Teoria dell'Informazione/RTI **Anno Accademico 2022–2023**

Lezione 9

Ugo Vaccaro

Abbiamo già visto che la entropia $H(X)$ di una v.c. X può essere vista come una misura dell'incertezza media a priori che abbiamo sui valore che la variabile casuale (v.c.) X potrà assumere. In questa lezione vedremo un'ulteriore interpretazione dell'entropia $H(X)$ di un v.c. X. Tale interpretazione è collegata al seguente problema: data una v.c. X con distribuzione $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, come possiamo in pratica realizzare un esperimento che abbia esiti $x_1, x_2, \ldots, x_n$ con probabilità $p_1, p_2, \ldots, p_n$, rispettivamente? Ovviamente, il tutto dipende dalle risorse a nostra disposizione. Effettuiamo un'ipotesi minimale al riguardo, ovvero supponiamo di avere a nostra disposizione solo una moneta bilanciata che genera TESTA (o, equivalentemente, 0), e CROCE (o, equivalentemente, 1), con probabilità 1/2 ciascuno. Consideriamo il seguente esempio.

Esempio 1 *Avendo a disposizione una moneta bilanciata, vogliamo generare la seguente variabile casuale X:*

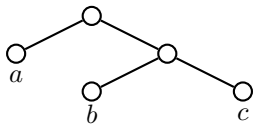
$$X = \begin{pmatrix} a & b & c \\ 1/2 & 1/4 & 1/4 \end{pmatrix}$$

Un metodo potrebbe essere il seguente. Lanciamo la moneta, se esce 0 *poniamo $X = a$, se esce* 1 *rilanciamo una seconda volta la moneta. Se esce* 10 *dai due lanci poniamo $X = b$, se esce* 11 *dai due lanci poniamo $X = c$. É chiaro che questo medtodo dà $P\{X = a\} = 1/2, P\{X = b\} = 1/4, P\{X = c\} = 1/4$, come desideravamo. Qual è il numero medio di lanci (bits 0 e 1) che abbiamo effettuato? Esso sarà pari a $1 \times (1/2) + 2 \times (1/4) \times 2 \times (1/4) = 1.5$. Guarda caso, $H(X) = (1/2)\log 2 + (1/4)\log 4 + (1/4)\log 4 = 1.5$.*

Per descrivere un generico algoritmo per generare distribuzioni di probabilità attraverso lanci di monete, utilizziamo una struttura ad albero, che è semplicemente un modo per "ricordare" gli esiti dei lanci della moneta. Quindi, un algoritmo di generazione è rappresentato da un albero in cui:

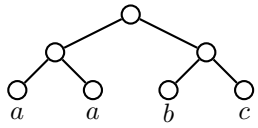
1. ogni nodo, tranne le foglie (che hanno zero figli), ha esattamente 2 figli (tali alberi vengono detti *completi*);
2. ogni foglia è etichettata con uno dei valori x assunti dalla variabile casuale X che vogliamo generare;
3. la regola è che ogni lancio di moneta che ha esito 0 ci fà andare dal nodo corrente dell'albero al suo figlio sinistro ed ogni lancio di moneta che ha esito 1 ci fà andare dal nodo corrente dell'albero al suo figlio destro. Quando ariviamo su di una foglia "emettiamo" il valore x ad essa associato e ripartiamo dalla radice.

Tale rappresentazione è perfettamente generale, e permette di descrivere un qualunque algoritmo di generazione di variabili casuali. L'albero per la generazione della variabile casuale dell'**Esempio 1** ottenuto attraverso il metodo appena descritto sarebbe:



Ci possono essere più alberi che generano la stessa variabile casuale. Ad esempio, anche quest'albero genera la

variabile casuale dell'**Esempio 1**



Questo secondo albero, però, usa in media 2 lanci di moneta per generare ciascun valore della v.c. X, quindi è peggiore dell'albero precedente che usava solo 1.5 lanci in media.

