Cosa accadrà domani?

Max Pierini

*info@maxpierini.it

May 29, 2020

Ci alziamo una mattina di sole e scopriamo che il tratto di strada visibile dalla nostra finestra è bagnato, ma il giorno passato non ha piovuto. Quante probabilità ci sono che abbia piovuto mentre dormivamo la scorsa notte?

Per rispondere a questa domanda possiamo affidarci alla *statistica inferenziale bayesiana* il cui scopo è determinare, con teoremi e formule più o meno complesse, di rispondere a domande simili a questa nei campi più disparati.

La categoria di "domanda" che abbiamo posto è chiamata probabilità **a posteriori** o **condizionata**: ci chiediamo infatti quale sia la probabilità dell'evento "pioggia la scorsa notte" (che chiameremo variabile X) condizionata dal, ovvero a posteriori del, fatto noto "tratto di strada bagnato" (che chiameremo parametro θ). Scriveremo

$$P(X|\theta) = ?$$

e leggeremo "probabilità di X dato θ". Il parametro θ rappresenta ciò che conosciamo (la strada è bagnata) e la variabile X la *verità* che non conosciamo e che vogliamo scoprire (ha piovuto la scorsa notte?).

Analizzando il problema, ci renderemo facilmente conto del fatto che la risposta dipende da una serie di fattori. Vediamoli uno per uno:

1. che probabilità c'è che una strada si bagni se piove?
2. che probabilità c'è, in generale, che piova una sola notte?
3. che probabilità c'è, in generale, che una strada si bagni?

La domanda 1. è ancora una probabilità condizionata (o a posteriori) ma è l'opposto della domanda iniziale: ci chiediamo infatti che probabilità ha una strada di bagnarsi a condizione (o a posteriori del fatto) che piova. Senza entrare troppo nel dettaglio, possiamo facilmente stabilire che sia piuttosto elevata: per non bagnarsi dovrebbe essere "coperta", ma allora ci troveremmo di fronte ad esempio a un porticato e non più ad un tratto di strada, oppure dovrebbe essere un tratto molto ristretto ad esempio coperto da un'auto parcheggiata che poi si è spostata, ecc. Le probabilità vanno da 0 a 1, ovvero da 0% a 100%.

Assegnamo pertanto a questa probabilità un valore prossimo al 100%, per esempio 99%. La probabilità a posteriori del parametro θ (strada bagnata) dato lo specifico valore della variabile X (pioggia) sarà

$$P(\theta|X) = 99.00\%$$

che leggeremo "probabilità di θ dato X".

La domanda 2. è invece una probabilità non condizionata ovvero **a priori**: ci si chiede qual'è la probabilità a priori di una sola notte di pioggia. Per rispondere a questa domanda potremmo

controllare gli archivi del servizio meteorologico del posto in cui viviamo e segnarci tutte le singoli notti di pioggia, diciamo, in un anno: saranno molto frequenti se viviamo, ad esempio, in Irlanda o Scozia ma molto meno se viviamo in Italia, dato che le condizioni climatiche portano più spesso a giornate intere di pioggia (magari consecutive). Supponiamo di scoprire che ci siano state 40 singole notti di pioggia nello scorso anno; la probabilità a priori della variabile pioggia X sarà pertanto

$$P(X) = 40/365 = 10.96\%$$

La domanda 3. è, come la precedente, una probabilità a priori ma più difficile da determinare. Perché? Ci chiediamo quale sia la probabilità a priori che una strada sia bagnata, dovremmo quindi tener conto di tutte le possibilità tali per cui una strada si bagni, dalle più probabili (la pioggia stessa, un camion delle pulizie stradali passato da poco ecc) alle più improbabili (il ribaltamento di un carretto di consegne di bottiglie di acqua con rottura di sufficienti bottiglie, l'esondazione di un canale di irrigazione abbastanza vicino, l'atterraggio di un'astronave aliena che usi acqua come propellente… ecc). Sembra quasi impossibile da determinare, ma esiste un metodo che può venirci in aiuto: possiamo semplicemente sommare tra loro

- la probabilità che una strada si bagni se piove, che abbiamo già stabilito infatti $P(\theta|X) = 99.00\%$, moltiplicato per la probabilità che piova una notte, che conosciamo già $P(X) = 10.96\%$
- la probabilità che una strada si bagni se non piove $P(\theta|\overline{X})$, che leggeremo "probabilità di θ dato non X" moltiplicato per la probabilità che non piova solo una notte $P(\overline{X})$, che leggeremo "probabilità di non X"

del primo termine conosciamo già tutto, ma come facciamo a determinare i valori del secondo termine? Grazie alle regole delle probabilità, la probabilità "non piova una sola notte" è complementare alla probabilità che "piova una sola notte", di cui abbiamo già stabilito il valore:

$$P(\overline{X}) = 1 - P(X) = 1 - 10.96\% = 89.04\%$$

Ci serve dunque assegnare un valore a $P(\theta|\overline{X})$. Osserviamo che, per le regole delle probabilità

$$P(\theta|\overline{X}) = 1 - P(\overline{\theta}|\overline{X})$$

ovvero: la probabilità che una strada *si bagni se non piove* è complementare alla probabilità che *non si bagni se non piove* che corrisponde alla probabilità che una strada sia asciutta se non piove.

