15 Stimatori Bayesiani

Ricordiamo innanzitutto la Formula di Bayes:

Theorem 123 (Bayes)**.** *Siano $H_1, ...H_m$ eventi disgiunti ed esaustivi, cioè $H_i \cap H_j = \emptyset$ per ogni $i \neq j$ e $\cup_{i=1}^m H_i = \Omega$. Sia inoltre $E \in \mathfrak{S}$ un evento con probabilità strettamente positiva; allora*

$$Pr(H_i|E) = \frac{Pr(E|H_i)Pr(H_i)}{\sum_{j=1}^m Pr(E|H_j)Pr(H_j)}.$$

Spiegazione Semplice

Cos'è il Teorema di Bayes?
Questo teorema è la formula matematica per "aggiornare una credenza" in modo logico quando si ricevono nuove informazioni.

- H_i **(Ipotesi):** Le possibili "cause" o "stati del mondo" (es. H_1 ="l'urna è la A", H_2 ="l'urna è la B").
- E **(Evidenza):** Il "dato" o "effetto" che osserviamo (es. "ho estratto una pallina bianca").
- $Pr(H_i)$ **(Probabilità "a priori"):** La nostra credenza iniziale sull'ipotesi H_i, *prima* di vedere l'evidenza E.
- $Pr(E|H_i)$ **(Verosimiglianza):** La probabilità di osservare l'evidenza E, *supponendo* che l'ipotesi H_i sia vera.
- $Pr(H_i|E)$ **(Probabilità "a posteriori"):** La probabilità *aggiornata* dell'ipotesi H_i, *dopo* aver visto l'evidenza E.

Il teorema ci dice come passare dalla credenza iniziale $P(H_i)$ a quella aggiornata $P(H_i|E)$.

La derivazione della formula di Bayes è matematicamente banale (si riduce essenzialmente a ricordare che $Pr(E) = \sum_{j=1}^m Pr(E|H_j)Pr(H_j)$; l'interpretazione però è molto importante, perché permette di combinare in modo matematicamente la probabilità a priori di m cause disgiunte H_j e l'evidenza empirica sul fatto che E si sia verificato.

Example 124. *L'approccio Bayesiano all'inferenza statistica è profondamente diverso da quello che abbiamo seguito finora. L'idea di fondo è che non esista un "vero" parametro da stimare $\theta = \theta_0$, ma che il parametro sia esso stesso una variabile (o un vettore) aleatorio la cui distribuzione, che rappresenta il nostro stato di conoscenza, viene aggiornata tramite la formula di Bayes alla luce delle osservazioni. In altre parole, oltre alla legge delle osservazioni $f(X_1, ..., X_n|\theta)$ dobbiamo supporre di conoscere la legge $\pi(\theta)$ del parametro prima di aver effettuato osservazioni. Si noti come abbiamo scritto $f(X_1, ..., X_n|\theta)$ invece di $f(X_1, ..., X_n; \theta)$ perché ora ha senso parlare della legge di $X_1, ..., X_n$ condizionatamente al valore θ del parametro. L'oggetto centrale dell'inferenza diviene quindi la legge a posteriori, che attraverso la formula di Bayes è data da*

$$\pi(\theta|X_1, ..., X_n) = \frac{f(X_1, ..., X_n|\theta)\pi(\theta)}{\int_\Theta f(X_1, ..., X_n|\theta)\pi(\theta)d\theta}$$

ed analogamente nel caso discreto.

Spiegazione Semplice

La Filosofia Bayesiana (Frequentista vs. Bayesiano)
Questo è il cambio di mentalità fondamentale dell'approccio Bayesiano.

- **Approccio Classico (Frequentista):** Il parametro θ (es. la probabilità p di una moneta) è un numero *fisso* e sconosciuto. L'unica cosa casuale sono i *dati* X. Il nostro obiettivo è "indovinare" quel numero.

- **Approccio Bayesiano:** Il parametro θ è *esso stesso* una variabile aleatoria. Non ha un "solo" valore vero, ma una *distribuzione di probabilità* $\pi(\theta)$ che esprime la nostra incertezza su di esso.

L'inferenza Bayesiana consiste nell'usare i dati per "aggiornare" la nostra distribuzione di credenza.

- $\pi(\theta)$ **(Priore):** È la distribuzione che θ ha *prima* di vedere i dati (la nostra "credenza iniziale").
- $f(X|\theta)$ **(Verosimiglianza):** È la probabilità dei dati, dato un certo θ. È la stessa funzione $L(\theta)$ del Capitolo 11.
- $\pi(\theta|X)$ **(Posteriore):** È la distribuzione (aggiornata) di θ *dopo* aver visto i dati X. Questa è la nostra "risposta".

La formula $\pi(\theta|X) = \frac{\text{Verosimiglianza} \times \text{Priore}}{\text{Costante}}$ è il Teorema di Bayes applicato ai parametri.

Example 125. *Consideriamo $X_1, ..., X_n$ variabili Bernoulliane di parametro p; per quest'ultimo, assumiamo che abbia una distribuzione a priori di tipo Beta con parametri α, β, cioè*

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad p \in [0,1].$$

Ricordiamo innanzitutto i valori di valor medio e varianza

$$E[p] = \frac{\alpha}{\alpha+\beta} \quad Var[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

infatti

$$\begin{aligned} E[p] &= \int_0^1 p \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} \int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} p^{\alpha}(1-p)^{\beta-1} dp \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

e

$$\begin{aligned} E[p^2] &= \int_0^1 p^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha+2)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+2)} \int_0^1 \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+2)\Gamma(\beta)} p^{\alpha+1}(1-p)^{\beta-1} dp \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} \end{aligned}$$

$$Var[p] = \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2}$$

$$= \frac{\alpha(\alpha+1)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Abbiamo inoltre, scrivendo $y = \sum_{i=1}^n X_i$

$$\begin{aligned} f(y|p)\pi(p) &= \binom{n}{y} p^y (1-p)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}. \end{aligned}$$

La marginale a denominatore si ottiene come segue:

$$\begin{aligned} f(y) &= \int_0^1 \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} \end{aligned}$$

Infine la distribuzione a posteriori è data da

$$\begin{aligned} \pi(p|y) &= \frac{\binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}{\binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}, \end{aligned}$$

cioè è ancora una beta, con parametri aggiornati. Quando la distribuzioni a posteriori assume la stessa forma dell'a priori si parla di leggi coniugate.

