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

Lezione 11

Ugo Vaccaro

Nella lezione scorsa abbiamo esaminato il problema di valutare l'evoluzione nel tempo del capitale T di uno scommettitore coinvolto in un gioco di scommesse. Dette $X_1, X_2, \ldots$ le variabili casuali (vc) che descrivono l'esito della gara $1, 2, \ldots$, abbiamo assunto che esse fossero indipendenti ed indenticamente distribuite, ovvero che valesse

$$\forall i \geq 1 \qquad X_k = X = \begin{pmatrix} 1 & 2 & \ldots & m \\ p_1 & p_2 & \ldots & p_m \end{pmatrix}, \tag{1}$$

dove con p_i denotiamo la probabilità che li-esimo cavallo vinca la gara k-esima, e tale probabilità è sempre la stessa per ogni $k \geq 1$.

Inoltre abbiamo descritto la strategia del giocatore mediante il vettore $\mathbf{b} = (b_1, \ldots, b_m)$ e quella dell'agenzia mediante il vettore $\mathbf{o} = (o_1, \ldots, o_m)$ con i significati già illustrati nella scorsa lezione.

Detto

$$W(\mathbf{p}, \mathbf{b}, \mathbf{o}) = \sum_{i=1}^{m} p_i \log o_i b_i$$

avevamo osservato che, per n molto grande, valeva che il capitale del giocatore era pari a

$$\approx T 2^{nW(\mathbf{p},\mathbf{b},\mathbf{o})}.$$

Inoltre, la strategia ottimale (ovvero quella che massimizzava il capitale) la si otteneva scegliendo $\mathbf{b} = \mathbf{p}$, ed in tal caso valeva che il capitale del giocatore era pari a

$$\approx T 2^{n[\sum_i p_i \log o_i - H(\mathbf{p})]}.$$

Infine, sotto l'ipotesi che l'agenzia fosse "onesta", ovvero che valesse $\sum_i (1/o_i) = 1$, avevamo ottenuto che, per *arbitraria* strategia $\mathbf{b}$ del giocatore, ed *arbitraria* strategia $\mathbf{r} = (r_1, \ldots, r_m)$ dell'agenzia (dove $r_i = (1/o_i)/(\sum_k (1/o_k))$, valeva che il capitale del giocatore era pari a

$$\approx T 2^{n[D(\mathbf{p}||\mathbf{r}) - D(\mathbf{p}||\mathbf{b})]}.$$

La domanda che ci poniamo in questa lezione (ed a cui vogliamo fornire una risposta) è: come cambia la velocità di crescita del capitale se disponiamo di una qualche informazione ausiliaria, descritta da una vc Y?

Preliminare alla risoluzione della domanda di sopra è l'introduzione (motivata!) di una misura dell'informazione che una vc Y fornisce su di un'altra vc X. Come abbiamo (lungamente...) argomentato a lezione, una ragionevole misura di quanta informazione Y fornisce su X potrebbe essere data dalla divergenza informazionale $D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y))$ tra la congiunta $\mathbf{p}(xy)$ della coppia di vc XY, e la distribuzione di probabilità che avremmo su XY se esse fossero indipendenti, ovvero tra la distribuzione $\mathbf{p}(xy)$ e la distribuzione $\mathbf{p}(x)\mathbf{p}(y)$. Infatti, sappiamo che $D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y)) = 0$ *se e solo se* $\mathbf{p}(xy) = \mathbf{p}(x)\mathbf{p}(y)$ (per tutti i possibili valori x che la vc X può assumere e per tutti i possibili valori y che la vc Y può assumere). Detto in altre parole, $D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y)) = 0$ se e solo se X e Y sono *indipendenti* (e ciò soddisfa la nostra intuizione, in quanto se X e Y fossero indipendenti, allora è ovvio che l'informazione che Y ci può fornire su X è pari a zero). D'altra parte, la divergenza informazionale $D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y))$ "aumenta" man mano che la congiunta $\mathbf{p}(xy)$ differisce dal prodotto $\mathbf{p}(x)\mathbf{p}(y)$, ovvero aumenta all'aumentare della dipendenza tra X e Y (ed anche ciò soddisfa la nostra intuizione, in quanto se X e Y sono "molto" dipendenti, allora è ovvio che l'informazione che Y ci può fornire su X è "grande", e nel

