Teoria dell'Informazione

Simone Alessandro Casciaro
08 Ottobre 2024

Lezione 3: Disuguaglianza di Kraft

Codici Istantanei

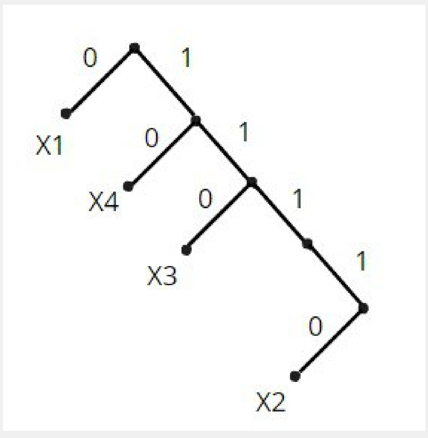
I codici istantanei, oltre alla possibilità di essere immediatamente decodificati, presentano un'altra proprietà.
È possibile capire se esiste un codice istantaneo per un messaggio generato dalla sorgente, anche senza conoscere il codice.
NOTA: Senza il codice, non è possibile determinare se il codice usato dalla sorgente è istantaneo, ma solo capire se può esisterne uno. Questa tecnica è più utile per capire se un codice NON è istantaneo, con operazioni meno onerose computazionalmente rispetto alla ricerca esaustiva.

Per fare ciò, è necessario il vettore l, che contiene la lunghezza delle codifiche delle parole del messaggio.

> Esempio
> $$l = \begin{cases} 1 & 0 \\ 5 & 11110 \\ 4 & 1100 \\ 2 & 10 \end{cases}$$
> Questo codice è istantaneo, quindi con $l = (1, 5, 4, 2)$ è possibile costruire un codice istantaneo. Non è detto che quello usato dalla sorgente lo sia, ad esempio:
> $$l = \begin{cases} 1 & A \\ 5 & AAAAA \\ 4 & BBAA \\ 2 & BB \end{cases}$$
> che non è istantaneo.

Albero di Codifica

Un albero di codifica è una rappresentazione grafica delle codifiche appartenenti all'immagine della funzione c.

> Esempio
> $c(x_1) = 0$
> $c(x_2) = 11110$
> $c(x_3) = 1100$
> $c(x_4) = 10$
>
> 
>

Disuguaglianza di Kraft

Ipotesi

Dati:

- una sorgente $X = \{x_1, \dots, x_m\}$
- d base del codice
- m numero di elementi del messaggio
- $l_c = (l_1, \dots, l_m) > 0$ lunghezze delle codifiche dei simboli

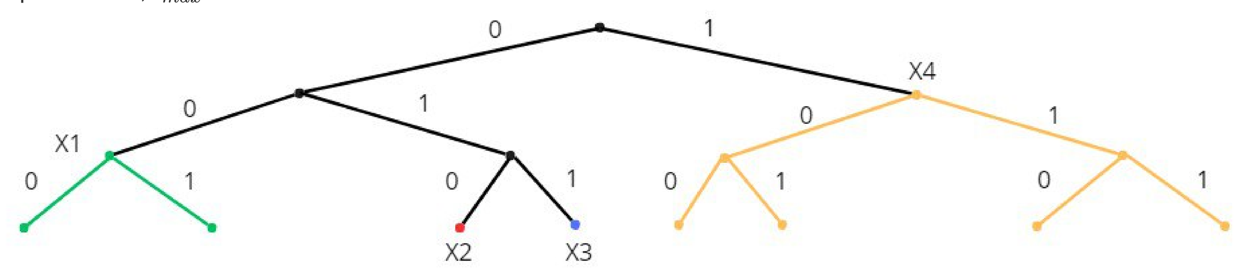
Tesi

$$\exists \text{ codice istantaneo } c : X \to D^+ \text{t.c. } l_c(x_i) = l_i \; \forall i = 1, \dots, m$$
$$\iff \sum_{i=1}^{m} d^{-l_i} \leq 1$$

Dimostrazione

$\boxed{\implies}$
Supponiamo di avere un codice istantaneo c. Dato che questa è una dimostrazione per costruzione, lo dimostriamo con un esempio.
$c(x_1) = 00$
$c(x_2) = 010$
$c(x_3) = 011$
$c(x_4) = 1$

Sia $l_{max} = \max_{i=1,\dots,m} l_c(x_i)$ la massima altezza dell'albero di codifica. Disegnamo l'albero di codifica di c.
In questo caso, $l_{max} = 3$.



