

Similarity measures between distributions



Claude
Shannon

- Shannon postulated that any measure of the informativeness of an event, x , should satisfy three conditions:
 1. An event with probability 1 yields no information
 2. The probability of an event and the information it yields vary inversely with each other
 3. The total information coming from independent events is purely additive
- Which he used to define self-information: $I(x) = -\log P(x)$
- Shannon entropy: $H(P) = \mathbb{E}_{x \sim P}[I(x)] = -\mathbb{E}_{x \sim P}[\log P(x)]$
- Kullback-Leibler (KL) divergence: $D_{KL}(P || Q) = \mathbb{E}_{x \sim P} \left[\log \frac{P(x)}{Q(x)} \right]$
- Cross entropy: $H(P, Q) = H(P) + D_{KL}(P || Q) = -\mathbb{E}_{x \sim P}[\log Q(x)]$

Maximum likelihood estimation

- Estimates the parameters, θ , of a distribution using a likelihood function, $\mathcal{L}(\mathbf{D}; \theta)$, given some data, \mathbf{D} .

- We maximize the likelihood function by minimizing its -logarithm:

$$\mathcal{L}(\theta | \mathbf{D}) = P(\mathbf{D}; \theta)$$

$$= \left(\prod_{i=1}^M P(\mathbf{y}_i | \mathbf{x}_i; \theta) \right)^{\frac{1}{M}}$$

$$= \frac{1}{M} \sum_{i=1}^M \log P(\mathbf{y}_i | \mathbf{x}_i; \theta)$$

Note: technically natural log, but true to within a constant

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^M -\log P(\mathbf{y}_i | \mathbf{x}_i; \theta)$$

- This expresses an optimization problem; the form of $p(\mathbf{D}; \theta)$ dictates how we solve it
 - In deep learning, this function is a neural network; we compute its gradient w.r.t. θ , and then estimate $\hat{\theta}$ using stochastic gradient descent (SGD).