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

Lezione 16

Ugo Vaccaro

La motivazione del problema che vogliamo considerare nella lezione odierna è la seguente. Supponiamo di avere un esperto che ci fornisce delle predizioni, vorremmo stabilire un sistema di "ricompense" in modo tale che esso sia incentivato a fornirci predizioni che siano il più possibili corrette.

Siano $E_1, \ldots, E_n$ eventi mutualmente esclusivi (che, ad esempio, potrebbero corripondere ad esiti di un esperimento, esiti di una corsa di cavalli, eventi di tipo finanziario, di tipo meteorologico, medico ect.), e siano $p_1, \ldots, p_n$ le rispettive probabilitità di verificarsi, ovvero

$$p_i = \text{Probabilità che si verifichi l'evento } E_i,\ i = 1, \ldots, n.$$

Noi non conosciamo $\mathbf{p} = (p_1, \ldots, p_n)$, quindi chiediamo ad un esperto di darci una valutazione di $\mathbf{p}$. Denotiamo con $\mathbf{q} = (q_1, \ldots, q_n)$ la *valutazione* che l'esperto fà di $\mathbf{p}$. Ovviamente, noi vorremmo che $\mathbf{q}$ fosse uguale a $\mathbf{p}$. Ci accordiamo con l'esperto in modo tale che lui sarà pagato ogni volta, in base all'esito che l'esperimento avrà, ed in base alla sua predizione. Stabiliamo di corrispondere all'esperto una somma pari a $f(q_k)$ (dove f è una funzione da *noi* decisa, visto che siamo noi a pagare), se e solo se si verifica l'evento E_k. Di conseguenza, il guadagno medio dell'esperto sarà pari a

$$\sum_{k=1}^{n} p_k f(q_k). \tag{1}$$

Come prima detto, il nostro obiettivo è far sì che l'esperto ci dia una valutazione $\mathbf{q} = (q_1, \ldots, q_n)$ che sia uguale alla vera distribuzione di probabilità $\mathbf{p} = (p_1, \ldots, p_n)$ con cui gli eventi $E_1, \ldots, E_n$ si verificano. Però l'esperto ha un differente obiettivo: *massimizzare il suo guadagno*. Pertanto, potrebbe fornirci una $\mathbf{q} \neq \mathbf{p}$ se ciò gli convenisse maggiormente. A mò di esempio, assumiamo che $\mathbf{p} = (1/2, 1/4, 1/4)$ sia la vera distribuzione di probabilità e che la funzione di pagamento dell'esperto da noi scelta sia $f(x) = x^2$. Se l'esperto ci fornisse $\mathbf{q} = \mathbf{p}$, il suo guadagno medio, dalla (1) sarebbe pari a

$$\frac{1}{2} \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{16} + \frac{1}{4} \times \frac{1}{16} = \frac{5}{32}, \tag{2}$$

mentre se ci fornisse una $\mathbf{q} \neq \mathbf{p}$, ad esempio $\mathbf{q} = (1 - 2\epsilon, \epsilon, \epsilon)$, il suo guadagno medio sarebbe

$$\frac{1}{2} \times (1 - 2\epsilon)^2 + \frac{\epsilon^2}{4} + \frac{\epsilon^2}{4} \underset{\epsilon \to 0}{\longrightarrow} \frac{1}{2} > \frac{5}{32}. \tag{3}$$

Ovvero, l'esperto sarebbe incentivato a darci una predizione sbagliata, pur di massimizzare il suo guadagno! Stesso problema se scegliessimo una funzione decrescente come $f(x) = 1/x$. Sempre con $\mathbf{p} = (1/2, 1/4, 1/4)$, l'esperto guadagnerebbe in media

$$\frac{1/2}{1/2} + \frac{1/4}{1/4} + \frac{1/4}{1/4} = 3, \tag{4}$$

se ci fornisse $\mathbf{q} = \mathbf{p}$, mentre guadagnerebbe in media

$$\frac{1/2}{\epsilon} + \frac{1/4}{\epsilon} + \frac{1/4}{1 - 2\epsilon} \underset{\epsilon \to 0}{\longrightarrow} \infty, \tag{5}$$

se ci fornisse $\mathbf{q} = (\epsilon, \epsilon, 1 - 2\epsilon) \neq \mathbf{p}$.