DIciamo che ogni $c(x_i)$ "copre" il suo sottoalbero e chiamiamo A_i l'insieme delle foglie all'interno del sottoalbero generato dal nodo $c(x_i)$. Dato che c è un codice istantaneo, tutti gli insiemi A sono disgiunti tra loro.
Il numero di foglie totali è $d^{l_{max}}$.

$$\sum_{i=1}^{m} d^{l_{max} - l_i} = \sum_{i=1}^{m} |A_i| \leq d^{l_{max}}$$

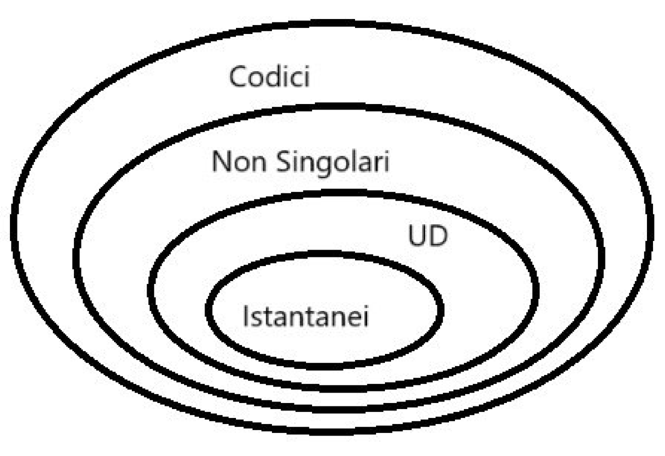
Prendendo gli estremi di questa disuguaglianza e dividendo tutto per $d^{l_{max}}$, otteniamo

$$\sum_{i=1}^{m} d^{-l_i} \leq 1$$

$\boxed{\impliedby}$
Sempre per costruzione, partiamo dal vettore $l = (2, 1, 3, 3)$
Controlliamo la sommatoria per vedere se rispettiamo le ipotesi della disuguaglianza.
$\sum_{i=1}^{m} d^{-l_i} = \frac{1}{2^2} + \frac{1}{2^1} + \frac{1}{2^3} + \frac{1}{2^3} = \frac{1}{4} + \frac{1}{2} + \frac{1}{8} + \frac{1}{8} = 1 \leq 1$
Possiamo costruire un albero di codifica in modo tale che il codice che rappresenta sia istantaneo (Ad esempio quello usato per la prima parte della dimostrazione). Dato che si riesce a costruirlo, $\exists c$ codice istantaneo. $\square$

Codici di Shannon

In seguito a tutto il discorso detto, abbiamo scoperto come i codici istantanei siano migliori di quelli univocamente decodificabili, che sono migliori dei non singolari.



Tuttavia, restringere l'insieme dei codici possibili comporta un innalzamento del minimo che noi stiamo cercando.
Ricordiamo infatti, che noi vogliamo minimizzare il valore atteso delle lunghezze delle codifiche.
L'idea di Shannon è di minimizzare il valore atteso delle lunghezze avendo però un codice che rispetta la disuguaglianza di Kraft.

$$\begin{cases} \min_{l_1,\dots,l_m} \sum_{i=1}^{m} l_i p_i & \text{Minimizzare Valore Atteso} \\ \sum_{i=1}^{m} d^{-l_i} \leq 1 & \text{Disuguaglianza di Kraft} \end{cases}$$

