Compresión de Datos sin Pérdida

Códigos para distribuciones de probabilidad conocidas

Álvaro Martín

[1]Instituto de Computación,
Facultad de Ingeniería
almartin@fing.edu.uy

[2]PEDECIBA Informática

Códigos

Definición (Código fuente)

- Un *código de fuente* (o simplemente un *código*) C para una variable aleatoria X es una función,

$$C : \mathcal{X} \to \mathcal{B}^* ,$$

  donde $\mathcal{B} = \{0, 1\}$ y $\mathcal{B}^*$ es el conjunto de cadenas binarias de largo finito.
- A cada $x \in \mathcal{X}$ se asigna la *palabra de código* $C(x)$.
- Al *largo* de $C(x)$ lo denotamos $l(x)$ o $|C(x)|$.

Ejemplo
Para X definida sobre $\mathcal{X} = \{x_1, x_2, x_3\}$, $C(x_1) = 0$, $C(x_2) = 10$, $C(x_3) = 11$ es un código de fuente.

Largo medio de código

Definición (Largo medio de un código)
El *largo medio* de un código C, denotado $L(C)$, para una variable aleatoria $X \sim p$ se define como

$$L(C) = E[l(X)] = \sum_{x \in \mathcal{X}} p(x) l(x) .$$

Ejemplo

X	$p(X)$	$C(X)$
x_1	$1/2$	0
x_2	$1/4$	10
x_3	$1/8$	110
x_4	$1/8$	111

$$L(C) = H(X) = 1{,}75 \text{ bits.}$$

Clasificación de códigos

Definición (Código no singular)
Un código es *no singular* si es una función inyectiva, es decir, cada elemento de $\mathcal{X}$ se mapea a una palabra de código diferente

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)\,.$$

Definición (Extensión de un código)
La *extensión* C^* de un código C es un mapeo de $\mathcal{X}^*$ en $\mathcal{B}^*$ definida por

$$C^*(x_1 x_2 \ldots x_n) = C(x_1)C(x_2)\ldots C(x_n)\,,$$

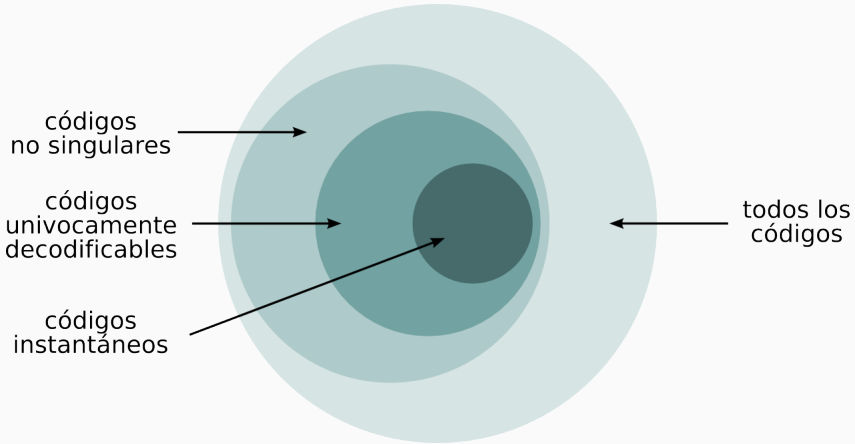
donde $C(x_1)C(x_2)\ldots C(x_n)$ es la concatenación de las palabras de código $C(x_1), C(x_2), \ldots, C(x_n)$.

Clasificación de códigos

Definición (Código unívocamente decodificable)
Un código es *unívocamente decodificable* si su extensión es no singular.

O sea, no hay ambigüedades al momento de decodificar una secuencia codificada.

Definición (Código instantáneo)
Un código es *instantáneo* o *de prefijo* si ninguna palabra de código es prefijo de otra palabra de código.



Ejemplos de códigos de diferentes clases

$\mathcal{X}$	C_1	C_2	C_3	C_4	C_5
x_1	0	0	10	0	0
x_2	0	010	00	10	10
x_3	0	01	11	110	110
x_4	0	10	110	1110	111

- C_1 es singular.
- C_2 es no singular pero no es unívocamente decodificable (UD). La secuencia 010 puede decodificarse como x_2, x_1x_4 o x_3x_1.
- C_3 es UD aunque no es instantáneo. Si se recibe $\ldots 110 \ldots$, decodificamos x_3 o x_4 dependiendo de la paridad de la cantidad de ceros entre 11 y el siguiente 1.
- C_4 es instantáneo (el 0 marca el final de cada palabra).
- C_5 es instantáneo.

