Fundamentals of Information Theory

1. Information: Degree of surprise after observing x.

$$h(x) = -\log_2 p(x) = \log_2 \frac{1}{p(x)}$$

Base of the log determines the unit of information. (2: bits, e: nats)

2. Entropy: Entropy is a measure of the uncertainty of a random variable.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Entropy is also defined as expected amount of information.

$$H(X) = \mathbf{E}[-\log_2 p(x)]$$

Entropy in the continuous domain is called the differential entropy:

$$H[X] = -\int p(x) \log p(x) dx$$

Differential entropy is typically measures in nats.

3. Joint Entropy: A measure of the uncertainty associated with a set of variables.

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, Y)$$

4. Conditional Entropy: The entropy of a random variable conditioned on another one.

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x)$$

If X and Y are independent:

$$H(X|Y) = H(X)$$

Conditional entropy is lower if random variables are dependent.

From the chain rule, joint entropy can be written as:

$$H(X, Y) = H(X) + H(Y|X)$$

General Form:

$$H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_2|X_1, ..., X_{n-1})$$

5. Mutual Information: Measures mutual dependence between two random variables

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Since $H(X|X) = 0$:

$$H(X) = H(X) - H(X|X) = I(X, X)$$

6. Pointwise Mutual Information: A measure of association

$$I(x, y) = \log \frac{p(x, Y)}{p(x)p(Y)}$$

7. Relative Entropy - Kullback-Leibler Divergence:
Measures the difference between two probability distribution.

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Since $D(p||q) \neq D(q||p)$, KL divergence is not defined as a metric.

8. Conditional Relative Entropy

$$D(p(y|x)||q(y|x)) = \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log \frac{p(y|x)}{q(y|x}$$

From chain rule:

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

9. Cross Entropy Measures entropy of a distribution p, under another distribution q.

$$H(X,q) = H(X) + D(p||q)$$

10. Perplexity A measurement of how well a probability distribution or probability model predicts a sample

$$Perplexity(W,m) = 2^{H(W,m)}$$