(F1X) Teoria dell'Informazione e della Trasmissione

Codici di Huffman

Docente: *Nicolò Cesa-Bianchi* versione 15 marzo 2016

Si noti che mentre il codice di Shannon è quasi ottimo in media, può essere altamente inefficiente per quanto riguarda la lunghezza delle singole parole di codice. Infatti, consideriamo una sorgente $\mathcal{X} = \{x_1, x_2\}$ con $p(x_1) = 2^{-10}$ e $p(x_2) = 1 - 2^{-10}$. Allora, nel caso di codifica binaria, Shannon produrrà un codice istantaneo c con lunghezze $\ell_c(x_1) = \lceil \log_2 2^{10} \rceil = 10$ e $\ell_c(x_2) = \left\lceil \log_2 \frac{2^{10}}{2^{10}-1} \right\rceil = 1$. D'altra parte, un codice istantaneo più ovvio è $c(x_1) = 0$ e $c(x_2) = 1$ con lunghezze $\ell_c(x_1) = 1$ e $\ell_c(x_2) = 1$.

Terminata l'analisi del codice di Shannon, passiamo al problema di trovare il codice istantaneo "ottimo". Ovvero, dato un modello di sorgente $\langle \mathcal{X}, p \rangle$ con $|\mathcal{X}| = m$, il codice istantaneo D-ario c^* le cui lunghezze $\ell_1^*, \dots, \ell_m^*$, risolvono il problema di ottimizzazione

$$\min_{\ell_1,\dots,\ell_m \in \mathbb{N}} \sum_{i=1}^m p_i \ell_i \qquad \text{tale che} \quad \sum_{i=1}^m D^{-\ell_i} \le 1 \ .$$

Il codice istantaneo ottimo è il codice di Huffman (si veda l'esempio in Figura 1). L'algoritmo per la costruzione del codice nel caso generale $D > 1$ è il seguente:

1. i simboli sorgente vengono ordinati in base alle probabilità;
2. si crea un nuovo modello di sorgente in cui i D simboli meno frequenti sono rimpiazzati da un nuovo simbolo con probabilità pari alla somma delle loro probabilità;
3. se la nuova sorgente contiene più di un simbolo si ricomincia dal passo 1.

L'albero di codifica viene costruito come segue: all'inizio, tutti i simboli sorgente sono foglie. Ogni volta che D simboli vengono rimpiazzati da un nuovo simbolo si crea un albero avente il nuovo simbolo come radice e le radici degli alberi corrispondenti ai simboli rimpiazzati come figli (in ordine arbitrario, quindi il codice di Huffman non è unico). È facile vedere che l'intera procedura richiede un tempo dell'ordine $\mathcal{O}(|\mathcal{X}| \log |\mathcal{X}|)$.

Dato che ad ogni passo la nuova sorgente ha $D - 1$ simboli in meno della sorgente precedente, perché l'algoritmo termini in modo corretto —cioè con una sorgente di esattamente un simbolo— è necessario che $|\mathcal{X}|$ sia divisibile per $D - 1$ col resto di 1. Ovvero, $|\mathcal{X}| = (D-1)k + 1$ per un qualche intero positivo k. Se non esiste un tale k, allora aggiungiamo alla sorgente un numero sufficiente di simboli "dummy" con probabilità pari a zero.

Procediamo ora a dimostrare l'ottimalità del codice di Huffman nell'ambito dei codici sorgente istantanei. Prima di dimostrare il teorema, dobbiamo però fare una semplice osservazione preliminare. Ovvero: da un codice di Huffman D-ario per una sorgente di $m - D + 1$ simboli possiamo ricavare un codice di Huffman D-ario per una sorgente di m simboli semplicemente sostituendo un simbolo sorgente con D nuovi simboli cosicché le probabilità assegnate ad essi siano tutte più piccole di quelle dei rimanenti $m - D$ vecchi simboli.

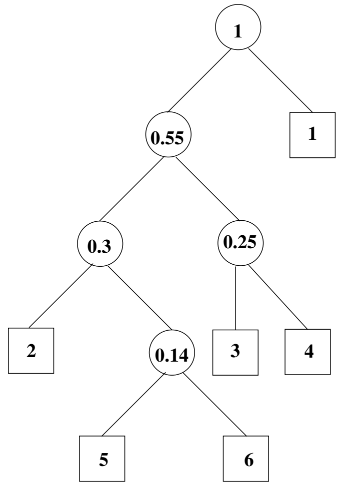

Figura 1: Albero del codice binario di Huffmann per la sorgente $\langle \mathcal{X}, p \rangle$ con $\mathcal{X} = \{1, \ldots, 6\}$ e probabilità $p_1 = 0.45$, $p_2 = 0.16$, $p_3 = 0.13$, $p_4 = 0.12$, $p_5 = 0.09$ e $p_6 = 0.05$.