Di conseguenza, allo scopo di costringere l'esperto ad essere onesto (o quanto meno, ad invogliarlo...), vorremmo trovare una funzione f per cui valga la seguente diseguaglianza:

$$\forall \mathbf{p}, \forall \mathbf{q} \qquad \sum_{k=1}^{n} p_k f(q_k) \leq \sum_{k=1}^{n} p_k f(p_k). \tag{6}$$

con uguaglianza se e solo se $\mathbf{q} = p$. In questo modo, se l'esperto vuole massimizzare il suo guadagno, è forzato a darci una $\mathbf{q} = \mathbf{p}$, ovvero una $\mathbf{q}$ uguale alla vera distribuzione di probabilità con cui gli eventi $E_1, \ldots, E_n$ si verificano.

Ci risulta più comodo risolvere il seguente problema: *determinare funzioni h tali che*

$$\forall \mathbf{p}, \forall \mathbf{q} \qquad \sum_{k=1}^{n} p_k h(p_k) \leq \sum_{k=1}^{n} p_k h(q_k). \tag{7}$$

Infatti, ciò che cerchiamo in (6) non sono altro che funzioni f che sono uguali a $-h$, dove h sono le funzioni che soddisfano (7).

Cerchiamo ora di capire come sono fatte le funzioni per cui la (7) vale. Scegliamo due arbitrarie distribuzioni di probabilità $\mathbf{p} = (p_1, \ldots, p_n)$ e $\mathbf{q} = (q_1, \ldots, q_n)$ tali che

$$p_k = q_k, \quad \text{per } k > 2. \tag{8}$$

Poichè $\sum_{k=1}^{n} p_k = \sum_{k=1}^{n} q_k = 1$, varrà anche che

$$p_1 + p_2 = q_1 + q_2 = r, \qquad \text{per qualche } r \in]0, 1[. \tag{9}$$

Per semplicità di notazione poniamo $p_1 = p, q_1 = q$, da cui $p_2 = r - p$ e $q_2 = r - q$. Dalla (8) otteniamo che $p_k h(p_k) = p_k h(q_k)$, per ogni $k > 2$, di conseguenza, la (7) è equivalente a richiedere che

$$p_1 h(p_1) + p_2 h(p_2) \leq p_1 h(q_1) + p_1 h(q_2) \tag{10}$$

ovvero, ricordando la definizione di p, q, r, otteniamo che

$$\forall p \in]0, 1[, \quad q \in]0, 1[, \quad r \in]0, 1[\qquad ph(p) + (r - p)h(r - p) \leq ph(q) + (r - p)h(r - q), \tag{11}$$

che può essere anche riscritta come

$$\forall p \in]0, 1[, \quad q \in]0, 1[, \quad r \in]0, 1[\qquad p[h(q) - h(p)] \geq (r - p)[h(r - p) - h(r - q)]. \tag{12}$$

Visto che l'intervallo dei valori della p e q è lo stesso nella (12), possiamo anche dire che vale la sua versione simmetrica, ovvero

$$q[h(p) - h(q)] \geq (r - q)[h(r - q) - h(r - p)]. \tag{13}$$

Moltiplichiamo la (12) per $(r - q) \geq 0$ e la (13) per $(r - p) \geq 0$, per ottenere le diseguaglianze

$$p(r - q)[h(q) - h(p)] \geq (r - q)(r - p)[h(r - p) - h(r - q)], \tag{14}$$

$$q(r - p)[h(p) - h(q)] \geq (r - p)(r - q)[h(r - q) - h(r - p)]. \tag{15}$$

Addizioniamo il membro sinistro della prima diseguaglianza con il corrispondente membro sinistro della seconda diseguaglianza, ed analogamente con i membri destri. Semplificando, otterremo che