In generale, si pone quindi il seguente problema: *data un'arbitraria variabile casuale X, qual è il minimo numero medio di lanci di moneta necessari e sufficienti per generare X?*

Dalla definizione di albero che genera variabili casuali, si evince immediatamente che per ogni foglia f che appare ad un qualche livello k dell'albero (ovvero per cui esiste un percorso dalla radice dell'albero ad f composto da k rami dell'albero), vale che la probabilità $P(f)$ della foglia è pari a 2^{-k}. Inoltre, ciò che deve ulteriormente valere per un albero che rapresenta un corretto algoritmo che genera la v.c. X, è che per ogni valore x della variabile casuale X si deve avere:

$$\sum_{f:f \text{ è etichettata con } x} P(f) = P\{X = x\} = P(x). \tag{1}$$

Il numero medio di lanci di monete necessario sarà ovviamente

$$\sum_{f:f \text{ foglie di } T} P(f) \times (\text{livello cui appare la foglia } f \text{ in } T). \tag{2}$$

Denotiamo tale ultima quantità con il simbolo $E[T]$. Essa può essere vista come il valor medio dei livelli cui le foglie dell'albero T appaiono (pesati dalle probabilità delle rispettive foglie), e viene generalmente chiamata *altezza media* di T.

Per un dato albero T e foglia f di T, denotiamo con $k(f)$ il livello cui appare f in T. Consideriamo la distribuzione di probabilità indotta sulle foglie di T dai lanci di monete. Come abbiamo già detto, ogni foglia avrà probabilità $P(f) = 2^{-k(f)}$ per cui l'entropia H di tale distribuzione sarà pari a

$$H = \sum_{f:f \text{ foglia di } T} 2^{-k(f)} \log \frac{1}{2^{-k(f)}} = \sum_{f:f \text{ foglia di } T} k(f) 2^{-k(f)} = E[T] \tag{3}$$

dove l'ultima eguaglianza è a causa della (2).

Quest'ultimo fatto ha già un'interessante conseguenza. Se la v.c. X che vogliamo generare è di tipo molto particolare, ovvero per se per ogni valore x che la v.c. X può assumere vale che la probabilità $P\{X = x\}$ è della forma $2^{-k(x)}$, per qualche intero $k(x)$ (tali distribuzioni di probabilità vengono dette *diadiche*), allora un semplice albero T che genera X consiste nell'albero completo T avente una foglia $f(x)$ associata ad ogni possibile valore x che la variabile casuale X può assumere, e la profondità della foglia $f(x)$ nell'albero è esattamente pari a $k(x)$ (nell'**Esempio 1** la variabile casuale X considerata era diadica e l'albero che la generava era ottenuto attraverso la considerazione appena fatta). In più, l'altezza media $E[T]$ dell'albero T così ottenuto e che genera X (e quindi il numero medio di lanci di monete che ci occorrono per generare X mediante T) è esattamente pari all'entropia $H(X)$ di X (ciò in virtù della (3)). Studiamo ora il caso *generale* di v.c. X in cui le probabilità $P\{X = x\}$ *non* sono di tipo diadico. Presentiamo prima un risultato intermedio.

Lemma 1 *Sia Y una v.c. e f una funzione. Vale*

$$H(f(Y)) \leq H(Y).$$

Dimostreremo questo risultato in seguito, come conseguenza di risultati più generali. Per il momento, diamo il seguente argomento per convincerci della sua correttezza. Supponiamo che la v.c. X assuma valori in $\{x_1, \ldots, x_n\}$.

Innanzitutto, osserviamo che se f è iniettiva, allora ovviamente la v.c. $f(Y)$ ha la stessa distribuzione di X (in quanto la f non fà altro che "rietichettare" i valori y assunti dalla Y nei valori $f(y)$ assunti dalla $f(Y)$, senza cambiare le rispettive probabilità. Se invece f non è iniettiva, allora esisteranno un certo numero $k \geq 2$ di valori $x_{i_1}, \ldots, x_{i_1}$ assunti dalla X, con probablità $P\{X = x_{i_1}\} = p(x_{i_1}), \ldots, P\{X = x_{i_k}\} = p(x_{i_k})$ rispettivamente, per cui $f(x_{i_1}) = \ldots = f(x_{i_k}) = y$. Varrà che $P\{f(X) = y\} = p(x_{i_1}) + \ldots + p(x_{i_k})$. Supponiamo che la f sia iniettiva sul resto di X, ovvero su $X \setminus \{x_{i_1}, \ldots, x_{i_1}\}$ (l'argomento si estende facilmente al caso generale, raggruppando in insiemi disgiunti i valori assunti dalla v.c. X che hanno la stessa immagine secondo la funzione f). Allora varrà