Fatto 1 *Sia c' un codice D-ario di Huffman per la sorgente $\mathcal{X}' = \{x_1, \ldots, x_{m-D+1}\}$ con probabilità $p_1 \geq \cdots \geq p_{m-D+1}$. Sia $\mathcal{X}$ la sorgente di m simboli $\{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{m+1}\}$ ottenuta da $\mathcal{X}'$ togliendo x_k e aggiungendo D nuovi simboli $x_{m-D+2}, \ldots, x_{m+1}$ con probabilità $p_{m-D+2}, \ldots, p_{m+1}$ tali che $0 < p_{m-D+2}, \ldots, p_{m+1} < p_{m-D+1}$ e $p_{m-D+2} + \cdots + p_{m+1} = p_k$. Allora il codice*

$$c(x) = \begin{cases} c'(x) & \text{se } x \in \{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{m-D+1}\}, \\ c'(x_k)i & \text{se } x = x_{m-D+i+2} \text{ per } i = 0, \ldots, D-1. \end{cases} \tag{1}$$

è un codice di Huffman per la sorgente $\mathcal{X}$.

Dimostrazione. La dimostrazione è ovvia considerando che il dopo il primo passo nella costruzione del codice di Huffman per $\mathcal{X}$ otteniamo $\mathcal{X}'$ come nuova sorgente. Quindi i due codici differiscono solo per le codifiche ai D simboli $x_{m-D+2}, \ldots, x_{m+1}$ che sono quelli meno probabili in $\mathcal{X}$. Per definizione dell'algoritmo di Huffman, le codifiche dei simboli meno probabili di $\mathcal{X}$ sono definite in termini del codice di Huffman per $\mathcal{X}'$ esattamente come descritto da (1). $\square$

Teorema 2 *Data una sorgente $\langle \mathcal{X}, p \rangle$ e dato $D > 1$, il codice D-ario c di Huffman minimizza $\mathbb{E}[\ell_c]$ fra tutti i codici D-ari istantanei per la medesima sorgente.*

Dimostrazione. Per semplicità, dimostriamo solo il caso particolare $D = 2$. La dimostrazione del caso generale $D > 1$ è lasciata come esercizio.

Procediamo per induzione su $|\mathcal{X}| = m$. Nel caso base $m = 2$ Huffman è ottimo. Infatti, è facile vedere che l'algoritmo di Huffman produce il codice $c(x_1) = 0$ e $c(x_2) = 1$ che è ottimo per ogni distribuzione di probabilità p su $\{x_1, x_2\}$.

Assumiamo quindi $m > 2$ con l'ipotesi induttiva che Huffman sia ottimo per $k \leq m - 1$. Fissiamo $\langle \mathcal{X}, p \rangle$ arbitraria. Siano $u, v \in \mathcal{X}$ tale che $p(u)$ e $p(v)$ sono minime. Definiamo la sorgente $\langle \mathcal{X}', p' \rangle$ dove $u, v \in \mathcal{X}$ sono rimpiazzati da $z \in \mathcal{X}'$ e dove

$$p'(x) = \begin{cases} p(x) & \text{se } x \neq z, \\ p(u) + p(v) & \text{se } x = z. \end{cases}$$

Sia c' il codice di Huffman per $\langle \mathcal{X}', p' \rangle$. Dato che $|\mathcal{X}'| = m - 1$, c' è ottimo per ipotesi induttiva. Definiamo ora il codice c per $\mathcal{X}$ come

$$c(x) = \begin{cases} c'(x) & \text{se } x \notin \{u, v\}, \\ c'(z)0 & \text{se } x = u, \\ c'(z)1 & \text{se } x = v. \end{cases}$$

Per il Fatto 1, c è di Huffman per $\langle \mathcal{X}, p \rangle$. Vogliamo ora dimostrare che c è anche ottimale. Cominciamo col dimostrare la seguente relazione che ci servirà in seguito

$$\begin{aligned} \mathbb{E}[\ell_c] &= \sum_{x \in \mathcal{X}} \ell_c(x) p(x) \\ &= \sum_{x \in \mathcal{X}'} \ell_{c'}(x) p'(x) - \ell_{c'}(z) p'(z) + \ell_c(u) p(u) + \ell_c(v) p(v) \\ &= \mathbb{E}[\ell_{c'}] - \ell_{c'}(z) p'(z) + \big(\ell_{c'}(z) + 1\big) p(u) + \big(\ell_{c'}(z) + 1\big) p(v) \\ &= \mathbb{E}[\ell_{c'}] - \ell_{c'}(z) p'(z) + \ell_{c'}(z) p'(z) + p'(z) \\ &= \mathbb{E}[\ell_{c'}] + p'(z) \ . \end{aligned} \tag{2}$$

Per dimostrare l'ottimalità di c consideriamo un altro codice istantaneo c_2 per $\langle \mathcal{X}, p \rangle$ e verifichiamo che $\mathbb{E}[\ell_c] \leq \mathbb{E}[\ell_{c_2}]$. Fissato c_2, siano $r, s \in \mathcal{X}$ tali che $\ell_{c_2}(r)$ e $\ell_{c_2}(s)$ sono massime.

