Divergenza informazionale

Due d.d.p. $p = (p_1, \dots, p_K) \in \mathbb{R}^K$ e $q = (q_1, \dots, q_K) \in \mathbb{R}^K$

$$D(p\|q) = \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i}$$

Convessità

per ogni $\lambda \in [0,1]$ e $x_1, x_2 \in (a,b)$

$f : \mathbb{R} \to \mathbb{R}$ è convesse in (a,b) se $f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$

strettamente convessa se $f(\lambda x_1 + (1-\lambda) x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$

per ogni $\lambda \in (0,1)$ e $x_1, x_2 \in (a,b)$ $(x_1 \neq x_2)$

Se $f''(x) \geq 0$ per ogni $x \in (a,b)$ allora f è convessa in (a,b)

Se $f''(x) > 0$ per ogni $x \in (a,b)$ allora f è strettamente convessa in (a,b)

Disuguaglianza di Jensen

Se f è una funz. convessa da $\mathbb{R}$ in $\mathbb{R}$ e X è una v.a. a valori reali, allora

$p_1, p_2, \dots, p_k$ ← probabilità

$$\circledast \quad \mathbb{E} f(X) \geq f(\mathbb{E}X)$$

$X \in \{x_1, x_2, \dots, x_k\}$ ← valori

(Inoltre, se f è strettamente convessa, si ha $\mathbb{E}f(X) = f(\mathbb{E}X)$ solo se la v.a. X è costante.)

Dim. Se $k=1$: banale $\overbrace{p_1 f(x_1)}^{\mathbb{E}f(X)} = \overbrace{f(p_1 x_1)}^{f(\mathbb{E}X)}$ $\quad k=1 \Rightarrow p_1 = 1$

Se $k=2$: $p_1 f(x_1) + p_2 f(x_2) \overset{?}{\geq} f(p_1 x_1 + p_2 x_2)$

Prendo $\lambda = p_1$, $1-\lambda = p_2$ $\quad (p_1 + p_2 = 1)$

$$\circledast \quad \sum_{i=1}^{k} p_i f(x_i) \geq f\left(\sum_{i=1}^{k} p_i x_i\right)$$

$\lambda f(x_1) + (1-\lambda) f(x_2) \overset{?}{\geq} f(\lambda x_1 + (1-\lambda) x_2)$

vero per la convessità di f.

Se $K > 2$, per induzione

$$\mathbb{E} f(X) = \sum_{i=1}^{k} p_i f(x_i) \overset{?}{\geq} f\left(\sum_{i=1}^{k} p_i x_i\right)$$

$$\sum_{i=1}^{k} p_i f(x_i) = p_K f(x_K) + (1-p_K) \sum_{i=1}^{K-1} \underbrace{\frac{p_i}{1-p_K}}_{q_i} f(x_i) \geq \text{(per induzione)}$$

$$\sum_{i=1}^{k-1} q_i = \frac{\sum_{i=1}^{k-1} p_i}{1-p_k} = \frac{1-p_k}{1-p_k} = 1$$

$$\geq p_K f(x_K) + (1-p_K) f\left(\sum_{i=1}^{k-1} q_i x_i\right) \geq \text{(per convessità di } f\text{)}$$

$$\geq f\left(p_k x_K + \cancel{(1-p_K)} \sum_{i=1}^{k-1} \frac{p_i}{\cancel{1-p_K}} x_i\right) = f\left(\sum_{i=1}^{K} p_i x_i\right) = f(\mathbb{E}X) \qquad \square$$

Disuguaglianza della somma logaritmica.

Se $a_1, \dots, a_K \geq 0$, $b_1, \dots, b_K > 0$ (non tutti zero)

allora $$\sum_{i=1}^{K} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{K} a_i\right) \log \frac{\sum_{i=1}^{K} a_i}{\sum_{i=1}^{K} b_i}$$

(con la convenzione che $0 \cdot \log 0 = 0$, giustificata dal fatto che $\lim_{x \to 0^+} x \log x = 0$)

Dim. Considerarmo $f(x) = x \log x$ nell'intervallo $(0, +\infty)$

$f'(x) = x \cdot 1/x + 1 \cdot \log x = 1 + \log x$

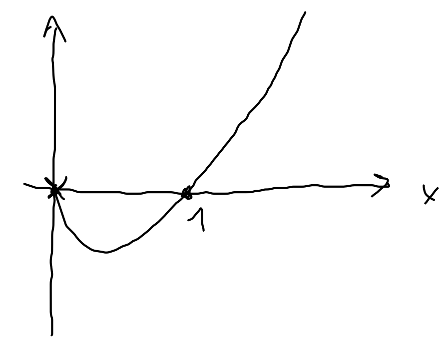
$f''(x) = 1/x > 0$ per ogni $x \in (0, +\infty)$

$\to f$ è strettamente convessa

$x_i \stackrel{def}{=} \frac{a_i}{b_i}$ per $i = 1, \dots, K$

$$\sum_i p_i f(x_i) \geq f\left(\sum_i p_i x_i\right)$$

(Jensen)