Desigualdad de Kraft

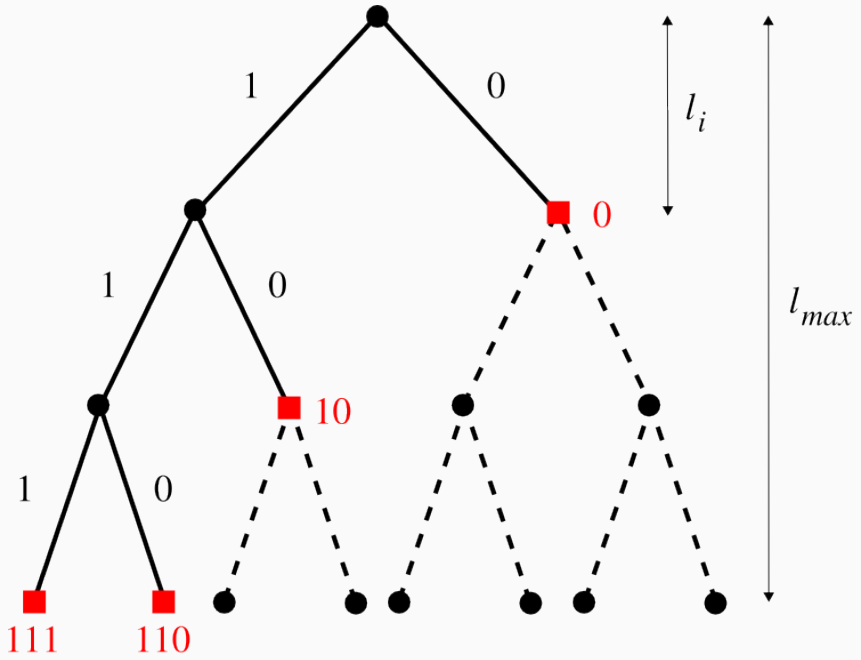
Teorema (Desigualdad de Kraft)
Para todo código instantáneo se cumple

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1 \,.$$

Recíprocamente, dado un conjunto de largos de palabra de código que satisfacen esta desigualdad, existe un código instantáneo con esos largos.

- Las longitudes de las palabras no pueden ser todas "cortas"; si hay una muy corta debe haber otras más largas.

Desigualdad de Kraft



$$\sum_{x \in \mathcal{X}} 2^{l_{\text{máx}} - l(x)} \leq 2^{l_{\text{máx}}}$$

Desigualdad de Kraft extendida

Teorema (Desigualdad de Kraft extendida)
Los largos de palabra de un código instantáneo definido para un alfabeto numerable, $\{l_i\}_{i\geq 1}$, satisfacen

$$\sum_{i=1}^{+\infty} 2^{-l_i} \leq 1.$$

Recíprocamente, dado un conjunto numerable de largos de palabra de código que satisfacen esta desigualdad, existe un código instantáneo con esos largos.

Desigualdad de Kraft extendida

Demostración.

- Sea $y_1 y_2 \dots y_{l_i}$ la i-ésima palabra, y sea

$$0.y_1 y_2 \dots y_{l_i} = \sum_{j=1}^{l_i} y_j 2^{-j} \,.$$

- $[0.y_1 y_2 \dots y_{l_i}, 0.y_1 y_2 \dots y_{l_i} + 2^{-l_i})$ contiene todos los reales con representación binaria finita que empieza con $0.y_1 y_2 \dots y_{l_i}$
- Todos estos intervalos son disjuntos.
- Por lo tanto, la suma de los anchos de estos intervalos debe ser menor o igual que la del intervalo $[0, 1]$ que los cubre,

$$\sum_{i=1}^{+\infty} 2^{-l_i} \leq 1 \,.$$

Desigualdad de Kraft extendida (recíproco)

- Ordenamos los largos de menor a mayor, $l_1 \leq l_2 \leq \ldots$.
- Para $m \geq 1$, definimos la palabra de código c_m como los l_m dígitos binarios a la derecha de la coma en la representación del número

$$f_m = \sum_{i=1}^{m-1} 2^{-l_i} .$$

- Sean m y m' arbitrarios, con $m < m'$. Tenemos

$$f_{m'} \geq \sum_{i=1}^{m} 2^{-l_i} = f_m + 2^{-l_m} .$$

- La representación binaria de $f_{m'}$ difiere de la de f_m en alguno de los primeros l_m dígitos binarios a la derecha de la coma.
- Por lo tanto c_m y $c_{m'}$ no pueden ser una prefija de la otra.

Desigualdad de Kraft para códigos UD I

La clase de los codigos UD es más grande que la de los instantáneos, sin embargo no presentan ninguna ventaja respecto a la longitud de las palabras de código.

