Information Theory

1 Entropy

Entropy quantifies the uncertainty of a probability distribution—lower uncertainty means lower entropy, while higher uncertainty means higher entropy.

The formula for **information content**:

$$I(x) = -\log P(x)$$

2 Information Entropy

Information Entropy and **Entropy** refer to the same concept, differing only in naming across disciplines.

Information Entropy is computed as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

where $p(x_i)$ is the probability that the random variable X takes the value x_i. **Information Entropy is the weighted average of information content.**

3 Relative Entropy or KL Divergence

Relative Entropy or **KL divergence** is used to measure the difference between two probability distributions. (**Non-symmetric measure**)

3.1 Mathematical Representation

The formula for KL divergence is shown below:

Discrete form:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

Continuous form:

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

3.2 Derivation

For the same random variable X, if **the true probability distribution is** $p(x)$ and **the predicted distribution is** $q(x)$, the difference between their information entropies can be used to quantify the divergence between these two distributions.

Entropy of true distribution:

$$H(p) = -\sum_{x} p(x) \log p(x)$$

Entropy of predicted distribution:

$$H(q) = -\sum_{x} p(x) \log q(x)$$

KL divergence:

$$D_{\mathrm{KL}}(P \parallel Q) = H(q) - H(p) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

3.3 Properties

Non-negativity:

$$D_{\mathrm{KL}}(P \parallel Q) \geq 0$$

Asymmetry:

$$D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$$

4 Cross Entropy

Cross-entropy measures the difference between the **predicted distribution** $Q(X_i)$ and the **true probability distribution** $P(X_i)$ of the same random variable X.

Discrete form:

$$H(P,Q) = \sum_{x} p(x) \log p(x) - \sum_{i=1}^{n} p(x_i) \log q(x_i) = constant - \sum_{i=1}^{n} p(x_i) \log q(x_i) = -\sum_{i=1}^{n} p(x_i) \log q(x_i)$$

The derivative of a constant is zero, therefore constants can be disregarded.

Continuous form:

$$H(P,Q) = -\int_{X} p(x) \log q(x) \, dr(x)$$

The more accurate the prediction, the smaller the cross-entropy.