$p_i \stackrel{def}{=} \frac{b_i}{\sum_{i=1}^{K} b_i} \geq 0$ $\left(\to \sum_{i=1}^{K} p_i = 1\right)$

$$\sum_i \overset{p_i}{\boxed{\frac{\cancel{b_i}}{\sum_{j=1}^{k} b_j}}} \; \overset{f(x_i)}{\boxed{\frac{a_i}{\cancel{b_i}} \log \frac{a_i}{b_i}}} \geq f\left(\sum_i \frac{\cancel{b_i}}{\sum_j b_j} \frac{a_i}{\cancel{b_i}} \right)$$

$$= \frac{1}{\cancel{\sum_i b_i}} \sum_i a_i \log \frac{a_i}{b_i} \geq f\left(\frac{\sum_i a_i}{\sum_j b_j} \right) = \frac{(\sum_i a_i)}{\cancel{\sum_i b_i}} \log \left(\frac{\sum_i a_i}{\sum_i b_i} \right) . \quad \square$$

(Inoltre, nella disuguaglianza della somma logaritmica ho uguaglianza se e solo se a_i / b_i = costante $\left(a_1/b_1 = a_2/b_2 = \dots = a_k/b_k \right)$)

Disuguaglianza di Gibbs:

$D(p\|q) \geq 0$ per ogni coppia di d.d.p $p, q \in \mathbb{R}^k$

$$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \geq (\overbrace{\sum_i p_i}^{1}) \log \left(\frac{\overbrace{\sum_i p_i}^{1}}{\underbrace{\sum_i q_i}_{1}} \right) = 1 \cdot \log \frac{1}{1} = 0 .$$

$D(p\|q) = 0$ solo se p_i / q_i = costante

$\rightarrow$ solo se $p = q$.

$D(p\|q)$ può divergere a $+\infty$ se $\exists i$ tale che $p_i > 0$ ma $q_i = 0$

In generale, $0 \leq D(p\|q) \leq +\infty$ $\begin{pmatrix} 0 & \text{se } p = q \\ +\infty & \text{se } \exists i \text{ tale che } p_i > 0,\ q_i = 0 \end{pmatrix}$

In generale, $D(p\|q) \neq D(q\|p)$

Due v.a. X, Y si dicono indipendenti se

$$\underbrace{p_{XY}}(x,y) = \underbrace{p_X(x) \cdot p_Y(y)} \quad \text{per ogni } x \in \mathcal{X}, y \in \mathcal{Y}$$

La quantità

$$\underset{\| \text{def}}{D(p_{XY} \| p_X \cdot p_Y)} = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{XY}(x,y) \cdot \log \frac{p_{XY}(x,y)}{p_X(x) \cdot p_Y(y)} = \sum_{\substack{x \in \mathcal{X},\\ y \in \mathcal{Y}}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$I(X;Y)$ si chiama mutua informazione tra X e Y

Osserviamo che:

1) $I(X;Y) \geq 0$ (per la disug. di Gibbs)

2) $I(X;Y) = I(Y;X)$

$$\begin{array}{ccc} I(X;Y) & = & I(Y;X) \\ \| & & \| \\ D(p_{XY} \| p_X \cdot p_Y) & = & D(p_{YX} \| p_Y \cdot p_X) \end{array}$$

3) $I(X;Y) = 0$ se e solo se X e Y sono indipendenti

$$\begin{array}{c|cccc} & x_1 & & \cdots & x_K \\ \hline x_1 & \boxed{p_1} & 0 & 0 & 0 \\ x_2 & 0 & \boxed{p_2} & 0 & 0 \\ \vdots & 0 & 0 & \boxed{p_3} & 0 \\ x_K & 0 & 0 & 0 & \boxed{p_K} \end{array}$$

Quanto vale $I(X;X)$?

$$I(X;X) = \sum_{x' \in \mathcal{X},\, x'' \in \mathcal{X}} p(x', x'') \log \frac{p(x', x'')}{p(x')\, p(x'')} =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} = \sum_{i=1}^{K} p_i \log \frac{\cancel{p_i}}{p_i^{\cancel{2}}} = \overbrace{-\sum_{i=1}^{k} p_i \log p_i}^{H(X)}$$

entropia di X

$$H(X) \geq 0 \qquad (\text{poiché } I(X;X) \geq 0)$$

Inoltre, $H(X) = 0$ se e solo se X è costante

$$\Downarrow$$

$$x_1, \ldots, x_k$$

$$p_X = (0, 0, 1, 0, \ldots, 0)$$

$$\begin{aligned} H(X) &= -\sum_i p_i \log p_i \\ &= -0 \log 0 - 0 \log 0 - \ldots - 1 \log 1 - 0 \log 0 \\ &= 0 \end{aligned}$$

$$H(X) = \mathbb{E}_{X \sim p_X}\Big[\overbrace{-\log p(X)}^{\text{autoinformazione}}\Big] = \sum_i p_i(-\log p_i)$$

$$H(X) = 0 = + \sum_i \overbrace{p_i}^{>0} \underbrace{\log \overbrace{\frac{1}{p_i}}^{>1}}_{>0}$$

$\leftarrow$ possibile solo se X = costante

$H_2(X) = -\sum_i p_i \log_2 p_i$ entropia misurata in <u>bit</u>

$H_e(X) = -\sum_i p_i \log_e p_i$ " " " <u>nat</u>