Per comodità, indichiamo con p_i la probabilità della parola x_i di lunghezza l_i.
Per la disugaglianza di Kraft, sappiamo che $\sum_{i=1}^{m} d^{-l_i} \leq 1$. Per il secondo assioma di Kolmogorov, $1 = \sum_{i=1}^{m} p_i$.
Dato che vogliamo che questa proprietà valga per la sommatoria, possiamo farla valere per ogni elemento (richiesta più stringente).
$d^{-l_i} \leq p_i \quad \forall i = 1, \dots, m$ e dunque

$$l_i \geq \log_d \frac{1}{p_i}$$

Scegliamo cioè le lunghezze della codifica in base alla probabilità con cui un elemento del messaggio è presente nella sorgente.
Dato che voglio minimizzare il valore atteso, sceglierò il più piccolo l_i possibile, ovvero

$$l_i = \lceil \log_d \frac{1}{p_i} \rceil$$

Se $l_i = \log_d \frac{1}{p_i} \quad \forall i$, cioè senza approssimare, allora si ha che:

$$\mathbb{E}(l_c) = \sum_{i=1}^{m} l_i p_i = \sum_{i=1}^{m} p_i * \log_d \frac{1}{p_i} = H$$

chiamiamo questo valore Entropia.

Esercizi

1. Costruiamo un codice per la sorgente $X = \{x_1, x_2, x_3, x_4\}$ con $d = 2$ e $\mathbb{P} = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right\}$

   Seguendo la disuguaglianza di Kraft, scegliamo:
   $l_1 = \lceil \log_2 2 \rceil = 1$
   $l_2 = \lceil \log_2 4 \rceil = 2$
   $l_3 = \lceil \log_2 8 \rceil = 3$
   $l_4 = \lceil \log_2 8 \rceil = 3$

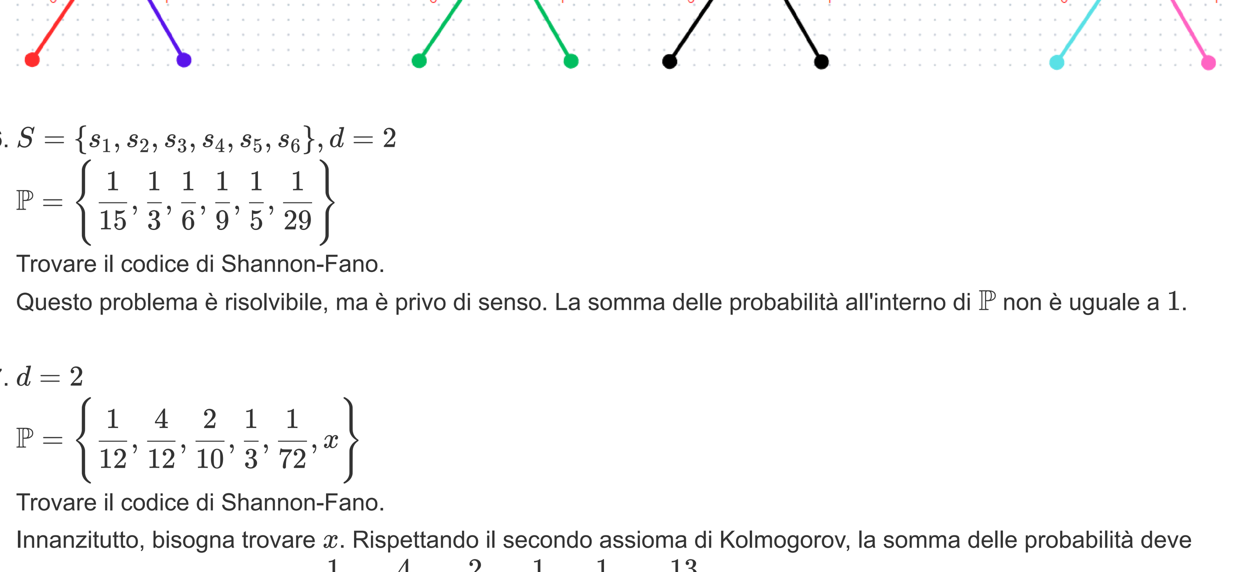
   Assegniamo ai vari x_i la loro codifica.
   $c(x_1) = 0$
   $c(x_2) = 10$
   $c(x_3) = 110$
   $c(x_4) = 111$