caso estremo in cui la conoscenza di Y ci permette di determinare *esattamente* X dovremmo avere la massima informazione possibile). Calcoliamo esplicitamente $D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y))$ per chiarire ulteriormente la questione. Si ha

$$
\begin{aligned}
D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y)) &= \sum_{x,y} p(xy)\log\frac{p(xy)}{p(x)p(y)} && (2)\\
&= \sum_{x,y} p(xy)\log\frac{p(x|y)}{p(x)} \qquad \text{(in quanto } \frac{p(xy)}{p(y)} = p(x|y)\text{)} && (3)\\
&= \sum_{x,y} p(xy)\log\frac{1}{p(x)} - \sum_{x,y} p(xy)\log\frac{1}{p(x|y)} && (4)\\
&= \sum_{x} p(x)\log\frac{1}{p(x)} - \sum_{x,y} p(xy)\log\frac{1}{p(x|y)} \qquad \text{(in quanto } \sum_y p(xy) = p(x)\text{)} && (5)\\
&= \sum_{x} p(x)\log\frac{1}{p(x)} - \sum_{y} p(y)\sum_x p(x|y)\log\frac{1}{p(x|y)} \qquad \text{(in quanto } p(xy) = p(y)p(x|y)\text{)} && (6)\\
&= H(X) - \sum_{y} p(y)\sum_{x} p(x|y)\log\frac{1}{p(x|y)}. && (7)
\end{aligned}
$$

Osserviamo ora che il termine

$$\sum_x p(x|y)\log\frac{1}{p(x|y)}$$

di fatto è una entropia, per ogni valore fissato y. Ovvero, è l'entropia della vc X condizionata dal fatto che $Y = y$. La denotiamo quindi con $H(X|Y = y)$ e scriviamo

$$\sum_y p(y)\sum_x p(x|y)\log\frac{1}{p(x|y)} = \sum_y p(y)H(X|Y = y).$$

Chiameremo il valor medio delle entropie

$$\sum_y p(y)H(X|Y = y)$$

che compare nella (7) *entropia di X condizionata da Y* e la denoteremo con $H(X|Y)$. D'altra parte, sappiamo che la divergenza informazionale $D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y))$ è sempre non negativa, ed è pari a zero se e solo se le vc X e Y sono indipendenti. Abbiamo quindi:

$$D(\mathbf{p}(xy)||\mathbf{p}(x)\mathbf{p}(y)) = H(X) - H(X|Y) \geq 0$$

da cui otteniamo che, per ogni vc X e Y vale la diseguaglianza

$$H(X) \geq H(X|Y)$$

con uguaglianza se e solo se X e Y sono indipendenti. Il tutto ha un'ovvia ed intuitiva interpretazione: *ogni qualvolta condizioniamo i valori di una vc casuale X con valori di una vc Y, la sua entropia (media) può solo decrescere* (e rimane inalterata se e solo se X e Y sono indipendenti, come è ragionevole aspettarsi). Denotiamo (per comodità di notazione) con $I(X;Y)$ la quantità

$$H(X) - H(X|Y) = I(X;Y) \geq 0.$$

Visto che

$$H(X|Y) = H(X) - I(X;Y)$$

si ha che $I(X;Y)$ rappresenta di quanto decresce la nostra incertezza su X (ovvero, $H(X)$) quando conosciamo Y (che sarà ora $H(X|Y)$). Poichè solo "informazione" può far decrescere l'incertezza, è ragionevole chiamare $I(X;Y)$ la (mutua) informazione di Y su X. Il temine *mutua* si riferisce al fatto che, ovviamente, dalla definizione si evince che la $I(\cdot;\cdot)$ è simmetrica nelle sue variabili, ovvero

$$I(X;Y) = I(Y;X).$$

Notiamo, infine, che varrà $I(X;X) = H(X) - H(X|X) = H(X)$, per cui la mutua informazione è, in un certo senso, una estensione dell'entropia $H(X)$.