$$[h(p) - h(q)]r(q - p) \geq 0. \tag{16}$$

Dalla (16), otteniamo quindi che una *qualsiasi* funzione h che soddisfa la (7) è tale che

$$q > p \Rightarrow h(q) \leq h(p), \tag{17}$$

ovvero *ogni* h che soddisfa la (7) è necessariamente monotona non crescente. Dalla (12) otteniamo che

$$h(q) - h(p) \geq \frac{(r - p)}{p}[h(r - p) - h(r - q)] \tag{18}$$

ovvero, sotto l'ipotesi che $q > p$

$$\frac{h(q)-h(p)}{q-p} \geq \frac{(r-p)}{p}\frac{[h(r-p)-h(r-q)]}{q-p} = \frac{(r-p)}{p}\frac{[h(r-p)-h(r-q)]}{(r-p)-(r-q)} \tag{19}$$

Analogamente, dalla (13) otteniamo

$$\frac{h(q)-h(p)}{q-p} \leq \frac{(r-q)}{q}\frac{[h(r-p)-h(r-q)]}{(r-p)-(r-q)} \tag{20}$$

Ovviamente, se $q < p$ le diseguaglianze appena ottenute sono invertite. Mettendo insieme la (19) e (20) otteniamo

$$\frac{(r-p)}{p}\frac{[h(r-p)-h(r-q)]}{(r-p)-(r-q)} \leq \frac{h(q)-h(p)}{q-p} \leq \frac{(r-q)}{q}\frac{[h(r-p)-h(r-q)]}{(r-p)-(r-q)} \tag{21}$$

Se la funzione h è differenziabile in $r-p \in]0,1-p[$, allora

$$\lim_{q\to p}\frac{[h(r-p)-h(r-q)]}{(r-p)-(r-q)} = h'(r-p). \tag{22}$$

Di conseguenza, i due termini esterni della (21) tendono, per $q \to p$, allo stesso limite, pari a

$$\frac{r-p}{p}h'(r-p). \tag{23}$$

Ciò implica che anche il termine interno della (21) tende, per $q \to p$, allo stesso limite (23). Di conseguenza, se h è differenziabile in $r-p \in]0,1-p[$, allora h è differenziabile *anche* in p e vale

$$h'(p) = \frac{r-p}{p}h'(r-p). \tag{24}$$

Ovvero,

$$ph'(p) = (r-p)h'(r-p), \qquad p \in]0,1[, \quad r \in]p,1[. \tag{25}$$

Ripetendo, se h è differenziabile in $(r-p)$, allora h è anche differenziabile in p. Detto in modo equivalente, se h *non fosse* differenziabile in p, allora h non lo sarebbe nemmeno in alcun $r-p \in]0,1-p[$, ovvero h non sarebbe differenziabile nell'intero intervallo $]0,1-p[$. Però noi sappiamo che h è monotona e quindi è necessariamente differenziabile (a meno di un insieme di misura nulla). La raggiunta contraddizione ci permette di dire che h è derivabile in tutto $]0,1-p[$.

Proviamo ora che $\forall p \in]0,1[$ la (25) implica che

$$ph'(p) = \gamma \quad \text{per qualche costante } \gamma. \tag{26}$$

Innanzitutto osserviamo che, se la (25) vale, allora necessariamente $\gamma \leq 0$ in quanto h è non crescente. Supponiamo innanzitutto che $p \in]0,1/2[$. Scegliamo $r = p + 1/2$. Dalla (25) otteniamo

$$\forall p \in]0,1/2[\qquad ph'(p) = \frac{1}{2}h'\Big(\frac{1}{2}\Big) = \gamma \tag{27}$$