$$\begin{aligned}
H(f(Y)) - H(Y) &= (p(x_{i_1}) + \ldots + p(x_{i_k})) \log \frac{1}{p(x_{i_1}) + \ldots + p(x_{i_k})} - p(x_{i_1}) \log \frac{1}{p(x_{i_1})} - \ldots - p(x_{i_k}) \log \frac{1}{p(x_{i_k})} \\
&= p(x_{i_1}) \log \frac{1}{p(x_{i_1}) + \ldots + p(x_{i_k})} + \ldots + p(x_{i_k}) \log \frac{1}{p(x_{i_1}) + \ldots + p(x_{i_k})} \\
&\qquad - p(x_{i_1}) \log \frac{1}{p(x_{i_1})} - \ldots - p(x_{i_k}) \log \frac{1}{p(x_{i_k})} \\
&= p(x_{i_1}) \log \frac{p(x_{i_1})}{p(x_{i_1}) + \ldots + p(x_{i_k})} + \ldots + p(x_{i_k}) \log \frac{p(x_{i_k})}{p(x_{i_1}) + \ldots + p(x_{i_k})} \\
&\leq 0
\end{aligned}$$

dove l'ultima diseguaglianza è conseguenza del fatto che gli argomenti del logaritmo sono tutti ≤ 1. Possiamo ora provare il seguente risultato.

Teorema 1 *Per ogni albero T che genera la v.c. X, il numero medio di lanci di monete usate è necessariamente pari* almeno *all'entropia di X, ovvero*

$$E[T] \geq H(X). \tag{4}$$

Dimostrazione. Dato l'albero T, sia Y la variabile casuale con distribuzione di probabilità

$$P = \{P(f) : f \text{ è foglia di } T\},$$

dove ricordiamo $P(f) = 2^{-k(f)}$ e $k(f)$ è la profondità della foglia f nell'albero T. Per quanto prima osservato, vale che $H(Y) = E[T]$. Ora, sulla base delle considerazioni che ci hanno portato alla (1), possiamo osservare la v.c. X è ottenibile mediante l'applicazione di una funzione f alla v.c. Y (di fatto la funzione f mappa uno o più valori della v.c. Y, ovvero una o più foglie di T, in un qualche singolo valore della v.c. X). Dal Lemma 1 otteniamo

$$E[T] = H(Y) \geq H(f(Y)) = H(X).$$

□

Quindi, per generare una qualsiasi v.c. X ci occorrono in media un numero di lanci di moneta pari almeno all'entropia di X. Ricordiamo che se la v.c. X assume valori x con probabilità della forma $2^{-k(x)}$, con $k(x)$ intero, (ovvero se la distribuzione è diadica), allora un albero che genera X e che usa un numero medio di lanci di moneta proprio pari a $H(X)$ (e quindi il *migliore* possibile, in virtù del Teorema 1) lo sappiamo costruire.

Consideriamo ora il caso di v.c. generali, ovvero con distribuzioni non necessariamente diadiche. L'idea è di esprimere ogni probabilità $P(x)$, con cui la generica v.c. X assume valore x, come somma di potenze di $1/2$, ovvero di valori della forma $2^{-k_1}, 2^{-k_2}, \ldots$. In altri termini, l'idea è di ottenere inannzitutto la rappresentazione del numero $P(x)$ in base $1/2$. Ad esempio, se $P(x) = 7/8$, potremmo scrivere $7/8 = 1/2 + 1/4 + 1/8$, successivamente potremmo costruire un albero T che ha una foglia etichettata x al livello 1 (ovvero di probabilità $1/2$), una foglia etichetatta x al livello 2 (e quindi di probabilità $1/4$), ed infine una foglia etichetatta x al livello 3 (e quindi di probabilità $1/8$). Cosicchè, usando l'albero T e lanciando la moneta bilanciata, arriveremo in una qualsiasi delle tre foglie etichettate x con probablità pari a $1/2 + 1/4 + 1/8 = 7/8$, che è esattamente la probabilità che la v.c. X assuma valore x. Chiaramente, questo procedimento lo possiamo ripetere per ogni valore x che la

