Modelli descrittivi, statistica e simulazione

Master per Smart Logistics specialist

Roberto Cordone
(roberto.cordone@unimi.it)

Probabilità condizionata

Spesso succede di avere informazioni parziali sul risultato dell'esperimento
Ad esempio, si sa che un dato evento E si è verificato o si verificherà

- l'insieme dei casi possibili si restringe da Ω ad E
- quindi il valore della probabilità in genere non è più lo stesso

Esempio: si lancia un dado

1. la probabilità che esca un numero ≥ 4 (evento A) è $p(A) = 3/6 = 0.5$;
2. se si sa che il numero uscito è pari (evento B),
   la probabilità diventa $p(A|B) = 2/3 = 0.\overline{6}$

Esempio: tre clienti sono in coda, e si sa che uno è tesserato e gli altri due no:

- la probabilità dell'evento $A =$ "il primo è tesserato" è $p(A) = 1/3$
- la probabilità dell'evento $B =$ "il secondo è tesserato" è $p(B) = 1/3$, ma
  - se si sa che il primo era tesserato, diventa $p(B|A) = 0$
  - se si sa che il primo non era tesserato, diventa $p\left(B|\bar{A}\right) = 1/2$

Probabilità condizionata

Si dice probabilità condizionata $p(A|B)$ la probabilità che si verifichi A avendo la certezza che si verifichi B

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

La definizione è coerente con la teoria classica
Infatti, il rapporto fra il numero dei casi positivi e quello dei casi totali è

$$p(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{p(A \cap B)}{p(B)}$$

La definizione è coerente anche con la teoria frequentista:
il rapporto fra due limiti che hanno lo stesso denominatore
coincide con il limite del rapporto dei numeratori

(*Vedi Esercizio 4-1*)

Teorema di Bayes

Siccome l'operazione di intersezione di due insiemi è simmetrica

$$p(C \cap E) = p(C|E)\, p(E) = p(E|C)\, p(C) = p(E \cap C)$$

Sembra un gioco formale, ma se ne ricava un potente metodo di inferenza statistica

Consideriamo una situazione in cui

- l'evento C è una possibile causa (ad es., un pezzo è guasto)
- l'evento E è un possibile effetto (ad es., un test rileva un guasto)

Non stiamo parlando di fenomeni deterministici

- è possibile che l'effetto non si verifichi, e invece la causa sì (falso negativo: il pezzo è guasto, ma il test non lo rileva)
- è possibile che l'effetto si verifichi, e invece la causa no (falso positivo: il pezzo non è guasto, ma il test lo rileva come tale)

Grazie al teorema di Bayes

$$p(C|E) = p(E|C)\,\frac{p(C)}{p(E)}$$

se si è verificato E, possiamo valutare la probabilità che a provocare l'effetto E sia stata la causa C, $p(C|E)$, se conosciamo le probabilità

1. che la causa verificandosi provochi l'effetto $p(E|C)$
2. che si verifichi la causa $p(C)$
3. che si verifichi l'effetto $p(E)$

Esempio (dati)

Un turista vuol sapere se farà tempo bello (evento A) o brutto (evento $\bar{A}$) per decidere come vestirsi durante una gita

Consulta il barometro, che può dare i seguenti responsi:

- sereno (evento S)
- variabile (evento V)
- pioggia (evento P)

partizione dello spazio campione (coprono tutti i casi e sono incompatibili)

Possiamo pensare

- al tipo di tempo come causa
- al responso del barometro come effetto

Storicamente, si conoscono le probabilità assolute dei due tipi di tempo

tempo	$p\,(\text{tempo})$
A	0.40
$\bar{A}$	0.60

e le probabilità condizionate dei tre responsi rispetto ai due tipi di tempo

$p\,(\text{resp.}\vert\text{tempo})$	S	V	P
A	0.60	0.25	0.15
$\bar{A}$	0.20	0.30	0.50

Svolgimento (1)

A priori, è una situazione molto incerta con tendenza al brutto ($40\% - 60\%$)
Se però si conosce il responso, le cose possono cambiare

Calcoliamo le probabilità congiunte $p(A \cap E)$ e $p(\bar{A} \cap E)$ per $E \in \{S, V, P\}$

$p(\text{tempo} \cap \text{resp.})$	S	V	P
$p(A \cap E)$	0.24	0.10	0.06
$p(\bar{A} \cap E)$	0.12	0.18	0.30

Le loro somme riga per riga sono le probabilità assolute dei tipi di tempo, mentre le somme colonna per colonna sono le probabilità assolute dei responsi

$p(\text{tempo} \cap \text{resp.})$	S	V	P	$p(\text{tempo})$
$p(A \cap E)$	0.24	0.10	0.06	0.40
$p(\bar{A} \cap E)$	0.12	0.18	0.30	0.60
$p(\text{resp.})$	0.36	0.28	0.36	1.00

Ora si calcolano le probabilità condizionate dei tipi di tempo rispetto ai responsi, dividendo ogni elemento per la somma sulla corrispondente colonna

Svolgimento (2)

