Teoria dell'informazione e inferenza statistica

Sia $\{p_\vartheta(x)\}_{\vartheta}$ una famiglia di d.d.p. parametrizzata da un parametro $\vartheta \in \mathbb{R}^k$

Esempio: Lancio di una moneta

$$p_\vartheta(x) = \begin{cases} \vartheta & \text{se } x \text{ è Testa } (0) \\ 1-\vartheta & \text{se } x \text{ è croce } (1) \end{cases}$$

Sia X una v.a. campionata da p_ϑ : $X \sim p_\vartheta$ $\quad X \sim p_\vartheta^n$

Una statistica è una funzione $T(X)$ del campione X

Esempio: media campione : $T(\underbrace{X_1, X_2, \dots, X_n}_{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$

Poiché $T(X)$ è una funzione del campione, vale la catena di Markov : $\vartheta \to X \to T(X)$

$$\left(Pr[T(X) \mid \vartheta, X] = Pr[T(X) \mid X] \right)$$

$\Rightarrow$ Per il 2° teorema di elaborazione dati, $I(\vartheta; T(X)) \overset{?}{\leq} I(\vartheta; X)$

Quando $I(\vartheta; T(X)) = I(\vartheta; X)$, $T(X)$ è detta statistica sufficiente

Questo è vero quando

$$Pr[X \mid \vartheta, T(X)] = Pr[X \mid T(X)]$$

ovvero quando vale la catena di Markov: $\vartheta \to T(X) \to X$

Esempio. $X_1, X_2, \dots, X_n$ v.a. indipendenti, binarie, identicamente distribuite; con $p(1) = \vartheta$, $p(0) = 1-\vartheta$ $(\vartheta \in [0,1])$

$$Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n)] = \vartheta^{\sum_i x_i} (1-\vartheta)^{n - \sum_i x_i}$$

$$T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$$

$$\underbrace{Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n) \mid \textstyle\sum_{i=1}^n X_i = k]}_{Pr[X \mid T(X)]} = \begin{cases} 1/\binom{n}{k} & \text{se } \sum_{i=1}^n x_i = k \\ 0 & \text{se } \sum_{i=1}^n x_i \neq k \end{cases}$$

↳ non dipende da ϑ

$$\Rightarrow \textstyle\sum_i X_i \text{ è una statistica sufficiente}$$

$$Pr[(X_1, X_2, \dots, X_5) = 1000\,1] = \vartheta(1-\vartheta)^3\vartheta = \vartheta^2(1-\vartheta)^3$$

$$Pr[\underbrace{0100\,1}_{\binom{5}{2}}] = \underline{\vartheta^2(1-\vartheta)^3} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Stima a massima verosimiglianza (MLE: Maximum Likelihood Estimation)

Scenario: Ho una famiglia di d.d.p $\{p_\vartheta(x)\}_\vartheta$; esiste un modello corretto $p_{\vartheta^*}(x)$ ma non conosco ϑ^*

Qual è il valore di ϑ maggior consistente con le osservazioni?

Una possibilità è di misurare la consistenza di ϑ (rispetto al valore corretto ϑ^*) come segue:

$$D(p_{\vartheta^*} \| p_\vartheta) = \mathop{\mathbb{E}}_{X\sim p_{\vartheta^*}}\left[\log \frac{p_{\vartheta^*}(X)}{p_\vartheta(X)}\right] = \underbrace{\mathop{\mathbb{E}}_{X\sim p_{\vartheta^*}}[\log p_{\vartheta^*}(X)]}_{\text{non dipende da } \vartheta} - \mathop{\mathbb{E}}_{X\sim p_{\vartheta^*}}[\log p_\vartheta(X)]$$

$$= \text{costante (rispetto a } \vartheta) - \mathop{\mathbb{E}}_{X\sim p_{\vartheta^*}}[\log p_\vartheta(X)]$$

Non la conosco, ma posso stimarla:

per la legge dei grandi numeri,

$$\frac{1}{n}\sum_{i=1}^n \log p_\vartheta(X_i) \xrightarrow{\text{prob.}} \mathop{\mathbb{E}}_{X\sim p_{\vartheta^*}}[\log p_\vartheta(X)] \quad (n\to\infty)$$

Considero:

$$\hat{D}(p_{\vartheta^*} \| p_\vartheta) = \text{costante (rispetto a } \vartheta) - \frac{1}{n}\sum_{i=1}^n \log p_\vartheta(X_i)$$

Scelgo ϑ in modo da minimizzare $\hat{D}(p_{\vartheta^*} \| p_\vartheta)$

In generale,

$$\min_\vartheta \hat{D}(p_{\vartheta^*} \| p_\vartheta) \iff \min_\vartheta \left[-\frac{1}{n} \sum_{i=1}^n \log p_\vartheta(X_i) \right]$$

$$\iff \min_\vartheta \left[-\frac{1}{n} \log \prod_{i=1}^n p_\vartheta(X_i) \right]$$

$$\iff \max_\vartheta \left[\log \prod_{i=1}^n p_\vartheta(X_i) \right]$$

$\log(\cdot)$ è una funz. monotona

$$\iff \boxed{\max_\vartheta \prod_{i=1}^n p_\vartheta(X_i)}$$

$\prod_{i=1}^n p_\vartheta(X_i)$ → funzione di verosimiglianza (likelihood)