Avendo terminato la ricerca di una ragionevole una misura dell'informazione che una vc Y fornisce su di un'altra vc X, ed avendola individuata in $I(X;Y)$, ritorniamo al nostro problema di partenza. Cerchiamo quindi di quantificare come cambia la velocità di crescita del capitale se disponiamo di una qualche informazione ausiliaria sugli esiti della corsa (prima esclusivamente rappresentati dalla vc X in (1)), dove tale informazione ausiliaria è descritta da una vc Y.

La prima considerazione da fare è che adesso le probabilità di vittoria di ciascun cavallo i sono condizionate da i valori assunti dalla vc ausiliaria Y (imaginiamo i valori assunti da Y come delle informazioni che una "spia" ci fornisce). Pertanto, il punto di partenza sono adesso le probabilità di vittoria condizionate, date da

$$\Pr\{X = i|Y = y\} \qquad \forall i = 1, \ldots, m,\ \forall y$$

da interpretarsi come la probabilità che il cavallo i-esimo vinca la gara, dato che abbiamo saputo dalla "spia" la notizia y. Ciò comporta che anche la strategia del giocatore $\mathbf{b}$ potrà dipendere dai valori y assunti da Y. Di conseguenza, mentre prima avevamo solo la successione di vc $X_1, X_2, \ldots$, dove ciascuna X_i era uguale alla vc (1), adesso lo scenario è descritto da una successione di coppie di vc

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_i, Y_i), \ldots$$

Assumeremo che tale successione sia composta da vc indipendenti ed identicamente distribuite, il che implica anche che

$$\forall k \geq 1 \qquad \Pr\{X_k = x, Y_k = y\} = p(xy)$$

per ogni possibile cavallo x e possibile fatto y che la spia ci può dire. Ragionando come fatto nella lezione scorsa, abbiamo che il capitale iniziale T dopo n gare viene moltiplicato per un fattore S_n dato da[1]

$$S_n = \prod_{k=1}^{n} b(X_k|Y_k)o(X_k)$$

dove

$$\forall k \geq 1 \qquad o(X_k) = O(X) = \begin{pmatrix} o_1 & o_2 & \ldots & o_m \\ p_1 & p_2 & \ldots & p_m \end{pmatrix}.$$

e $\forall k \geq 1$ $b(X_k|Y_k)$ è una vc che assume valori $b(i|y)$ con probabilità $\Pr\{X = i, Y = y\}$. Ovvero, assume i valori pari alla percentuale del capitale che il giocatore decide di puntare sul cavallo i-esimo, dato che ha ricevuto dalla "spia" la notizia y. Formalmente, $b(X_k|Y_k)$ è la seguente vc:

$$b(X_k|Y_k) = b(X|Y) = \begin{pmatrix} b(1|y_1) & \ldots & b(m|y_1) & \ldots & b(1|y_s) & \ldots & b(m|y_s) \\ \Pr\{X = 1, Y = y_1\} & \ldots & \Pr\{X = m, Y = y_1\} & \ldots & \Pr\{X = 1, Y = y_s\} & \ldots & \Pr\{X = m, Y = y_s\} \end{pmatrix},$$

dove abbiamo denotato con $y_1, \ldots, y_s$ i possibili messaggi che la spia può fornire al giocatore. Siamo interessati a stimare il valore S_n quando il numero di gare n è molto grande, per cui effettuiamo le seguenti considerazioni:

[1] Qui usiamo una terminologia leggermente diversa da quella usata nella lezione scorsa, per evidenziare che la strategia del giocatore dipende dai valori assunti dalla vc Y, mentre le quote pagate dall'agenzia no (i valori assunti da Y sono noti al giocatore e quindi possono influenzare la sua strategia. I valori assunti da Y non sono noti all'agenzia di scommesse...).