In questo caso specifico possiamo solo assegnare un valore arbitrario a $P(\overline{\theta}|\overline{X})$ risparmiandoci solamente il compito di immaginare le situazioni più improbabili (come l'astronave aliena) e considerando che "è molto probabile" che una strada sia asciutta in assenza di pioggia, ad esempio 97%:

$$P(\theta|\overline{X}) = 1 - P(\overline{\theta}|\overline{X}) = 1 - 97.00\% = 3.00\%$$

in altre situazioni invece è un valore noto (come nel caso dei test diagnostici in campo medico che vedremo in seguito).

Possiamo pertanto determinare che la probabilità a priori che una strada sia bagnata è pari al

$$P(\theta) = P(\theta|X)P(X) + P(\theta|\overline{X})P(\overline{X}) = 99.00\% \cdot 10.96\% + 3.00\% \cdot 89.04\% = 13.52\%$$

Questa operazione appena effettuata

$$P(\theta) = \sum_i \Big(P(\theta|X_i)P(X_i) \Big)$$

è chiamata *marginalizzazione del parametro* θ.

Arrivati a questo punto chiediamoci: quale relazione hanno le tre probabilità determinate con la probabilità che vogliamo ottenere $P(X|\theta)$?.

Possiamo dire che

- la probabilità che una strada sia bagnata se ha piovuto $P(\theta|X)$ e la probabilità che piova una notte sola $P(X)$ sono direttamente proporzionali alla probabilità che abbia piovuto la scorsa notte data la strada bagnata: entrambe queste probabilità contribuiscono in modo positivo alla risposta che cerchiamo
- la probabilità che una strada sia bagnata in generale $P(\theta)$ è inversamente proporzionale alla probabilità che abbia piovuto la scorsa notte se la strada è bagnata: ci potrebbero essere una serie di altri motivi per cui la strada sia bagnata e noi vogliamo solo quelli direttamente collegati alla pioggia

riassumendo

$$P(X|\theta) = \frac{P(\theta|X)P(X)}{P(\theta)}$$

l'equazione appena scritta è il *Teorema di Bayes*.

Assegnamo ai termini del teorema i valori ricavati in precedenza e calcoliamo

$$P(X|\theta) = \frac{P(\theta|X)P(X)}{P(\theta)} = \frac{99.00\% \cdot 10.96\%}{13.52\%} = 80.24\%$$

Siamo giunti pertanto a determinare, grazie alle regole delle probabilità e al Teorema di Bayes, che la probabilità che la scorsa notte abbia piovuto dato il fatto che vediamo un tratto di strada bagnata è pari al 80.24%.

Nel caso specifico appena trattato, la risposta sembra piuttosto ovvia e vano l'intero sforzo per arrivare al risultato, ma in altre situazioni il calcolo bayesiano della probabilità a posteriori può dare responsi anche molto lontani da quelli che ci suggerirebbe il semplice intuito.

Supponiamo ad esempio che un paziente venga sottoposto ad un test di screening per una malattia M piuttosto rara, ad esempio con una prevalenza **PV** nella popolazione di appartenenza dello 0.002%, ovvero 2 casi su 100,000.

Il test cui verrà sottoposto è di tipo qualitativo e fornirà un responso dicotomico: positivo $\oplus$ oppure negativo $\ominus$. Sappiamo, dalla letteratura medica che abbiamo consultato, che il test ha sensibilità **SE** 85% e specificità **SP** 99% (per approfondimenti, si veda Link).

Dal punto di vista della statistica bayesiana sensibilità e specificità sono due probabilità a posteriori, rispettivamente:

$$\mathbf{SE} = P(\oplus|M) = 85.00\%$$

$$\mathbf{SP} = P(\ominus|\overline{M}) = 99.00\%$$

ovvero la probabilità a posteriori di ottenere un test positivo se si è malati e la probabilità a posteriori di ottenere un test negativo se non si è affetti dalla relativa patologia.

La prevalenza **PV** della malattia in questione invece è, in questo caso, una probabilità a priori: dato che non abbiamo alcuna informazione sul paziente se non quella di appartenere alla popolazione scelta per lo screening, è la probabilità a priori che il paziente sia affetto dalla malattia $P(M)$.

Quello che vogliamo ottenere sottoponendo il paziente al test è la probabilità a posteriori dato il risultato $\odot$ del test

$$P(M|\odot)$$

applicando quindi il Teorema di Bayes otteniamo che

$$P(M|\odot) = \frac{P(\odot|M)P(M)}{P(\odot)}$$

dove M è la variabile della *verità* (malattia) che vogliamo determinare e $\odot$ è il parametro osservato.

