Codificación de la fuente: definiciones

Docente: *Nicolò Cesa-Bianchi*
Traducción: *Mario Román*
 versión 10 de julio de 2016

Los mensajes a transmitir son generados desde una entidad abstracta llamada fuente. Sea $\mathcal{X}$ el conjunto finito de símbolos que componen los mensajes generados por la fuente. Un mensaje $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ de longitud n es una secuencia de n símbolos fuente. Una función de codificación leva símbolos fuente a palabras de código. Una palabra de código es una secuencia de números del conjunto $\{0, \dots, D-1\}$ de los símbolos de código, donde $D > 1$ es la base del código. Por ejemplo, con $D = 2$ obtenemos los códigos binarios. En este sentido, podemos representar una función de codificación por un código fuente con

$$c : \mathcal{X} \to \{0, \dots, D-1\}^+$$

donde $\{0, \dots, D-1\}^+$ representa el conjunto de las secuencias sobre $\{0, \dots, D-1\}$ de longitud mayor o igual a uno. Formalmente,

$$\{0, \dots, D-1\}^+ = \bigcup_{n=1}^{\infty} \{0, \dots, D-1\}^n \ .$$

Por ejemplo, $(2, 1)$ y $(4, 7, 3)$ pertenecen ambas a $\{0, \dots, D-1\}^+$ con $D = 8$.

Ejemplo 1 *Dado $\mathcal{X} = \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$, un ejemplo de código binario $c : \mathcal{X} \to \{0, 1\}^+$ es el siguiente:*

$$c(\heartsuit) = 0 \quad c(\diamondsuit) = 010 \quad c(\clubsuit) = 01 \quad c(\spadesuit) = 10 \ .$$

Dado que el objetivo de un código fuente es maximizar la compresión, estamos interesados en medir la cantidad $\ell_c(x)$ definida como la longitud de la palabra de código $c(x)$ para el símbolo $x \in \mathcal{X}$. En el ejemplo precedente, $\ell_c(\diamondsuit) = 3$. El objetivo de un código fuente es el de minimizar la longitud media de la palabra de código utilizada para codificar los símbolos fuente.

La intución en la base de la construcción de códigos fuente es la misma del alfabeto Morse: utilizar palabras de código cortas para los símbolos que son generados frecuentemente por la fuente. Para poder analizar de modo riguroso debemos crear un modelo formal de la fuente. La propuesta de Shannon es la de definir una distribución de probabilidad p fijada sobre símbolos fuente y entonces asumir que $p(x)$ represente la probabilidad de que la fuente genere el símbolo $x \in \mathcal{X}$. Un **modelo de fuente** está entonces definido por la pareja $\langle \mathcal{X}, p \rangle$.

Nótese que, en realidad, la cantidad interesante aquí es la distribución de probabilidad sobre los mensajes $\boldsymbol{x}$ más que sobre los símbolos x, dado que los mensajes son los objetos que queremos

transmitir. Dada una distribución p sobre símbolos $\mathcal{X}$ definimos entonces una distribución P_n sobre los mensajes $\mathcal{X}^n$ de longitud n como

$$P_n(x_1, \ldots, x_n) = p(x_1) \times \cdots \times p(x_n) \ .$$

Esta definición de P_n corresponde a asumir que la fuente genera un mensaje a través de extracciones indipendendientes de símbolos. En generale, sin embargo, esta asunción no es muy plausibile. Por ejemplo, pensemos en un mensaje de texto en castellano donde los símbolos fuente son las letras del alfabeto incluyendo espacios y punto. Claramente, hay fuertes dependencias entre una letra del mensaje y las letras que están alrededor y tales dependencias no son capturadas por la P_n definida como antes. Por otro lado, el análisis matematico se ve muy facilitado por la asunción de independencia. En lo que sigue, asumiremos entonces la independencia de los símbolos fuente, teniendo sin embargo en mente que códigos fuente más realistas y sofisticados pueden obtenerse sin esta asunción.

