① Sorpresa como función de probabilidad.

Para un resultado x al que un modelo asigna una probabilidad p, se define su sorpresa como

$$S(p) \quad : p \in (0,1]$$

② Axiomas:

- Monotonía: Si un evento es menos probable, sorprende más.

$$p_1 < p_2 \longrightarrow S(p_1) > S(p_2)$$

- Cero sorpresa para lo seguro:

$$p = 1 \longleftrightarrow S(p) = 0$$

- Aditividad para eventos independientes:

$$S(p_1 p_2) = S(p_1) + S(p_2)$$

- Regularidad suave:

pequeños cambios en p no deben producir grandes cambios de sorpresa.

$$S(p) = -\log p$$

③ Entropía de Shannon:

Medida de "costo"/incertidumbre

$$H(p) = \mathbb{E}[S(p)] = -\sum_i p_i \log p_i$$

El costo promedio óptimo de describir muestras $x \sim P$ usando la expresión de $p(x)$ es:

$$\mathbb{E}_{x\sim p}[-\log p(x)]$$

Ahora, si los datos vienen de P, pero se modela respecto a Q

$$\mathbb{E}_{x\sim p}[-\log q(x)]$$

Entonces, la diferencia de describir a x con Q cuando $x \sim P$ es

$$\mathbb{E}_{x\sim p}[-\log q(x)] - \mathbb{E}_{x\sim p}[-\log p(x)]$$

$$= \mathbb{E}_{x\sim p}[-\log q(x)] + \mathbb{E}_{x\sim p}[\log p(x)]$$

$$= \mathbb{E}_{x\sim p}[-\log q(x) + \log p(x)]$$

$$= \mathbb{E}_{x\sim p}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] \triangleq D_{KL}(P \| Q)$$