Supponiamo che il test sia risultato positivo $\oplus$, la nostra formula diventerà pertanto

$$P(M|\oplus) = \frac{P(\oplus|M)P(M)}{P(\oplus)}$$

Osserviamo che

- $P(\oplus|M)$ altro non è che la sensibilità **SE** del test
- $P(M)$ è, come detto in precedenza, la prevalenza della malattia

Come determiniamo $P(\oplus)$ ovvero la probabilità a priori di ottenere un test positivo?

Possiamo procedere *marginalizzando* il parametro $\oplus$ come fatto in precedenza per la probabilità a priori di osservare una strada bagnata

$$P(\oplus) = P(\oplus|M)P(M) + P(\oplus|\overline{M})P(\overline{M})$$

come prima, conosciamo il primo termine della somma (sensibilità e prevalenza) e sappiamo che

$$P(\overline{M}) = 1 - P(M) = 1 - 0.002\% = 99.998\%$$

ma, a differenza dell'esempio precedente, possiamo determinare con precisione il secondo termine coi dati a disposizione infatti

$$P(\oplus|\overline{M}) = 1 - P(\ominus|\overline{M}) = 1 - \mathbf{SP} = 1 - 99.00\% = 1.00\%$$

Possiamo quindi ottenere

$$P(\oplus) = P(\oplus|M)P(M) + P(\oplus|\overline{M})P(\overline{M}) = \mathbf{SE}\cdot\mathbf{PV} + (1-\mathbf{SP})(1-\mathbf{PV}) = 85.00\%\cdot 0.002\% + 1.00\%\cdot 99.998\% = 1.002\%$$

Abbiamo perciò tutti i termini necessari a determinare la probabilità di malattia a posteriori dato il risultato positivo del test. Visto il valore molto elevato di sensibilità e specificità ci aspetteremmo di ottenere un risultato altrettanto notevole ma applicando il teorema di Bayes otteniamo

$$P(M|\oplus) = \frac{P(\oplus|M)P(M)}{P(\oplus)} = \frac{85.00\% \cdot 0.002\%}{1.002\%} = 0.17\%$$

ovvero, in seguito a risultato positivo del test, il nostro paziente avrà una probabilità a posteriori di essere malato pari soltanto al 0.17%…!

Perché è così bassa nonostante sensibilità e specificità fossero piuttosto elevate? Il risultato dipende dal fatto che siamo partiti con un'esigua probabilità di malattia a priori $\mathbf{PV} = 0.002\%$.

Questo esempio spiega perciò più chiaramente in cosa consista l'analisi bayesiana:

> L'analisi bayesiana consiste nella *ridistribuzione* della probabilità *a priori* delle nostre convinzioni riguardo la probabilità di un determinato evento X in base all'osservazione di un certo parametro θ

Ripensando all'esempio precedente della strada, cosa abbiamo fatto in realtà? Abbiamo *ridistribuito* la probabilità a priori circa le nostre convinzioni di una notte isolata di pioggia in base all'osservazione del tratto di strada bagnata. Se non avessimo osservato θ o se avessimo osservato qualcosa di differente (ad esempio, strisce bagnate sulla strada o dei lavori in corso sulla tubazioni, ecc) la probabilità a priori di pioggia la notte scorsa sarebbe stata modificata di conseguenza.

Ma come può il ragionamento seguito finora portare ad una stima di previsione riguardo la probabilità un determinato evento X?

Supponiamo che la variabile X di cui vogliamo calcolare la probabilità sia il Tasso di Riproduzione Effettivo R_t di COVID-19 in Italia e il parametro θ osservato siano i nuovi casi giornalieri k di COVID-19 in Italia dal primo all'ultimo giorno dei dati registrati.

I calcoli saranno più complessi visto che non ci stiamo più occupando di eventi dicotomici, come malattia o salute, strada bagnata o asciutta, notte di pioggia o non, risultato positivo o negativo, bensì una scala continua dei possibili valori di R_t da un minimo di 0 ad un massimo da definire e una scala discreta di valori dei nuovi casi k osservati da un minimo di 0 ad un massimo virtualmente

illimitato, ma ciò di cui trattiamo sono comunque probabilità. Infatti, ciò che vogliamo ottenere è $P(R_t|k)$ ovvero, a partire da una convinzione a priori circa le probabilità di ottenere tutti i valori possibili di R_t, ridistribuire per ogni giorno in archivio le probabilità dei valori di R_t in base all'osservazione dei k nuovi casi registrati. Usando la notazione ormai nota e il Teorema di Bayes:

$$P(R_t|k) = \frac{P(k|R_t)P(R_t)}{P(k)}$$

dove

- $P(k|R_t)$ sono le probabilità a posteriori di osservare i possibili nuovi casi per ciascuno dei possibili valori di R_t
- $P(R_t)$ sono le probabilità a priori per ogni valore possibile del numero di riproduzione effettivo
- $P(k)$ è la probabilità a priori di osservare il numero di nuovi casi giornalieri effettivamente misurato per ogni giorno in archivio

La probabilità a posteriori di R_t calcolata per ciascun giorno diventa la probabilità a priori per il giorno successivo, che verrà nuovamente ridistribuita in base ai nuovi casi giornalieri osservati, ecc… fino all'ultimo giorno (presumibilmente oggi).