Esaminiamo ora le posizioni delle foglie r e s nell'albero di codifica per c_2 (che esiste perché c_2 è istantaneo). Se r e s sono fratelli, non facciamo nulla. Se r o s hanno un fratello (diciamo r ha un fratello f), allora possiamo scegliere r e f tali che $\ell_{c_2}(r)$ e $\ell_{c_2}(f)$ sono massime invece di r e s. Se invece né r né s hanno un fratello nell'albero, allora possiamo sostituire alla codifica di ciascun nodo la codifica del padre finché ci riportiamo nella situazione in cui r e s hanno entrambi un fratello. Così facendo riduciamo $\mathbb{E}[\ell_{c_2}]$. Quindi, senza perdita di generalità, possiamo assumere che c_2 sia tale che r e s sono fratelli.

Ora trasformiamo c_2 in un codice $\widetilde{c}_2$ per la stessa sorgente scambiando la codifica di u con quella di r e la codifica di v con quella di s. Ovvero,

$$\widetilde{c}_2(x) = \begin{cases} c_2(x) & \text{se } x \notin \{u, v, r, s\}, \\ c_2(u) & \text{se } x = r, \\ c_2(r) & \text{se } x = u, \\ c_2(v) & \text{se } x = s, \\ c_2(s) & \text{se } x = v. \end{cases}$$

Ora esaminiamo la differenza fra la lunghezza media di $\widetilde{c}_2$ e c_2,

$$\begin{aligned}
\mathbb{E}[\ell_{\widetilde{c}_2}] - \mathbb{E}[\ell_{c_2}] &= \sum_{x\in\mathcal{X}} p(x)\Big(\ell_{\widetilde{c}_2}(x) - \ell_{c_2}(x)\Big) \\
&= p(r)\ell_{c_2}(u) + p(u)\ell_{c_2}(r) + p(s)\ell_{c_2}(v) + p(v)\ell_{c_2}(s) \\
&\quad - p(u)\ell_{c_2}(u) - p(r)\ell_{c_2}(r) - p(v)\ell_{c_2}(v) - p(s)\ell_{c_2}(s) \\
&= \underbrace{\Big(p(r) - p(u)\Big)}_{\geq 0}\underbrace{\Big(\ell_{c_2}(u) - \ell_{c_2}(r)\Big)}_{\leq 0} + \underbrace{\Big(p(s) - p(v)\Big)}_{\geq 0}\underbrace{\Big(\ell_{c_2}(v) - \ell_{c_2}(s)\Big)}_{\leq 0} \\
&\leq 0 \ .
\end{aligned}$$

I segni delle differenze sono determinati dalla scelta di u, v, r, s in quanto

$$\max\big\{p(u), p(v)\big\} \leq \min\big\{p(r), p(s)\big\} \qquad \min\big\{\ell_{c_2}(r), \ell_{c_2}(s)\big\} \geq \max\big\{\ell_{c_2}(u), \ell_{c_2}(v)\big\} \ .$$

Quindi abbiamo dimostrato che $\mathbb{E}[\ell_{\widetilde{c}_2}] \leq \mathbb{E}[\ell_{c_2}]$.

Notiamo ora che, dopo lo scambio con r e s, u e v sono diventati fratelli in $\widetilde{c}_2$. Quindi esiste $w \in \{0,1\}^*$ tale che $\widetilde{c}_2(u) = w0$ e $\widetilde{c}_2(v) = w1$. Allo scopo di applicare l'ipotesi induttiva, introduciamo un codice c_2' per $\langle \mathcal{X}', p' \rangle$ definito come segue

$$c_2'(x) = \begin{cases} \widetilde{c}_2(x) & \text{se } x \neq z, \\ w & \text{se } x = z. \end{cases}$$

Possiamo allora scrivere, ricordando che $p'(z) = p(u) + p(v)$,

$$\begin{aligned}
\mathbb{E}[\ell_{\widetilde{c}_2}] &= \sum_{x\in\mathcal{X}' : x\neq z} p'(x)\ell_{\widetilde{c}_2}(x) + p(u)\big(\ell_{c_2'}(z) + 1\big) + p(v)\big(\ell_{c_2'}(z) + 1\big) \\
&= \sum_{x\in\mathcal{X}' : x\neq z} p'(x)\ell_{\widetilde{c}_2}(x) + p'(z)\ell_{c_2'}(z) + p'(z) \\
&= \mathbb{E}[\ell_{c_2'}] + p'(z) \ .
\end{aligned}$$

Ricordando allora le disuguaglianze precedentemente ottenute, e utilizzando l'ipotesi induttiva per stabilire $\mathbb{E}[\ell_{c'}] \leq \mathbb{E}[\ell_{c_2'}]$, prossiamo quindi scrivere

$$\mathbb{E}[\ell_c] = \mathbb{E}[\ell_{c'}] + p'(z) \leq \mathbb{E}[\ell_{c_2'}] + p'(z) = \mathbb{E}[\ell_{\widetilde{c}_2}] \leq \mathbb{E}[\ell_{c_2}] \ .$$

Questo conclude la dimostrazione. $\square$