Azzolini Riccardo 2020-03-17

Probabilità condizionata e indipendenza

1 Probabilità condizionata

Spesso, nel calcolo della probabilità di un evento aleatorio, si ha una condizione che riduce lo spazio campionario. Un esempio di tale situazione è illustrato nel seguente problema.

Problema: Si giocano alla roulette i numeri 3, 13 e 22.

$$A = \text{“esce il numero 3, 13 o 22”} = \{3, 13, 22\} \quad \# A = 3$$

Poiché i possibili risultati sono 37,

$$\Omega = \{0, 1, 2, \dots, 36\} \quad \# \Omega = 37$$

e si può supporre che la probabilità sia uniforme, la probabilità di vincere (ovvero che esca uno dei numeri giocati) è $P(A) = \frac{3}{37}$. Si viene però a sapere che la roulette è truccata, in modo che possano uscire solo numeri dispari. In altre parole, si verificherà certamente l'evento

$$B = \{1, 3, \dots, 35\} \quad \# B = 18$$

(che può essere considerato come un nuovo spazio campionario, “ridotto” dalla conoscenza che la roulette sia truccata). Qual è ora la probabilità di vincere?

La probabilità che si verifichi l'evento A, sapendo che B si verifica certamente, è chiamata **probabilità condizionata** (o **condizionale**) di A rispetto a B, ed è indicata con la notazione $P(A \mid B)$. Tale notazione, però, non rappresenta la probabilità di un nuovo tipo di evento, $A \mid B$, bensì corrisponde a una nuova mappa di probabilità, $P(\_ \mid B)$, dove “al posto” di _ può essere inserito qualunque evento (dello spazio di probabilità considerato). Perciò, il modo di calcolare tale probabilità deve rispettare le proprietà richieste dalla definizione di mappa di probabilità:

1. $P(\Omega \mid B) = 1$
2. se $\{A_n\}$ è una successione di eventi $A_n \in \mathscr{A}$ disgiunti ($A_i \cap A_j = \varnothing \ \forall i \neq l$), allora

$$P\left(\bigcup_{n=1}^{\infty} A_n \,\middle|\, B\right) = \sum_{n=1}^{\infty} P(A_n \mid B)$$

Un primo tentativo di calcolo potrebbe essere

$$P(A \mid B) = P(A \cap B)$$

cioè, in questo caso, considerare semplicemente la probabilità dell'evento

$$A \cap B = \{3, 13\}$$

Questa soluzione, però, *non rispetta* la proprietà 1:

$$P(\Omega \mid B) = P(\Omega \cap B) = P(B) \neq 1$$

Il modo più semplice per riportare a 1 tale risultato è dividerlo per $P(B)$, ottenendo così la definizione[1]

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

che soddisfa la proprietà 1,

$$P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

e si può dimostrare che verifica anche la 2, quindi è la soluzione corretta.

Tornando al problema, la definizione appena individuata può essere usata per calcolare la probabilità di vincita, cioè dell'evento A, sulla roulette truccata:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{3, 13\})}{P(\{1, 3, \dots, 35\})} = \frac{\frac{2}{37}}{\frac{18}{37}} = \frac{2}{18} = \frac{1}{9}$$

2 Probabilità totale

Oltre a permettere di modellare i casi in cui una nuova informazione riduce lo spazio campionario, la probabilità condizionata è utile anche per effettuare calcoli sull'intero spazio, se alcuni dei dati del problema sono probabilità condizionate.

Problema: Una popolazione è composta per il 40 % da fumatori, e per il restante 60 % da non fumatori. Si sa che il 25 % dei fumatori e il 7 % dei non fumatori sono affetti da una malattia respiratoria cronica. Qual è la probabilità che un individuo scelto a caso sia affetto dalla malattia?

[1] Per poter applicare questa definizione, è necessario avere $P(B) > 0$. Comunque, a livello intuitivo, non avrebbe senso parlare di probabilità condizionata rispetto a B se $P(B) = 0$, perché in tal caso B, l'evento “dato per certo”, sarebbe uno che invece non si verifica mai.