v.c. assume, ed ottenere quindi un albero che genera la corrispondente distribuzione di probabilità $P = \{P(x) : \forall x \text{ valore assunto da } X\}$.

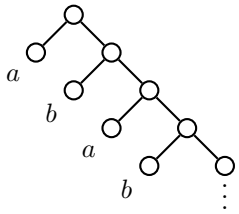
Esempio 2 *Sia*

$$X = \begin{pmatrix} a & b \\ 2/3 & 1/3 \end{pmatrix}$$

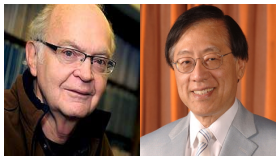
Scrivendo $2/3$ *ed* $1/3$ *in binario otteniamo*

$$\frac{2}{3} = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \dots \qquad \frac{1}{3} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \dots$$

e l'albero che genera X sarebbe



che in questo caso è infinito, in quanto sia la rappresentazione binaria di $2/3$ che di $1/3$ ha un numero infinito di termini (ma ciò non ci disturba affatto).



D.E. Knuth A.C. Yao

Il seguente Teorema rappresenta il risultato principale della lezione e và sotto il nome di Teorema di Knuth e Yao, dal nome degli scopritori.

Teorema 2 *Il minimo numero medio* $E[T]$ *di lanci di moneta richiesto da ogni algoritmo che genera la v.c.* X *è tale che*

$$H(X) \leq E[T] < H(X) + 2,$$

dove $H(X) = -\sum_x P\{X = x\} \log\{X = x\}$ *è l'entropia della v.c.* X.

Dimostrazione. La diseguaglianza $E[T] \geq H(X)$ l'abbiamo già dimostrata nel Teorema 1. Sia $\mathbf{p} = (p_1, p_2, \dots, p_n)$ la distribuzione di probabilità della v.c. X. Esprimiamo ogni p_i come somma di potenze di $1/2$. Più precisamente scriviamo

$$p_i = p_i^{(1)} + p_i^{(2)} + \dots \tag{5}$$

dove ogni termine $p_i^{(j)}$ o è pari a 0 oppure è un'opportuna potenza di $1/2$. Usando tutti i temini $p_i^{(j)}$, per ogni i e per ogni j, scriviamo la nuova distribuzione di probabilità (che può avere un numero infinito di termini ma, ancora, ciò non ci disturba affatto) data da:

$$(p_1^{(1)}, p_1^{(2)}, \dots, p_2^{(1)}, p_2^{(2)}, \dots, p_n^{(1)}, p_n^{(2)}, \dots) \tag{6}$$

e costruiamo un albero completo T che ha una foglia associata ad ogni una delle potenze di $1/2$ che compare in (5). Ovvero, se $p_i^{(j)} = 2^{-k_{ij}}$, per qualche intero k_{ij}, allora l'albero T avrà una foglia alla profondità k_{ij}. Questa costruzione è sicuramente possibile a causa della Diseguaglianza di Kraft, mostrata nella Lezione 4.[1] Infatti vale

[1] Tecnicamente, abbiamo dimostrato la Diseguaglianza di Kraft nel caso *finito*, ma è semplice verificare che la dimostrazione a suo tempo data funziona anche nel caso infinito.