$\max_\vartheta \mathcal{L}(\vartheta | X)$

Esempio: $p_\vartheta(X) = \begin{cases} \vartheta & \text{se } X=1 \\ 1-\vartheta & \text{se } X=0 \end{cases}$ $(\vartheta \in [0,1])$

$$\mathcal{L}(\vartheta | \overbrace{X_1, \dots, X_n}^{X}) = \prod_{i=1}^n p_\vartheta(X_i) = \vartheta^k (1-\vartheta)^{n-k}$$ dove $k = \sum_{i=1}^n X_i$

Osservazione: $\mathcal{L}(\vartheta | X) \geq 0$ sempre

$\mathcal{L}(0|X) = 0 = \mathcal{L}(1|X)$

Calcolo $\mathcal{L}'(\vartheta | X) = k\vartheta^{k-1}(1-\vartheta)^{n-k} - (n-k)\vartheta^k (1-\vartheta)^{n-k-1}$

$$\mathcal{L}'(\vartheta|X) = k\vartheta^{k-1}(1-\vartheta)^{n-k} + (n-k)\vartheta^{k}(1-\vartheta)^{n-k-1}$$

$$= \underbrace{\vartheta^{k-1}}\ \underbrace{(1-\vartheta)^{n-k-1}}\ \underbrace{\left[k(1-\vartheta) - (n-k)\vartheta\right]}$$

Quando ho $\mathcal{L}'(\vartheta|X) = 0$? $\vartheta = 0$ oppure $\vartheta = 1$

oppure $k(1-\vartheta) - (n-k)\vartheta = 0$

$\Updownarrow$

$$k - k\vartheta - (n-k)\vartheta = 0$$

$$k - n\vartheta = 0 \quad \Rightarrow \quad \boxed{\vartheta = k/n}$$

Esercizio. Codifica run-length

Siano $X_1, \dots, X_n$ n v.a. $\in \{0,1\}$, non necessariamente indipendenti

Sia $R = (R_1, R_2, \dots)$ la sequenze delle run

$X =$ 000 11 00 1 0 0 ↕ codifica

111 00 11 0 1 1 → $R = (3, 2, 2, 1, 2)$

3 2 2 1 2

Si confrontino:

→ $H(X_1, \dots, X_n)$ (cioè $\underline{H(X)}$), $H(R_1, R_2, \dots)$ (cioè $H(R)$), e $\underline{H(X_n, R)}$

e limitare le differenze tra queste 3 quantità.

Osservo che $R = (R_1, R_2, \dots)$ è una funzione di X; quindi $H(R) \leq H(X)$

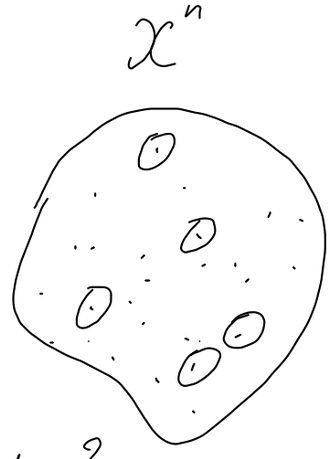
Inoltre: X è una funzione della coppia (X_n, R) → $H(X) \leq H(X_n, R)$

(X_n, R) è una funzione di X → $H(X_n, R) \leq H(X)$

} $H(X_n, R) = H(X)$

$$H(R) \leq H(X) = H(X_n, R) = H(R) + H(X_n | R) \leq H(R) + \underbrace{H(X_n)}_{\leq 1 = \log_2 |\{0,1\}|} = H(R) + 1 \dots$$

Esercizio. Siano $X_1, X_2, \dots, X_n$ di v.a. indipendenti e identicamente distribuite, con entropia $\bar{H}$.



Sia $C_n(t) = \{ (x_1, x_2, \dots, x_n) \in \mathcal{X}^n : p(x_1, \dots, x_n) \geq \underline{2^{-nt}} \}$

(a) Mostrare che $|C_n(t)| \leq 2^{nt}$

(b) Per quali valori di t si ha $\Pr[(X_1, \dots, X_n) \in C_n(t)] \xrightarrow{n\to\infty} 1$?

(a) $1 \geq \Pr[(X_1, \dots, X_n) \in C_n(t)] \geq |C_n(t)| \cdot \min_{x \in C_n(t)} p(x) \geq |C_n(t)| \cdot 2^{-nt}$

$\Rightarrow |C_n(t)| \leq 2^{nt}$.

(b) Per il principio di equipartizione asintotica:

$\underbrace{-\frac{1}{n} \log p(X)}_{\text{autoinformaz. normalizzata}} \xrightarrow{\text{prob.}} \bar{H}$ $\Rightarrow$ $p(X)$ si comporta come $2^{-n\bar{H}}$

Quindi $\Pr[p(X) \geq 2^{-nt}] \to 1$ se $t > \bar{H}$, $\to 0$ altrimenti (se $t < \bar{H}$)