Nella descrizione del problema si individuano i seguenti eventi:

$$\begin{aligned} F &= \text{“l'individuo scelto è fumatore”} \\ N &= \text{“l'individuo scelto è non fumatore”} \\ M &= \text{“l'individuo scelto è affetto dalla malattia”} \end{aligned}$$

Le probabilità date sono

$$P(F) = 0.4 \quad P(N) = 0.6 \qquad \begin{aligned} P(M \mid F) &= 0.25 \\ P(M \mid N) &= 0.07 \end{aligned}$$

ed è richiesto di calcolare $P(M)$.

Osservando che gli eventi F e N costituiscono una *partizione* di Ω, ovvero che

$$\begin{aligned} F \cup N &= \Omega \\ F \cap N &= \varnothing \end{aligned}$$

si possono sfruttare le proprietà delle operazioni insiemistiche e della probabilità per ricondurre il calcolo di $P(M)$ a quello di $P(M \cap F)$ e $P(M \cap N)$:

$$\begin{aligned} P(M) &= P(M \cap \Omega) \\ &= P(M \cap (F \cup N)) \\ &= P((M \cap F) \cup (M \cap N)) \\ &= P(M \cap F) + P(M \cap N) \qquad && (\text{perché } F \cap N = \varnothing \\ & && \implies (M \cap F) \cap (M \cap N) = \varnothing) \end{aligned}$$

Queste due probabilità possono essere ricavate dalle probabilità condizionate, mediante una formula inversa:

$$\begin{aligned} P(M \cap F) &= \frac{P(M \cap F)}{P(F)} \cdot P(F) = P(M \mid F) \cdot P(F) = 0.25 \cdot 0.4 = 0.1 \\ P(M \cap N) &= P(M \mid N) \cdot P(N) = 0.07 \cdot 0.6 = 0.042 \end{aligned}$$

Allora, il calcolo complessivo è:

$$\begin{aligned} P(M) &= P(M \cap F) + P(M \cap N) \\ &= P(M \mid F)\, P(F) + P(M \mid N)\, P(N) \\ &= 0.1 + 0.042 \\ &= 0.142 = 14.2\ \% \end{aligned}$$

Il principio che è stato sfruttato in questo calcolo è chiamato **probabilità totale**.

3 Formula di Bayes

Riprendendo il problema precedente, si vuole adesso calcolare la probabilità che un individuo affetto dalla malattia sia un fumatore, cioè $P(F \mid M)$.

Anche questo risultato si ricava attraverso una manipolazione della formula della probabilità condizionata:

$$\begin{aligned} P(F \mid M) &= \frac{P(F \cap M)}{P(M)} \\ &= \frac{P(F \cap M)}{P(M)} \cdot \frac{P(F)}{P(F)} \\ &= \frac{P(M \cap F)}{P(F)} \cdot \frac{P(F)}{P(M)} \\ &= \frac{P(M \mid F)\, P(F)}{P(M)} \\ &= \frac{0.25 \cdot 0.4}{0.142} \\ &\approx 0.70 = 70\ \% \end{aligned}$$

Il calcolo effettuato in questo esempio può essere generalizzato, ottenendo così il teorema riportato in seguito.

Teorema: Siano $A_1, A_2, \dots, A_n \in \mathscr{A}$ degli eventi che costituiscono una partizione di Ω, ovvero disgiunti e tali che

$$A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$$

e sia $B \in \mathscr{A}$ un altro evento qualsiasi. Vale allora la **formula di Bayes**:

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{P(B)} = \frac{P(B \mid A_i)\, P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)\, P(A_k)}$$

Nota: Il denominatore di questa formula calcola $P(B)$ secondo la regola della probabilità totale:

$$P(B) = \sum_{k=1}^{n} P(B \cap A_k) = \sum_{k=1}^{n} P(B \mid A_k)\, P(A_k)$$

4 Indipendenza

Due eventi A e B si dicono **indipendenti** se $P(A \mid B) = P(A)$, cioè, intuitivamente, se il verificarsi di B non dà informazioni che alterino la probabilità di A. Se e solo se A e B sono indipendenti, vale la formula

$$P(A \cap B) = P(A)\, P(B)$$

perché:

$$P(A \cap B) = \underbrace{P(A \mid B)}_{=P(A) \text{ per ipotesi}} P(B) = P(A)\, P(B)$$

4.1 Esempio

Problema: Il lancio di una moneta dà testa con probabilità p, $0 \le p \le 1$, e croce con probabilità $1 - p$. La moneta viene lanciata n volte. Qual è la probabilità di ottenere una sequenza fissata di teste e croci?