Spiegazione Semplice

Esempio Pratico: "Beta-Binomiale" e Leggi Coniugate
Questo è l'esempio "classico" di inferenza Bayesiana.

- **Verosimiglianza (Dati):** I dati sono lanci di moneta (Bernoulliani). La loro somma y (numero di Teste) segue una distribuzione *Binomiale*. $f(y|p) \propto p^y(1-p)^{n-y}$.
- **Priore (Credenza Iniziale):** Scegliamo una distribuzione *Beta* per la nostra credenza iniziale sul parametro p. La Beta è una distribuzione flessibile che vive solo tra 0 e 1, perfetta per una probabilità. I parametri α e β "modellano" la nostra credenza (es. $\alpha = \beta = 1$ significa "non so nulla", è una uniforme; $\alpha = 10, \beta = 1$ significa "sono quasi certo che p sia vicino a 1").
- **Posteriore (Risultato):** Dopo aver applicato la formula di Bayes e svolto l'algebra, la distribuzione posteriore $\pi(p|y)$ è *ancora* una Beta!

Questo è comodissimo. La distribuzione *Posteriore* è una $Beta(y+\alpha, n-y+\beta)$. Non abbiamo dovuto fare integrali complicati, abbiamo solo "aggiornato" i parametri.
Leggi Coniugate: Quando la distribuzione Priore (es. Beta) e la Verosimiglianza (es. Binomiale) sono "accoppiate" in modo tale che la Posteriore appartenga alla stessa famiglia della Priore, si parla di **leggi coniugate**. Questo rende i calcoli molto più facili.

Remark 126. *Una interpretazione rigorosa dell'approccio Bayesiano dovrebbe concludersi con la derivazione della distribuzione a posteriori: il calcolo di uno stimatore "puntuale "non ha strettamente senso, visto che il paramrto non ha un singolo valore. In pratica però il calcolo degli stimatori Bayesiani si conclude molto spesso con la derivazione di un singolo valore di sintesi, come ad esempio il valore che massimizza la distribuzione a posteriori o ancora più spesso il valore medio a posteriori. Nel caso della binomiale con a priori beta otteniamo*

$$\begin{aligned} \hat{p}_{Bayes} &= \int_0^1 p \cdot \pi(p|y) dp \quad \textit{(calcolo della media della posteriore)} \\ &= \frac{y+\alpha}{n+\alpha+\beta} = \frac{y}{n}\frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{n+\alpha+\beta} \end{aligned}$$

L'ultima espressione è illuminante, perchè rappresenta lo "stimatore Bayesiano" come una media ponderata di due elementi: lo stimatore classico di massima verosimiglianza (in questo caso, la semplice media aritmetica) ed il valore atteso a priori:

$$\frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n, \frac{\alpha}{\alpha+\beta}$$

Il peso dello stimatore di massima verosimiglianza converge ad 1 quando la dimensione del campione n diverge all'infinito; intuitvamente, le nostre opinioni a priori sono schiacciate dalla forza dell'evidenza empirica.

Spiegazione Semplice

Come si ottiene uno Stimatore Bayesiano?
Questo remark è cruciale. L'output "puro" Bayesiano non è un singolo numero, ma l'intera distribuzione posteriore $\pi(\theta|X)$ (che descrive la nostra incertezza aggiornata).
Tuttavia, spesso serve un singolo numero (uno "stimatore puntuale"). La scelta più comune è la **media della distribuzione posteriore** (il suo valore atteso).
Per l'esempio Beta-Binomiale, la media della distribuzione posteriore $Beta(y+\alpha, n-y+\beta)$ è:
$\hat{p}_{Bayes} = \frac{y+\alpha}{n+\alpha+\beta}$
L'Interpretazione Illuminante: Il testo riscrive questa formula in un modo molto intelligente:
$\hat{p}_{Bayes} = \left(\frac{y}{n}\right) \times \left(\frac{n}{n+\alpha+\beta}\right) + \left(\frac{\alpha}{\alpha+\beta}\right) \times \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)$
Questa è una **media ponderata** tra:

- **La Stima dei Dati (MLE):** $\frac{y}{n}$ (la media campionaria)
- **La Stima della Credenza Iniziale (Media Priore):** $\frac{\alpha}{\alpha+\beta}$

I "pesi" sono n (la quantità di dati reali) e $\alpha+\beta$ (la "quantità" di credenza iniziale).
Se n è piccolo (pochi dati), la stima Bayesiana è una via di mezzo tra i dati e la credenza iniziale.
Se n è enorme, $n \to \infty$, il peso $\frac{n}{n+\alpha+\beta} \to 1$. Questo significa che *l'evidenza empirica (i dati) "schiaccia" la credenza a priori*, e lo stimatore Bayesiano converge a quello classico (MLE).

Exercise 127. *Sia X una Gaussiana con valor medio θ e varianza σ^2, con θ essa stessa Gaussiana di parametri μ, τ^2. Abbiamo che*

$$\pi(\theta|X) \sim N(\frac{\tau^2}{\tau^2+\sigma^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}).$$