De ahora en adelante identificamos un símbolo emitido desde la fuente mediante la variable aleatoria $X : \mathcal{X} \to \mathbb{R}$. Fijado D (base del código) indicamos con $\mathcal{D}$ el conjunto $\{0, \ldots, D-1\}$ de los símbolos de código con base D. Así, una función de codificación, o código, es una función del tipo $c : \mathcal{X} \to \mathcal{D}^+$.

Estamos preparados para definir formalmente el problema de la codificación fuente: dado un modelo de fuente $\langle \mathcal{X}, p \rangle$ y una base $D > 1$, encontrar un código $c : \mathcal{X} \to \mathcal{D}^+$ tal que el valore medio

$$\mathbb{E}\big[\ell_c\big] = \sum_{x \in \mathcal{X}} \ell_c(x)\, p(x) \tag{1}$$

de la longitud de la palabra de código sea mínimo.

Formulado en estos términos, el problema de la codificación fuente se presta a una solución banal e inútil. Es obvio que el código $c : \mathcal{X} \to \mathcal{D}^+$ tal que $c(x) = 0$ para cada $x \in \mathcal{X}$ minimiza $\mathbb{E}\big[\ell_c\big]$ para cada modelo de fuente. Por tanto, hace falta imponer limitaciones sobre la clase de códigos que queremos utilizar para resolver (1).

Una primera limitación es la siguiente. Un código $c : \mathcal{X} \to \mathcal{D}^+$ es **no singular** si a símbolos fuente distintos les corresponden palabras de código distintas. Formalmente, para cada $x, x' \in \mathcal{X}$ tal que $x \neq x'$ se cumple $c(x) \neq c(x')$. Dicho de otra forma, la no singularidad del código es equivalente a la inyectividad de la función de codificación. Esta es claramente una propiedad necesaria para un código utilizable en la práctica.

Ahora introducimos un concepto natural: el de **extensión de un código**. La extensión se usa para definir de forma simple la palabra de código asociada a un mensaje de longitud dada, es decir, a una secuencia de símbolos fuente. Dado un código $c : \mathcal{X} \to \mathcal{D}^+$, su extensión es la función $C : \mathcal{X}^+ \to \mathcal{D}^+$ definida come $C(x_1, \ldots, x_n) = c(x_1) \cdots c(x_n)$, donde $c(x_1) \cdots c(x_n)$ indica la secuencia obtenida yuxtaponiendo las palabras de código $c(x_1), \ldots, c(x_n)$.

Ejemplo 2 *La extensión C del código definido en el Ejemplo 1 es tal que*

$$C(\heartsuit, \spadesuit, \clubsuit) = c(\heartsuit)c(\spadesuit)c(\clubsuit) = 01001 \ .$$

La propiedad de no singularidad no es suficientemente fuerte para garantizar que esta sea heredada también por la extensión de un código. De hecho, la extensión en el Ejemplo 2 es tal que

$$C(\diamondsuit) = C(\clubsuit, \heartsuit) = C(\heartsuit, \spadesuit) = 010 \ .$$

Por tanto, mientras el código c del Ejemplo 1 es no singular su extensión C no lo es.

Motivados por este ejemplo, introducimos la noción de código **unívocamente decodificable**, es decir, de código cuya extensión es no singular. Formalmente, c es unívocamente decodificable si C es una función inyectiva. En la práctica, esta propiedad permite decodificar los mensajes. De hecho, si c es unívocamente decodificable entonces para cada $\boldsymbol{y} \in \mathcal{D}^+$ encontraré como mucho un único mensaje $\boldsymbol{x} \in \mathcal{X}^+$ (la decodificación de $\boldsymbol{y}$) tal que $C(\boldsymbol{x}) = \boldsymbol{y}$. La verficación para determinar si un código dado c es unívocamente decodificable la realiza el algoritmo de Sardinas-Patterson en tiempo $\mathcal{O}(mL)$, donde m es el número de las palabras de código y L es la suma de sus longitudes.