Lo spazio campionario è costituito da tutte le n-uple di valori 1 (che indicano testa) e 0 (che indicano croce),

$$\Omega = \{(\omega_1, \dots, \omega_n) \mid \omega_i \in \{0, 1\}\ \forall i\}$$

e ha cardinalità $\#\,\Omega = 2^n$.

Nel caso di $p = \frac{1}{2} = 1 - p$, si ha probabilità uniforme: tutte le possibili sequenze di risultati diventano appunto equiprobabili, quindi la probabilità di ciascuna di esse è

$$\frac{1}{\#\,\Omega} = \frac{1}{2^n} = \left(\frac{1}{2}\right)^n$$

Invece, per calcolare la probabilità nel caso $p \ne \frac{1}{2}$, conviene considerare inizialmente solo le n-uple $\omega \in \Omega$ della forma

$$\omega = (\underbrace{1, 1, \dots, 1}_{n_1}, \underbrace{0, 0, \dots, 0}_{n_0 = n - n_1})$$

Indicando con le notazione $\omega_i = 1$ e $\omega_i = 0$ gli eventi "l'i-esimo risultato è testa" e "l'i-esimo risultato è croce", rispettivamente, che fissano solo uno degli n risultati, lasciando variare tutti gli altri,

$$\begin{aligned}
(\omega_i = 1) &= \{(\omega_1, \dots, \omega_n) \mid \omega_i = 1,\ \omega_j \in \{0, 1\}\ \forall j \ne i\} \\
(\omega_i = 0) &= \{(\omega_1, \dots, \omega_n) \mid \omega_i = 0,\ \omega_j \in \{0, 1\}\ \forall j \ne i\}
\end{aligned}$$

l'evento $\{\omega\}$, cioè il verificarsi della sequenza ω, può essere scritto come l'intersezione del verificarsi dei singoli risultati che la compongono:

$$\{\omega\} = \underbrace{(\omega_1 = 1) \cap (\omega_2 = 1) \cap \dots \cap (\omega_{n_1} = 1)}_{n_1} \cap \underbrace{(\omega_{n_1+1} = 0) \cap (\omega_{n_1+2} = 0) \cap \dots \cap (\omega_n = 0)}_{n - n_1}$$

Per ipotesi, le probabilità dei singoli eventi di quest'intersezione sono

$$\left.\begin{aligned}
P(\omega_i = 1) &= p \\
P(\omega_i = 0) &= 1 - p
\end{aligned}\right. \quad \forall i$$

Poi, siccome la conoscenza del risultato di un lancio non dà alcuna informazione utile a prevedere gli altri, tali eventi sono indipendenti, e allora la probabilità è complessivamente:

$$\begin{aligned} P(\{\omega\}) &= \underbrace{P(\omega_1 = 1)\, P(\omega_2 = 1) \cdots P(\omega_{n_1} = 1)}_{n_1} \underbrace{P(\omega_{n_1+1} = 0)\, P(\omega_{n_1+2} = 0) \cdots P(\omega_n = 0)}_{n-n_1} \\ &= \underbrace{p\, p \cdots p}_{n_1} \underbrace{(1-p)(1-p) \cdots (1-p)}_{n-n_1} \\ &= p^{n_1} (1-p)^{n-n_1} \end{aligned}$$

Infine, poiché la moltiplicazione è commutativa, questo risultato dipende solo dal numero di teste e croci presenti nella sequenza dei risultati, e non dalle loro posizioni, quindi è valido per qualsiasi n-upla $\omega \in \Omega$,

$$P(\{\omega = (\omega_1, \dots, \omega_n)\}) = p^{\#\{i \mid \omega_i = 1\}} (1-p)^{\#\{i \mid \omega_i = 0\}}$$

e il problema è risolto.