Ora si calcolano le probabilità condizionate dei tipi di tempo rispetto ai responsi

$$p\left(A|E\right) = \frac{p\left(A \cap E\right)}{p\left(E\right)} \qquad p\left(\bar{A}|E\right) = \frac{p\left(\bar{A} \cap E\right)}{p\left(E\right)}$$

$p\left(\text{tempo} \cap \text{resp.}\right)$	S	V	P	$p\left(\text{tempo}\right)$
$p\left(A \cap E\right)$	0.24	0.10	0.06	0.40
$p\left(\bar{A} \cap E\right)$	0.12	0.18	0.30	0.60
$p\left(\text{resp.}\right)$	0.36	0.28	0.36	1.00

$p\left(\text{tempo}|\text{resp.}\right)$	S	V	P
$p\left(A|\text{resp.}\right)$	0.67	0.36	0.17
$p\left(\bar{A}|\text{resp.}\right)$	0.33	0.64	0.83

da cui si deduce che, se il barometro dà:

- sereno (S), è più probabile che sia bello (67% contro 33%)
- variabile (V), è più probabile che sia brutto (64% contro 36%)
- pioggia (P), è molto più probabile che sia brutto (83% contro 17%)

Un'informazione ancora incerta, ma decisamente più utile e precisa

(*Vedi Esercizio 4-2 per lo svolgimento con Excel*)

Procedimento

Non sempre si parte dagli stessi dati: può capitare di avere

- le probabilità condizionate di un effetto rispetto a tutte le cause
  (*Vedi Esercizio 4-3*)
- le probabilità condizionate di esito positivo di un test rispetto al verificarsi di una causa e di esito negativo rispetto al non verificarsi
  (*Vedi Esercizio 4-4*)
- le frequenze assolute di tutte le combinazioni di effetti e cause
  (*Vedi Esercizio 4-5 ed Esercizio 4-6*)
- le probabilità congiunte di tutte le combinazioni di effetti e cause
- ...

In ogni situazione, il legame fra le quantità è lo stesso, ma

- si parte da dati diversi
- per ottenere risultati diversi

Falsi positivi

Si esegue il test dell'HIV su un gruppo di individui

- si stima che siano malati lo 0.25% degli individui;
- se uno è malato, il test lo rileva con probabilità 98% (sensibilità)
- se uno è sano, il test lo rileva con probabilità 99% (sensitività)

Con quale probabilità una persona è malata quando il test è positivo?

Apparentemente, dovrebbe essere molto alta: è forse il 98%?

Non è detto: sia $C =$ "individuo malato" ed $E =$ "test positivo"

$$p(E|C) = 98\% \qquad \text{Quanto vale } p(C|E)?$$

Ebbene

$$p(C|E) = p(C)\,\frac{p(E|C)}{p(E)} = 0.25\% \cdot \frac{98\%}{1.24\%} = 20\%$$

La probabilità a posteriori cresce molto grazie al test, ma
su eventi rari le campagne di test a tappeto rimangono inutili o dannose

Correlazione e causalità

Si è parlato sinora disinvoltamente di causa ed effetto

In realtà, la probabilità valuta correlazioni, non causalità

- i due eventi studiati possono essere entrambi effetti di un terzo
  (*ad es., numero di scarpa e abilità nella lettura sono legati all'età*)
- i due eventi studiati possono essere correlati senza motivo, solo perché lo spazio campionario che li contiene è così
  (*ad esempio, il lancio di un dado e il non essere un cubo perfetto*)

Quindi il fatto che due eventi siano causa ed effetto è un'ipotesi a priori, da dimostrare con altri mezzi

Le leggi della probabilità valgono però anche per eventi solo correlati

Eventi indipendenti

Due eventi A e B sono indipendenti tra loro quando
il verificarsi di uno dei due non modifica la probabilità dell'altro

$$p(A|B) = p(A) \qquad p(B|A) = p(B)$$

Esempio: nel lancio di un dado

- A: il numero estratto è pari
- B: il numero estratto è un quadrato perfetto

Abbiamo infatti

$$p(A|B) = p(A) = \frac{1}{2} \qquad p(B|A) = p(B) = \frac{1}{3}$$

Ne deriva che se A e B sono indipendenti

$$p(A \cap B) = p(A) \cdot p(B)$$

La cosa vale anche per n eventi $A_1, \ldots A_n$ indipendenti a coppie

Esempio

Durante la mattina, la probabilità che arrivi un ordine

- fra le 9 e le 11 è $p(A) = 0.9$
- fra le 10 e le 12 è $p(B) = 0.8$

Se supponiamo che gli arrivi degli ordini siano indipendenti, qual è la probabilità che

1. arrivi un ordine fra le 10 e le 11?

$$p(A \cap B) = p(A) \cdot p(B) = 0.9 \cdot 0.8 = 0.72$$

2. arrivi un ordine fra le 9 e le 12?

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) = 0.9 + 0.8 - 0.72 = 0.98$$

3. non arrivino ordini fra le 9 e le 12?

$$p\left(\overline{A \cup B}\right) = 1 - p(A \cup B) = 1 - 0.98 = 0.02$$