Giornale di Tecniche Nefrologiche & Dialitiche 2013; 25 (2): 150-152
DOI: 10.5301/GTND.2013.11227

Una digressione su probabilità, informatica e medicina

Michele Nichelatti, Maurizio Nordio, Umberto Maggiore, Maurizio Postorino, Aurelio Limido

a nome del Comitato Scientifico SIN-RIDT

A digression on probability, computer science and medicine
Abstract. The relationship between the probability of realization of an event and the information obtained from the evnt itself is presented.
The discussed example refers to codon and amino acids.

Key words: Probability, Computer science, Codon, Amino acids

Conflict of interest: None.

Ricevuto: 29 Aprile 2013; Accettato: 2 Maggio 2013

Ordine e disordine

Durante il nostro normale percorso di studi, tutti abbiamo avuto a che fare con il concetto di *entropia* (parola derivante dai termini del greco antico ἐν e τροπή, che possono essere tradotti come *trasformazione interna*), che, in termini pratici, significa *disordine*. Sappiamo, quindi, che tutte le cose tendono a evolvere verso lo stato di massimo disordine, che è lo stato di massima probabilità, a meno che non venga fatto del lavoro (lavoro in senso fisico, ovviamente) capace di compensare il disordine che si viene a creare e di diminuire localmente (cioè, in una zona ben delimitata dello spazio) l'entropia.
Sappiamo che gli esseri viventi sono costretti a svolgere del lavoro che consenta loro di mantenere l'ordine delle strutture (cellule, tessuti, organi), che, altrimenti, sarebbero rapidamente distrutte e disorganizzate dall'aumentare dell'entropia al loro interno. Per fare questo, al di là del lavoro fisico, è necessario che gli organismi viventi siano dotati di strutture particolari, capaci di mantenere l'ordine dei sistemi e di immagazzinare tutte le informazioni in grado di aiutare queste funzioni.
Si rende, quindi, necessario stabilire una misura che sia in grado di determinare lo stato di ordine presente in un sistema, esattamente come l'entropia ne misura il disordine: per fare questo è necessario, e forse anche sorprendente, ricorrere alla probabilità. Se l'entropia evolve verso lo stato di massima probabilità, è possibile ipotizzare che un sistema che si trovi in uno stato di bassa probabilità sia un sistema anche a bassa entropia e che quanto più bassa sia la probabilità dello stato di un sistema, tanto maggiore possa essere l'informazione che si può ricavare dall'analisi di quel sistema.

L'informazione e la probabilità

Assumiamo di estrarre a caso 1 pallina da un'urna che ne contiene 32: ogni pallina avrà una probabilità pari a 1/32 (cioè pari a circa il 3.1%) di essere estratta. Assumiamo ora che 1 pallina sia differente dalle altre, per esempio per il colore, e, quindi, che 31 palline siano bianche (e indistinguibili) e una sia nera: in questa situazione la probabilità di estrarre una qualsiasi pallina bianca e pari al 96.9%, mentre la probabilità di estrarre l'unica pallina nera rimane del 3.1%.
Ora immaginiamo di dividere le palline tra due urne differenti e di sapere (o meglio, di essere informati) in quale delle due urne sarà presente la pallina nera: in pratica, avremo due urne ciascuna contenente 16 palline, e sapremo in quale delle due urne sarà contenuta la pallina nera, quella che ci "interessa". Tenendo solo l'urna con la pallina nera e scartando quella che contiene solo palline bianche, ripetiamo ora l'operazione dividendo le 16 palline fra due urne da 8 ciascuna, e assumiamo anche questa volta di essere a conoscenza della posizione della pallina nera. Se continuiamo a effettuare questa operazione dividendo le 8 palline residue tra due urne da 4 e poi, ancora, tra due urne da 2 palline e due urne da 1 pallina ciascuna, alla fine arriveremo ad avere la pallina nera da sola nell'urna.
Abbiamo, perciò, diviso in due il numero delle palline per cinque volte consecutive, tenendo solo le palline contenute nell'urna che sapevamo contenere anche la pallina nera; tutto l'insieme di queste operazioni lo possiamo esprimere scrivendo $2^5 = 32$ oppure (il che è lo stesso) $\log_2 32 = 5$. Da quest'ultima espressione, si vede che 5 è il logaritmo in base 2 di 32, ma cos'è di preciso un logaritmo? Fondamentalmen-

te, è uno strumento matematico *inventato*[I] per lavorare con i numeri molto grandi e molto piccoli e che consente di trasformare una moltiplicazione in un'addizione e una divisione in una sottrazione, semplificando, quindi, molte situazioni. Con l'avvento dei calcolatori elettronici, l'importanza dei logaritmi, da questo punto di vista, non è affatto diminuita; anzi, possiamo affermare che, se non ci fossero i logaritmi, probabilmente non esisterebbero neppure i computer.
L'operazione più semplice compiuta nel corso delle cinque suddivisioni consecutive, è sicuramente l'ultima, ovvero quella che ha suddiviso le 2 palline rimanenti in due urne contenenti 1 pallina ciascuna, rispettivamente la nera e la bianca. Possiamo, quindi, ricavare una misura dell'informazione che ci è arrivata dal sistema costituito da 2 palline: questa informazione, equivalente alla più elementare informazione possibile, la definiamo pari a 1 bit (acronimo dall'inglese *binary digit*). L'informazione di 1 bit è quella che si ricava da un evento che ha la probabilità del 50% di realizzarsi (estrarre la pallina nera da un'urna che contiene solo 2 palline), quindi possiamo scrivere:

$$1 \text{ bit} = \log_2\left(\frac{1}{0.5}\right) = \log_2 2,$$

dove si nota che il logaritmo deve necessariamente avere base binaria (ovvero la base deve essere 2), in quanto $2^1 = 2$.
A questo punto, è possibile ricavare l'informazione che si ottiene dal verificarsi di un evento con una probabilità qualsiasi: estrarre la pallina nera da un'urna che contiene 4 palline ha una probabilità del 25%, quindi dato che $\log_2 (1/0.25) = \log_2 4$ e che $4 = 2^2$, l'informazione che si ottiene è pari a 2 bit; estrarre la pallina nera da un'urna che contiene 8 palline ha una probabilità del 12.5%, pertanto, essendo $1/0.125 = 8 = 2^3$, si ricava un'informazione pari a 3 bit.
Si potrebbe continuare in questo modo per qualsiasi valore della probabilità: basta ricordare la relazione generale, che dice:

$$\text{informazione (in bit)} = \log_2\left(\frac{1}{\text{probabilità}}\right)$$

Pertanto, più un evento è raro, maggiore è l'informazione che si ricava al verificarsi dell'evento; se vogliamo studiare l'informazione che si ricava dal verificarsi di due eventi indipendenti, dobbiamo ricordare che la probabilità di due eventi indipendenti è data dal prodotto delle due singole probabilità e che quindi, per esempio, pescare un asso di coppe da un mazzo di 40 carte e una qualsiasi carta che abbia per seme bastoni da un altro mazzo di 40 carte ha una probabilità pari al prodotto delle due probabilità elementari che la compongono. In questo caso, dato che pescare l'asso di coppe ha una probabilità pari a $1/40 = 0.025$ e che pescare una qualsiasi carta di bastoni ha una probabilità pari a $10/40 = 0.25$, la probabilità che si verifichino ambedue gli eventi diventa pari a $0.025 \times 0.25 = 0.00625 \approx 0.6\%$. Il verificarsi di un evento con bassa probabilità ci spinge immediatamente a ritenere che sotto ci sia qualcosa di anomalo, per esempio un trucco o una manipolazione: è quello che penseremmo istintivamente se qualcuno ci dicesse di essere in grado di estrarre l'asso di coppe da un mazzo di 40 carte al primo colpo, ritenendo la cosa altamente improbabile (dato che la probabilità è del 2.5%) e, quindi, nel caso l'evento si realizzasse, penseremmo di avere a che fare con un baro o un prestigiatore (e, in effetti, un prestigiatore usa dei trucchi).
Ma qual è l'informazione che si ricava dal verificarsi di due eventi indipendenti come quelli appena descritti? Se chiamiamo p_1 la probabilità del primo evento, I_1 l'informazione che ne deriva, p_2 la probabilità del secondo evento e I_2 l'informazione che ne deriva, ricordando le principali proprietà dei logaritmi (vedi l'appendice), l'informazione complessiva I diventa:

$$I = \log_2\left(\frac{1}{p_1 p_2}\right) = \log_2\left(\frac{1}{p_1} \times \frac{1}{p_2}\right) = \log_2\left(\frac{1}{p_1}\right) + \log_2\left(\frac{1}{p_2}\right) = I_1 + I_2.$$

In altre parole, le probabilità degli eventi indipendenti si moltiplicano, mentre le informazioni che se ne ricavano vanno sommate. Nel nostro esempio, dato che $I_1 = \log_2 (1/0.025) = \log_2 40 \approx 5.322$ (e, infatti, $2^{5.322} \approx 40$) e che $I_2 = \log_2 (1/0.025) = \log_2 4 = 2$ (essendo $2^2 = 4$), l'informazione complessiva che si ricava è pari a $I = I1+12 \approx 5.322 + 2 = 7.322$ bit[II].