$$
\begin{aligned}
\frac{1}{n}\log S_n &= \frac{1}{n}\log\prod_{k=1}^{n} b(X_k|Y_k)o(X_k)\\
&= \frac{1}{n}\sum_{k=1}^{n}\log\left(b(X_k|Y_k)o(X_k)\right)\\
&\underset{n\to\infty}{\longrightarrow} \sum_{x,y}\Pr\{X=x,Y=y\}\log b(X=x|Y=y)o(X=x) \qquad \text{(per la legge debole dei grandi numeri)}\\
&\overset{\text{def}}{=} W(\mathbf{b}(\cdot|\cdot),P_{X,Y},\mathbf{o})\\
&= \sum_{x,y}P_{X,Y}(xy)\log(b(x|y)o(x))\\
&= \sum_{x,y}P_{X,Y}(xy)\log b(x|y) + \sum_{x,y}P_{X,Y}(xy)\log o(x)\\
&= \sum_{x,y}P_{X,Y}(xy)\log\left(b(x|y)\frac{P_{X|Y}(x|y)}{P_{X|Y}(x|y)}\right) + \sum_{x}P_X(x)\log o(x)\\
&= \sum_{y}P_Y(y)\sum_{x}P_{X|Y}(x|y)\log P_{X|Y}(x|y) - \sum_{y}P_Y(y)\underbrace{\sum_{x}P_{X|Y}(x|y)\log\frac{P_{X|Y}(x|y)}{b(x|y)}}_{\geq 0 \text{ poichè è una divergenza}} + \sum_{i}p_i\log o_i\\
&\leq \sum_{i}p_i\log o_i + \sum_{y}P_Y(y)\sum_{x}P_{X|Y}(x|y)\log P_{X|Y}(x|y)\\
&= \sum_{i}p_i\log o_i - \sum_{y}P_Y(y)\sum_{x}P_{X|Y}(x|y)\log\frac{1}{P_{X|Y}(x|y)}\\
&= \sum_{i}p_i\log o_i - \sum_{y}P_Y(y)H(X|Y=y) = \sum_{i}p_i\log o_i - H(X|Y).
\end{aligned}
$$

In più, notiamo che vale l'uguaglianza se e solo se la divergenza che compare nell'espressione è pari a zero, ovvero se e solo se il giocatore ha usato come strategia $\mathbf{b}(\cdot|\cdot) = P_{X|Y}(\cdot|\cdot)$. Pertanto possiamo dire che

$$
\max_{\mathbf{b}(\cdot|\cdot)} W(\mathbf{b}(\cdot|\cdot),P_{X,Y},\mathbf{o}) = \sum_{i}p_i\log o_i - H(X|Y). \tag{8}
$$

Cerchiamo ora di comprendere di quanto è variato il max appena calcolato nell'espressione (8), rispetto al valore massimo di W calcolato nella lezione scorsa, quando *non* era nota alcuna informazione addizionale Y. Ricordiamo che tale "vecchio" valore massimo era pari a

$$
W = \sum_{i}p_i\log o_i - H(X).
$$

Denotando con Δ la differenza tra nuovo massimo e vecchio massimo otteniamo che

$$
\begin{aligned}
\Delta &= \left(\sum_{i}p_i\log o_i - H(X|Y)\right) - \left(\sum_{i}p_i\log o_i - H(X)\right)\\
&= H(X) - H(X|Y) = I(X;Y)
\end{aligned}
$$

Informalmente, possiamo dire che nella *nuova* situazione in cui disponiamo della conoscenza di una vc causale ausiliaria Y, allora usando la strategia ottima il capitale crescerà con un fattore moltiplicativo pari a

$$
\approx 2^{n[W+I(X;Y)]}, \tag{9}
$$

mentre prima il capitale cresceva con un fattore moltiplicativo pari a

$$\approx 2^{nW}. \tag{10}$$

Di nuovo, il risultato matematico ha una piacevole interpretazione intuitiva: *L'incremento della crescita del capitale, dall'espressione (10) all'espressione (9), dipende esclusivamente da quanta informazione addizionale il giocatore possiede sugli esiti delle gare, ed esso è quantificato esattamente dalla mutua informazione* $I(X;Y)$.