2. Supponiamo di avere una sorgente di due simboli tali per cui $p(x_1) = 0.1$ e $p(x_2) = 0.9$. Se dovessimo seguire Shannon, avremmo:
   $l_1 = \lceil \log_2 10 \rceil = 4$
   $l_2 = \lceil \log_2 \frac{1}{0.9} \rceil = 1$
   Dunque avremmo $c(x_1) = 0000$ e $c(x_2) = 1$ che non è ottimale.

3. $m = 4$
   $c : 1, 011, 01, 111$
   Com'è questo codice?
   Non è istantaneo (perché 01 è prefisso di 011, ad esempio)
   Non è univocamente decodificabile (perché la stringa 111, ad esempio, è ambigua)

4. $m = 5$
   $c : 1, 001, 0000, 01, 0001$
   Com'è questo codice?
   Questo è un codice a virgola con carattere terminatore 1, dunque è istantaneo.

5. $m = 5$
   $c : 000, 001, 01, 111, 110$
   Com'è questo codice?
   Costruendo l'albero di codifica, si nota che tutti i sottoalberi generati dagli x_i sono disgiunti tra loro. Questo rende il codice istantaneo. Tuttavia, dato che non tutte le foglie sono coperte da un nodo x_i, il codice non è ottimale.

6. $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}, d = 2$
   $\mathbb{P} = \left\{\frac{1}{15}, \frac{1}{3}, \frac{1}{6}, \frac{1}{9}, \frac{1}{5}, \frac{1}{29}\right\}$
   Trovare il codice di Shannon-Fano.
   Questo problema è risolvibile, ma è privo di senso. La somma delle probabilità all'interno di $\mathbb{P}$ non è uguale a 1.

7. $d = 2$
   $\mathbb{P} = \left\{\frac{1}{12}, \frac{4}{12}, \frac{2}{10}, \frac{1}{3}, \frac{1}{72}, x\right\}$
   Trovare il codice di Shannon-Fano.
   Innanzitutto, bisogna trovare x. Rispettando il secondo assioma di Kolmogorov, la somma delle probabilità deve essere 1. Dunque $x = 1 - \frac{1}{12} - \frac{4}{12} - \frac{2}{10} - \frac{1}{3} - \frac{1}{72} = \frac{13}{360}$
   Avendo tutte le probabilità, usiamo il metodo di Shannon per calcolare le lunghezze.
   $l_1 = \lceil \log_2 12 \rceil = 4$
   $l_2 = \lceil \log_2 3 \rceil = 2$
   $l_3 = \lceil \log_2 5 \rceil = 3$
   $l_4 = \lceil \log_2 3 \rceil = 2$
   $l_5 = \lceil \log_2 72 \rceil = 7$
   $l_6 = \lceil \log_2 \frac{360}{13} \rceil = 5$

   Verifichiamo che può esistere un codice istantaneo (ovvero, che rispetta la disuguaglianza di Kraft).
   $\frac{1}{2^4} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^2} + \frac{1}{2^7} + \frac{1}{2^5} = \frac{1}{16} + \frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{1}{128} + \frac{1}{32} = \frac{93}{128} \leq 1$

   Possiamo costruire un codice istantaneo.
   Lo facciamo ordinando gli l_i in ordine crescente.
   $l_2 = 2 \implies c(x_2) = 00$
   $l_4 = 2 \implies c(x_4) = 01$
   $l_3 = 3 \implies c(x_3) = 100$
   $l_1 = 4 \implies c(x_1) = 1010$
   $l_6 = 5 \implies c(x_6) = 10110$
   $l_5 = 7 \implies c(x_5) = 1011100$