L'informazione negli esseri viventi

Sappiamo tutti che l'informazione negli esseri viventi si trasmette di generazione in generazione tramite il DNA. Il DNA è composto da una sequenza predefinita di basi nucleotidiche per ciascuna proteina codificata dall'organismo; le basi nucleotidiche sono quattro (adenina, guanina, citosina e timina), mentre le proteine sono formate da sequenze predefinite di aminoacidi ed è noto che gli aminoacidi utilizzati per la sintesi proteica sono 20.
Ovviamente, l'uso di una sola base nucleotidica per volta non sarebbe in grado di individuare in modo univoco un determinato aminoacido, perché un alfabeto di quattro sole lettere, come è l'insieme delle quattro basi nucleotidiche, e che avesse parole lunghe soltanto una lettera potrebbe produrre solamente quattro parole. Se tentassimo di individuare in modo univoco uno dei 20 aminoacidi usando una sequenza di due basi nucleotidiche, cioè se tentassimo di usare un vocabolario costituito solo da parole di due lettere con un alfabeto di quattro lettere, potremmo costruire al massimo $4^2 = 16$ parole, ancora insufficienti per individuare in modo univoco un aminoacido dei 20 necessari. Per individuare un aminoacido in modo univoco bisogna, quindi, ricorrere a una sequenza di tre basi nucleotidiche (chiamata *codon*), con le quali è possibile comporre $4^3 = 64$ parole, largamente sufficienti per codificare i 20 aminoacidi. Un codon contiene un'informazione pari a 6 bit, in quanto $2^6 = 64$; d'altra parte, dato che ogni base nucleotidica contiene 2 bit, in quanto $2^2 = 4$, per quanto visto precedentemente circa la quantità di informazione fornita da eventi indipendenti, tre basi nucleotidiche in sequenza contengono $2 + 2 + 2 = 6$ bit.

[I]Si preferisce usare il termine "inventato", anziché "scoperto": infatti, se si dà retta al matematico tedesco Leopold Kronecker (1823-1891), che una volta ha detto "*Dio ha creato i numeri interi, tutto il resto è opera dell'uomo*", se ne ricava che la matematica è fatta più di invenzioni che di scoperte.

[II]Bisogna stare attenti a non confondere il *bit* con il *Byte*: un *Byte* è, infatti, un'unità di misura dell'informazione equivalente a 8 *bit*.

Come visto prima, gli aminoacidi sono 20, quindi un singolo aminoacido trasporta con sé un'informazione pari a circa 4.322 bit, in quanto $2^{4.322} \approx 20$: ma, allora, perché sono necessari 6 bit per un *codon*, quando per un aminoacido ne bastano poco più di quattro? E che fine ha fatto l'informazione in eccesso, dato che ogni volta che un codon viene tradotto in aminoacido ne viene persa una quota pari a circa 1.7 bit? La risposta sta nella necessità di fornire un'informazione ridondante (quindi in eccesso) per bilanciare gli effetti dell'entropia, che tenderebbero a disorganizzare la trasmissione delle informazioni dal DNA alle proteine.

Quesito

Abbiamo appena visto che $\log_2 20 = 4.322$, mentre, in precedenza, avevamo visto che $\log_2 40 = 5.322$. In pratica, $\log_2 40 - \log_2 20 = 1$. È solo una coincidenza oppure c'è un motivo?

Appendice: Brevissime note sui logaritmi

Se $b^a = c$, allora possiamo scrivere anche $a = \log_b c$, ovvero a è l'esponente a cui elevare la base b per ottenere il numero c; per esempio, dato che $10^2 = 100$, allora $2 = \log_{10} 100$.
La base più comune dei logaritmi è il numero e = 2.71828…, chiamato anche *numero di Nepero* o anche *base dei logaritmi naturali*; è un numero irrazionale, che, quindi, non può essere ottenuto da alcuna frazione *a/b*, con *a* e *b* interi. Sono molto usate anche la base binaria (2) e, specialmente in medicina e biologia, la base decimale (10).
Le principali proprietà dei logaritmi derivano in modo immediato dalla loro definizione; in particolare:

- Il logaritmo di un numero reale, in qualsiasi base, è definito solo se il numero è maggiore di zero, per cui il logaritmo di un numero negativo non esiste
- La base di un logaritmo può essere qualsiasi numero reale *n*, purché $n > 0$ e $n \neq 1$
- In qualsiasi base, risulta sempre $\log_a 0 = -\infty$, $\log_a 1 = 0$ e $\log_a a = 1$
- Il logaritmo di un prodotto è uguale alla somma dei logaritmi: $\log_a (mn) = \log_a m + \log_a n$
- Il logaritmo di una frazione è uguale alla differenza dei logaritmi:

$$\log_a \left(\frac{m}{n}\right) = \log_a m - \log_a n$$

- Dato il logaritmo in base *a* di un qualsiasi numero *n* (maggiore di zero), è possibile trasformarlo nel logaritmo dello stesso *n*, con altra base *b*, in quanto

$$\log_b n = \frac{\log_a n}{\log_a b}$$

Riassunto

Si presentano le relazioni esistenti tra la probabilità di un evento e l'informazione che si ricava dal verificarsi dell'evento stesso.
L'esempio discusso fa riferimento a codon ed amino acidi.

Parole chiave: Probabilità, Informatica, Codon, Aminoacidi

Dichiarazione di conflitto di interesse: Gli Autori dichiarano di non avere conflitto di interessi.

Indirizzo degli Autori:
Dr. Michele Nichelatti
Servizio di Biostatistica
Dipartimento di Ematologia e Oncologia
Ospedale Niguarda Ca' Granda
Piazza Ospedale Maggiore 3
20162 Milano
NICHELATTI@OspedaleNiguarda.it