Teorema (McMillan)
Los largos de palabra de código l_i de un código UD cumplen la desigualdad de Kraft,

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1 \,.$$

Desigualdad de Kraft para códigos UD II

Demostración
Asumimos primero que el código es finito y, para un natural k, escribimos

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k &= \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_k\in\mathcal{X}} 2^{-l(x_1)}2^{-l(x_2)}\ldots 2^{-l(x_k)} \\
&= \sum_{x_1,x_2,\ldots,x_k\in\mathcal{X}^k} 2^{-l(x_1)}2^{-l(x_2)}\ldots 2^{-l(x_k)} \\
&= \sum_{x^k\in\mathcal{X}^k} 2^{-l(x^k)}, \quad \text{donde } l(x^k)\triangleq\sum_{i=1}^{k} l(x_i) \\
&= \sum_{m=1}^{kl_{\text{máx}}} a(m)2^{-m},
\end{aligned}$$

donde $a(m)$ es la cantidad de secuencias x^k con $l(x^k)=m$.

Desigualdad de Kraft para códigos UD III

Demostración.
Como C es UD, su extensión de orden k es no singular, lo cual implica que $a(m) \leq 2^m$. Por lo tanto,

$$\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k = \sum_{m=1}^{kl_{\text{máx}}} a(m)2^{-m} \leq \sum_{m=1}^{kl_{\text{máx}}} 2^m 2^{-m} = kl_{\text{máx}}\,.$$

Entonces

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq (kl_{\text{máx}})^{\frac{1}{k}} \xrightarrow[k\to\infty]{} 1\,.$$

Finalmente, si el código no es finito, cualquier subconjunto finito de él es UD. Por lo tanto,

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} = \sum_{i=1}^{\infty} 2^{-l_i} = \lim_{N\to\infty} \sum_{i=1}^{N} 2^{-l_i} \leq 1\,.$$

Primer Teorema de Shannon

Teorema (Teorema de Codificación de Fuente (cota inferior))
El largo medio de un código unívocamente decodificable, C, para una variable aleatoria X, satisface

$$L(C) \geq H(X)\,,$$

con igualdad sii $p(x) = 2^{-l(x)}$ para todo $x \in \mathcal{X}$.

Demostración de cota inferior para $L(C)$

Demostración.
Sea $K = \sum_{x\in\mathcal{X}} 2^{-l(x)}$ y $q(x) = 2^{-l(x)}/K$, $x \in \mathcal{X}$.

$$\begin{aligned}
D(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) \\
&= -H(X) - \sum_{x\in\mathcal{X}} p(x)\log\frac{2^{-l(x)}}{K} \\
&= -H(X) - \sum_{x\in\mathcal{X}} p(x)\log 2^{-l(x)} + \sum_{x\in\mathcal{X}} p(x)\log K \\
&= -H(X) + \sum_{x\in\mathcal{X}} p(x)l(x) + \log K \\
&= -H(X) + L(C) + \log K \geq 0
\end{aligned}$$

La igualdad se da si $D(p||q) = 0$ y $K = 1$.

□

Código de Shannon

Los largos de código $l(x) = \lceil -\log p(x) \rceil$, $x \in \mathcal{X}$, satisfacen la desigualdad de Kraft

$$\sum_{x \in \mathcal{X}} 2^{-\lceil -\log p(x) \rceil} \leq \sum_{x \in \mathcal{X}} 2^{\log p(x)} = \sum_{x \in \mathcal{X}} p(x) = 1 \,.$$

- Se llama *código de Shannon* a un código con estos largos.
- Como $l(x) < -\log p(x) + 1$, el código de Shannon satisface

$$L(C) < H(X) + 1 \,.$$

- Por lo tanto, el largo de código medio óptimo, L^*, cumple

$$H(X) \leq L^* < H(X) + 1 \,.$$

Primer Teorema de Shannon (cota alcanzable)

Si consideremos una secuencia $x^n = (x_1, x_2, \ldots, x_n)$ como un símbolo individual en $\mathcal{X}^n$, el largo de código medio óptimo para x^n satisface

$$H(X_1, X_2, \ldots, X_n) \leq E\left[l(X_1, X_2, \ldots, X_n)\right] < H(X_1, X_2, \ldots, X_n)+1\,.$$

Si los símbolos $X_1, X_2, \ldots, X_n$ son i.i.d.,
$H(X_1, X_2, \ldots, X_n) = nH(X)$, de modo que el largo de código medio *por símbolo*,

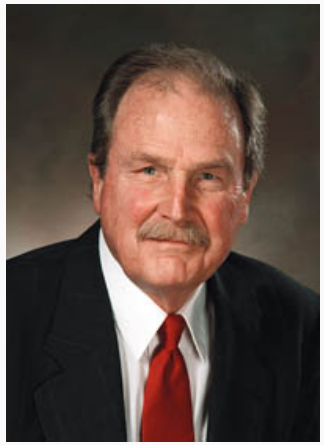
$$L_n \triangleq \frac{1}{n} E\left[l(X_1, X_2, \ldots, X_n)\right]\,,$$

satisface

$$H(X) \leq L_n < H(X) + \frac{1}{n}\,,$$

que puede hacerse arbitrariamente cercano a H.