Se invece $p \in [1/2,1[$, allora $r-p \in]0,1/2[$. Quindi, usando la (27) e la (25) otteniamo

$$ph'(p) = (r-p)h'(r-p) = \gamma. \tag{28}$$

Mettendo tutto insieme, abbiamo provato che una *qualsiasi* funzione h che soddisfa la (7), necessariamente soddisfa l'equazione differenziale

$$ph'(p) = \gamma \leq 0 \qquad \forall p \in]0,1[. \tag{29}$$

É semplice vedere che la soluzione all'equazione differenziale (29) è del tipo

$$h(p) = a \log_2 p + b, \tag{30}$$

per arbitrarie costanti a, b, ma con $a \leq 0$. Ricordiamo che noi eravamo interessati a funzioni f che soddisfacevano la (6). Per quanto appena mostrato, tali funzioni *necessariamente* devono avere la forma

$$f(p) = c \log_2 p + b, \tag{31}$$

per arbitrarie costanti c, b, ma con $c \geq 0$.

Abbiamo quindi scoperto come pagare l'esperto per far sì che esso, per massimizzare il suo guadagno (cosa che ci aspettiamo voglia fare...) sia costretto a darci la predizione corretta $\mathbf{p}$: *pagarlo usando la funzione f determinata dalla* (31). Calcoliamo adesso il guadagno medio dell'esperto usando la giusta funzione f (ovvero sostituiamo l'espressione della f trovata nella formula (31) nell'espressione (1) che rappresenta il guadagmo medio dell'esperto). Otterremo che tale guadagno medio è pari a

$$\begin{aligned} \sum_{k=1}^{n} p_k (c \log_2 p_k + b) &= b + c \sum_{k=1}^{n} p_k \log_2 p_k \\ &= b - c \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \\ &= b - cH(\mathbf{p}), \end{aligned} \tag{32}$$

dove $H(\mathbf{p})$ è (la solita) entropia di $\mathbf{p} = (p_1, \ldots, p_n)$. In conclusione, per avere la predizione corretta occorre pagare l'esperto una somma fissata b cui occorre sottrarre una quantità proporzionale al nostro livello di incertezza medio dopo aver ricevuto ogni predizione. Il fatto da sottolineare è che la funzione f nella formula (31) è *l'unica funzione* che ci garantisce che l'esperto sia incentivato a produrre predizioni corrette. La formula (32) ci porta ad un corrispondente risultato di "unicità" della funzione entropia!

Un altro modo equivalente per dire la stessa cosa, è il seguente. Se siamo alla ricerca di "misure di distanza" M tra arbitrarie distribuzioni di probabilità $\mathbf{p}$ e $\mathbf{q}$, con M della forma

$$M(\mathbf{p}, \mathbf{q}) = \sum_i p_i [h(p_i) - h(q_i)],$$

per *qualche* funzione h, che soddisfano la proprietà che $M(\mathbf{p}, \mathbf{q}) \geq 0$, e con uguaglianza a zero se e solo se $\mathbf{p} = \mathbf{q}$, allora l'*unica* M possibile è la divergenza informazionale!.

In realtà, il risultato è ancora più importante di ciò che sembra. Infatti, in Statistica (e nella moderna area di Apprendimento Automatico, alla base dei recenti risultati dell'Intelligenza Artificiale) si usano molto spesso delle misure per valutare la "penalità" in cui si incorre quando si afferma la probabilità che un evento x occorra sia pari a $P(x)$. É molto importante minimizzare la "penalità" media in cui si incorre quando si stima la forma di una distribuzione di probabilità. Tutto ciò che abbiamo prima provato (nella formulazione data dalla (7)), si può informalmente tradurre nel seguente modo: a meno di costanti additive e moltiplicative (che hanno poca importanza in pratica) l'*unica misura* di penalità per cui la penalità media è minimizzata quando si produce la *corretta* distribuzione di probabilità è la misura che assegna penalità $-\log P(x)$ ogni qualvolta si sostiene che il generico evento x occorra con probabilità $P(x)$.