Le situazioni come questa, nelle quali si ha una successione di esperimenti casuali tra loro indipendenti, ciascuno dei quali ha due possibili risultati, convenzionalmente chiamati successo (1) e insuccesso (0), sono dette *schemi successo-insuccesso* o *schemi di Bernoulli*.

4.2 Definizioni più precise

Gli eventi $A_1, \dots, A_n \in \mathscr{A}$ sono **a due a due indipendenti** se e solo se

$$P(A_i \cap A_j) = P(A_i)\, P(A_j) \quad \forall i \neq j$$

Inoltre, essi formano una **famiglia di eventi indipendenti** $\{A_1, \dots, A_n\}$ se e solo se, per ogni $k \le n$ e per ogni scelta di k indici tutti diversi $i_1, \dots, i_k \in \{1, \dots, n\}$, si ha che

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

(ovvero, se la probabilità dell'intersezione è uguale al prodotto delle probabilità dei singoli eventi *per ogni sottoinsieme* di eventi della famiglia).

Non è detto che degli eventi a due a due indipendenti formino anche una famiglia di eventi indipendenti. Ad esempio, considerando lo spazio campionario $\Omega = \{1, 2, 3, 4\}$, e considerando la probabilità uniforme, gli eventi

$$A_1 = \{1, 4\} \quad A_2 = \{2, 4\} \quad A_3 = \{3, 4\}$$
$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$$

sono a due a due indipendenti,

$$\begin{aligned} A_1 \cap A_2 &= \{4\} & P(A_1 \cap A_2) &= \frac{1}{4} = P(A_1)\,P(A_2) \\ A_1 \cap A_3 &= \{4\} & P(A_1 \cap A_3) &= \frac{1}{4} = P(A_1)\,P(A_3) \\ A_2 \cap A_3 &= \{4\} & P(A_2 \cap A_3) &= \frac{1}{4} = P(A_2)\,P(A_3) \end{aligned}$$

ma non formano invece una famiglia di eventi indipendenti:

$$A_1 \cap A_2 \cap A_3 = \{4\} \qquad P(A_1 \cap A_2 \cap A_3) = \frac{1}{4} \neq P(A_1)\,P(A_2)\,P(A_3) = \frac{1}{8}$$

5 Problema: celle di memoria

Problema: Un compilatore assegna a ognuna delle variabili di un programma una cella di memoria, scelta a caso e in modo indipendente dalle scelte precedenti. In caso di conflitto, cioè se due variabili vengono assegnate alla stessa cella di memoria, la procedura di assegnazione deve essere ripetuta.

Se ci sono n celle e k variabili, qual è in generale la probabilità che si verifichi un conflitto? In particolare, quanto vale la probabilità per $n = 1000$ e $k = 25$? Come si può valutare questa procedura in base a tali risultati?

5.1 Soluzione considerando l'evento complementare

Il modo più semplice e conveniente per risolvere questo problema è trattarlo analogamente al problema dei compleanni (presentato in una lezione precedente), calcolando la probabilità mediante lo studio dell'evento complementare.

Lo spazio campionario è dato da tutte le possibili assegnazioni, anche con ripetizioni, delle n celle di memoria alle k variabili:

$$\Omega = \{(\omega_1, \dots, \omega_k) \mid \omega_i \in \{1, \dots, n\}\} \qquad \#\,\Omega = n^k$$

L'evento considerato è

$$\begin{aligned} A &= \text{“si verifica un conflitto”} \\ &= \{(\omega_1, \dots, \omega_k) \in \Omega \mid \exists (i, j),\ i \neq j,\ \omega_i = \omega_j\} \end{aligned}$$

i cui elementi sono difficili da enumerare, quindi si tratta invece l'evento complementare,

$$\begin{aligned} A^{\mathsf{c}} &= \text{“non si verifica un conflitto”} \\ &= \text{“ogni variabile è assegnata a una cella diversa”} \\ &= \{(\omega_1, \dots, \omega_k) \in \Omega \mid \omega_i \neq \omega_j \ \forall i \neq j\} \end{aligned}$$

che corrisponde alle disposizioni semplici di k oggetti scelti da un insieme di n:

