Information Theory 101

Consider a source producing a discrete random variable x assuming values $\{1, 2, 3, 4 \ldots c\}$ with probabilities $p(x=k)$ for $1 \leq k \leq c$.

Information theory deals with a quantification of the information in this source.

Axiom : The information carried by an instance $x = k$ is

$$\log \frac{1}{p(x=k)} = 0$$

we know it is $x = k$ and hence 1, carries no information

But if the prob is a small number, it carries a lot of information.

Intuition → The higher the probability of x, the less is its 'uncertainity', and thus its information is smaller as well. Why log? Because the information in two independent event should be their sum.

The entropy of this source is defined by expected information

$$\boxed{H(x) = E_x\left(\log_2 \frac{1}{p(x)}\right) = \sum_{k=1}^{c} p(x=k) \log_2 \frac{1}{p(x=k)}}$$

$$\boxed{0 \leq H(x) \leq \log_2 c}$$

The notion of entropy can also be extended to continuous random variables.

Consider a source producing continuous random vectors $x \in \mathbb{R}^n$ with PDF $p(x)$.

The Differential Entropy of a random vector $x \sim p(x)$ is given by the expected information,

$$H(x) = E_x\left(\log \frac{1}{P(x)}\right) = \int P(x) \log \frac{1}{P(x)}\, dx$$

$\therefore$ This may assume negative values, as $P(x)$ can be greater than one.

→ Assume that $x \in R$ is a random variable with mean μ and variance σ^2. Then among all the possible PDF's of x, the gaussian distribution yields the maximal differential entropy.

$$\boxed{H(x) = \frac{1}{2} \log (2\pi e \sigma^2)}$$

☆ KL divergence

offers an asymmetric 'distance' measure between two distribution, $P(x)$ and $Q(x)$:

$$\boxed{KL(P \| Q) = E_{x\sim P}\left(\log \frac{P(x)}{Q(x)}\right) = \int P(x) \log \frac{P(x)}{Q(x)}\, dx}$$

→ $KL(P\|Q) \geq 0$ and $KL(P\|Q) \neq KL(Q\|P)$

→ $P(x) = Q(x)$ then $KL(P\|Q) = 0$

→ for $P(x) = N(x; \mu_p, \sigma_p)$ and $Q(x) = N(x; \mu_q, \sigma_q)$

$$\boxed{KL(P\|Q) \propto (\mu_p - \mu_q)^T \sigma_q^{-1} (\mu_p - \mu_q) + \log \frac{|\sigma_q|}{|\sigma_p|} + tr(\sigma_q^{-1}\sigma_p) - n}$$

→ For two random vectors $x, z \in R^n$ with a joint PDF, their Mutual information is defined by

$$I(x;z) = \int P(x,z) \log \frac{P(x,z)}{P(x)P(z)}\, dx = KL\left(P(x,z) \| P(x)P(z)\right)$$

$I(x;z)$ quantifies how dependent these two random vector are.

Few properties:

→ $x, z \in R^n$ are independent then $I(x;z) = 0$

→ $I(x;z) = I(z;x)$ is a symmetric function

→ Lower and upper bounds : $0 \leq I(x;z) \leq \min(H(x), H(z))$

→ If $z = f(x)$ where $f(\cdot)$ is a deterministic function,

$$\boxed{I(x;z) = H(x)}$$

An important alternative to KL-div is the wasserstein's distance $W_2(P(x), Q(x))$.

$W_2(P(x), Q(x))$ between two distribution, $P(x)$ and $Q(x)$, is given by:

$$W_2(P(x), Q(x)) = \inf_{G(x,z)} \int_x \int_z \|x-z\|_2^2 \, G(x,z) \, dx \, dz$$

where $P(x) = \int_z G(x,z) dz$ and $Q(z) = \int_x G(x,z) dx$

for two gaussians:

$$\boxed{W_2(N(x; \mu_p, \sigma_p), N(x; \mu_q, \sigma_q)) = \|\mu_p - \mu_q\|_2^2 + \text{trace}\left(\sigma_p + \sigma_q - 2\sqrt{(\sigma_p \sigma_q)}\right)}$$