che

$$1 = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \sum_{j \geq 1} p_i^{(j)} = \sum_{i=1}^{n} \sum_{j \geq 1 : p_i^{(j)} > 0} 2^{-k_{ij}},$$

quindi un albero con foglie nelle giuste profondità k_{ij} esiste. Sia Y la v.c con distribuzione di probabilità data dalla (6). Poichè la distribuzione (6) è diadica, sappiamo dalla (3) che

$$E[T] = H(Y). \tag{7}$$

Ora, la v.c che vogliamo generare è funzione della v.c. Y, nel senso che per ottenere X da Y etichettiamo tutte le foglie aventi probabilità $p_i^{(j)}$ che appaiono nell'espansione (5) di p_i con lo stesso valore x_i della v.c. X, ovvero con il valore x_i che ha probabilità di occorrere pari a p_i. Osserviamo che

$$H(Y) = -\sum_{i=1}^{n} \sum_{j \geq 1} p_i^{(j)} \log p_i^{(j)} = \sum_{i=1}^{n} \sum_{j : p_i^{(j)} > 0} j 2^{-j}$$

in quanto, come abbiamo più volte detto, ogni $p_i^{(j)}$ o è 0 oppure è una potenza di 1/2. Sia

$$T_i = \sum_{j : p_i^{(j)} > 0} j 2^{-j} \qquad \text{cosicchè} \qquad H(Y) = \sum_{i=1}^{n} T_i. \tag{8}$$

Notiamo che per ogni probabilità p_i possiamo sempre trovare un numero naturale n_i tale che $2^{-(n_i - 1)} > p_i \geq 2^{-n_i}$, ovvero

$$n_i - 1 < -\log p_i \leq n_i, \tag{9}$$

per cui $p_i^{(j)} > 0$ solo se $j \geq n_i$, e quindi possiamo riscrivere T_i come

$$T_i = \sum_{j : j \geq n_i, p_i^{(j)} > 0} j 2^{-j}.$$

Ricordando che

$$\forall i = 1, \ldots, n \qquad p_i = \sum_{j : j \geq n_i, p_i^{(j)} > 0} 2^{-j}$$

passiamo a provare che

$$T_i < -p_i \log p_i + 2 p_i. \tag{10}$$

Calcolando la differenza tra T_i e $-p_i \log p_i + 2p_i$ otteniamo:

$$\begin{aligned}
T_i + p_i \log p_i - 2p_i &< T_i - p_i(n_i - 1) - 2p_i \qquad \text{(dalla (9))} \\
&= T_i - (n_i - 1 + 2)p_i \\
&= \sum_{j:j\geq n_i, p_i^{(j)}>0} j2^{-j} - (n_i+1) \sum_{j:j\geq n_i, p_i^{(j)}>0} 2^{-j} \\
&= \sum_{j:j\geq n_i, p_i^{(j)}>0} (j - n_i - 1)2^{-j} \\
&= -2^{-n_i} + 0 + \sum_{j:j\geq n_i+2, p_i^{(j)}>0} (j - n_i - 1)2^{-j} \\
&= -2^{-n_i} + \sum_{k:k\geq 1, p_i^{(k+n_i+1)}>0} k2^{-(k+n_i+1)} \qquad \text{(cambiando variabili nella somma)} \\
&\leq -2^{-n_i} + \sum_{k:k\geq 1} k2^{-(k+n_i+1)} \qquad \text{(estendendo la somma su tutti i termini)} \\
&= -2^{-n_i} + 2^{-(n_i+1)} \sum_{k:k\geq 1} k2^{-k} \\
&= -2^{-n_i} + 2^{-(n_i+1)} \times 2 \qquad \text{(essendo } \sum_{k:k\geq 1} k2^{-k} = 2\text{)} \\
&= 0.
\end{aligned}$$

e quindi la (10) è vera. Ricordando, dalla (7) e (8), che

$$E[T] = H(Y) = \sum_{i=1}^{n} T_i,$$

otteniamo dalla (10) appena provata che che

$$E[T] = \sum_{i=1}^{n} T_i < -\sum_{i=1}^{n} (p_i \log p_i - 2p_i) = -\sum_{i=1}^{n} p_i \log p_i + 2\sum_{i=1}^{n} p_i = H(X) + 2.$$

Con questo la prova del Teorema è completata. □