$$\# A^c = D_{n,k} = \frac{n!}{(n-k)!}$$

Allora, la probabilità di *non* avere un conflitto è

$$P(A^c) = \frac{\# A^c}{\# \Omega} = \frac{n!}{(n-k)!\, n^k} = \frac{(n-1)!}{(n-k)!\, n^{k-1}}$$

e perciò quella di avere invece un conflitto è:

$$P(A) = 1 - P(A^c) = 1 - \frac{(n-1)!}{(n-k)!\, n^{k-1}}$$

Per $n = 1000$ e $k = 25$, la probabilità di un conflitto è

$$P(A) = 1 - \frac{999!}{975!\, 1000^{24}} \approx 1 - 0.74 = 0.26 = 26\ \%$$

cioè piuttosto elevata, quindi si può concludere che questa procedura di assegnazione delle celle di memoria alle variabili sia alquanto scadente.

5.2 Soluzione con la probabilità condizionata

Un metodo alternativo per risolvere questo problema, anche se più complicato (e quindi mostrato solo a scopo illustrativo), consiste nell'uso della probabilità condizionata per valutare la probabilità di conflitto a ogni assegnazione.

Per prima cosa, è necessario definire gli eventi

$$A_i = \text{“non si verificano conflitti nelle prime } i \text{ assegnazioni”}$$

per $i = 1, \dots, k$. La procedura si conclude senza conflitti se si verifica l'evento A_k, e ciò comporta anche il verificarsi di tutti gli A_i precedenti ($i = 1, \dots, k-1$), perché

$$A_k \subset A_{k-1} \subset \cdots \subset A_2 \subset A_1$$

(intuitivamente, se non si hanno conflitti, ad esempio, nelle prime 3 assegnazioni, cioè se si verifica A_3, allora non si sono avuti conflitti neanche nelle prime 2, né nella prima, ovvero si devono essere verificati anche A_2 e A_1).

Risulta conveniente ricavare la probabilità di ciascun A_i in forma condizionata rispetta ad A_{i-1}, in modo da potersi “concentrare” solo sull'i-esima assegnazione. Infatti, se le celle di memoria sono state assegnate alle prime $i-1$ variabili senza conflitti, il numero di celle “occupate” è appunto $i-1$, mentre ne rimangono “libere” $n-(i-1) = n-i+1$.

La probabilità di scegliere a caso una cella libera anche per la i-esima variabile, evitando un conflitto, è quindi:

$$P(A_i \mid A_{i-1}) = \frac{n-i+1}{n}$$

Inoltre, come caso particolare, si osserva che $P(A_1) = 1$: la prima assegnazione riesce sempre senza conflitti (non essendoci inizialmente celle già assegnate ad altre variabili).

Complessivamente, la probabilità che non ci siano conflitti è

$$\begin{aligned}
P(A_k) &= P(A_k \mid A_{k-1})\,P(A_{k-1}) \\
&= P(A_k \mid A_{k-1})\,P(A_{k-1} \mid A_{k-2})\,P(A_{k-2}) \\
&\;\vdots \\
&= P(A_k \mid A_{k-1})\,P(A_{k-1} \mid A_{k-2}) \cdots P(A_2 \mid A_1)\,P(A_1) \\
&= \frac{n-k+1}{n} \cdot \frac{n-(k-1)+1}{n} \cdots \frac{n-2+1}{n} \cdot 1 \\
&= \frac{n-k+1}{n} \cdot \frac{n-k+2}{n} \cdots \frac{n-1}{n} \cdots 1 \\
&= \frac{(n-k+1)(n-k+2)\cdots(n-1)}{\underbrace{n\,n\cdots n}_{k-1 \text{ volte}}} \\
&= \frac{(n-1)!}{(n-k)!\,n^{k-1}}
\end{aligned}$$

e, di conseguenza, la probabilità che invece ci sia un conflitto, richiesta dal problema, è

$$P(A_k^{\mathsf{c}}) = 1 - P(A_k) = 1 - \frac{(n-1)!}{(n-k)!\,n^{k-1}}$$

che è la stessa formula ottenuta nella soluzione precedente.