Códigos de Huffman



David A. Huffman propuso en 1952 un algoritmo para construir un código instantáneo óptimo para una distribución de probabilidad arbitraria sobre un alfabeto finito.

Es un procedimiento recursivo que en cada paso agrupa los dos símbolos menos probables para formar un nuevo símbolo.

Ejemplo
Código de Huffman para $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5\}$, $X \sim p$, con $p = (0{,}25, 0{,}25, 0{,}2, 0{,}15, 0{,}15)$.

Optimalidad de los códigos de Huffman

Consideramos un alfabeto $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ de símbolos de una fuente y asumimos, sin pérdida de generalidad, que las probabilidades están ordenadas $p_1 \geq p_2 \geq \cdots \geq p_m$.

Lema
Todo código instantáneo óptimo para $(p_1, p_2, \ldots, p_m)$ *satisface*

1. *si* $p_j > p_k$, *entonces* $l_j \leq l_k$,
2. *para toda palabra de código de largo máximo existe otra del mismo largo que solo difiere en el último símbolo.*

Existe un código óptimo que adicionalmente satisface:

3. *las palabras de código asociadas a* x_m *y* x_{m-1} *difieren entre sí solo en el último símbolo.*

Optimalidad de los códigos de Huffman

Demostración.

1. Si $p_j > p_k$ pero $l_j > l_k$, intercambiando las palabras j y k obtenemos un código con largo medio menor.
2. Sea w una palabra de código de largo máximo y sea w' la cadena que resulta de invertir el último símbolo de w. Si w' no es una palabra de código, entonces podemos eliminar el último símbolo de w, obteniendo un nuevo código que es de prefijo y tiene un largo medio menor.
3. Sea C un código óptimo y sean w, w' palabras de código de largo máximo que difieren entre sí solo en el último símbolo. Como $p_m \leq p_{m-1} \leq p_i$, $1 \leq i < m-1$, intercambiando estas palabras de código con las asignadas a los símbolos x_m y x_{m-1} el largo medio no aumenta.

Definición inductiva de código de Huffman

Un código de Huffman, $C_H^{(m)}$, para $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, $X \sim p$, con $p_1 \geq p_2 \geq \cdots \geq p_m$, se define inductivamente como:

1. Para $m = 2$, definimos $C_H^{(m)} = \{0, 1\}$.
2. Para $m > 2$, sean

$$\begin{aligned} \mathcal{Y} &= \{y_1, y_2, \ldots, y_{m-2}, y_{m-1}\}, \\ p_Y &= (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m), \end{aligned}$$

y sea $C_H^{(m-1)}$ un código de Huffman para $Y \sim p_Y$. Definimos

$$C_H^{(m)}(x_i) = \begin{cases} C_H^{(m-1)}(y_i), & 1 \leq i < m-1, \\ C_H^{(m-1)}(y_{m-1})0, & i = m-1, \\ C_H^{(m-1)}(y_{m-1})1, & i = m. \end{cases}$$

Demostración de la optimalidad del código de Huffman

Para $m = 2$ es obvio; para $m > 2$:

- Sean $\mathcal{Y}$, p_Y y $C_H^{(m-1)}$ como en la definición.
- Sea $C^{(m)}$ un código óptimo para X tal que

$$C^{(m)}(x_{m-1}) = w0\,, \quad C^{(m)}(x_m) = w1\,.$$

- Sea $C^{(m-1)}$ el código para Y (no necesariamente óptimo),

$$C^{(m-1)}(y_i) = \begin{cases} C^{(m)}(x_i)\,, & 1 \le i < m-1\,, \\ w\,, & i = m-1\,. \end{cases}$$

- Los largos esperados de estos códigos satisfacen

$$\begin{aligned} L_H^{(m)} &= L_H^{(m-1)} + p_{m-1} + p_m\,, \\ L^{(m-1)} &= L^{(m)} - (p_{m-1} + p_m)\,. \end{aligned}$$

  Sumando y reordenando obtenemos

$$L_H^{(m)} = L^{(m)} + \left(L_H^{(m-1)} - L^{(m-1)}\right) \le L^{(m)}\,.$$