Let A and B be finite sets $\{a_1, a_2...a_{N_A}\}$ and $\{b_1, b_2...b_{N_B}\}$

Let X and Y be discrete random variables with support A and probability measures P_x and P_y respectively.

Let Z be a discrete random variable with support B and probability measure P_z

Information

Information is a way of measuring changes in uncertainty. For example, when the outcome of a stochastic event is observed, the observer moves from a state of uncertainty to a state of certainty regarding the outcome of the event. Equivalently, we can say that the observer learned some information about the event. For an observation a from discrete random variable X the information associated with the observation is:

$$I(a) = -log(P_x(a)) \tag{1}$$

The units used to measure information are determined by the base of the logarithm in equation 1. The most common units are bits (base 2) and nats (base e).

Entropy

Entropy is a measure of the uncertainty associated with a random variable or probability distribution[1]. Equivalently, entropy measures the expected amount of information generated by sampling from a random variable or probability distribution. The entropy of random variable X is:

$$H(X) = \mathbb{E}_{P_x}(I(X)) = -\sum_{a \in A} P_x(a) log(P_x(a)) \tag{2}$$

Joint Entropy

The joint entropy of two variables is the entropy of their joint probability distribution. For variables X and Z with joint probability distribution $P_{xz}(a, b)$:

$$H(X, Z) = -\sum_{a \in A, b \in B} P_{xz}(a, b) log(P_{xz}(a, b)) \tag{3}$$

The two variables are independent iff:

$$H(X, Z) = H(X) + H(Z) \tag{4}$$

[1]For an axiomatic definition of entropy see Pattern Theory Ch1. Exercise 2.

Conditional Entropy

Conditional entropy measures the expected amount of information generated by observing a random variable when the outcome of a second random variable is fixed. It is defined as the expectation, with respect to the second variable, of the entropy of the conditional distribution of the first variable.

$$H(X|Z) = \mathbb{E}_{P_z}(H(X|Z=b)) = \sum_{a\in A, b\in B} P_{xz}(a,b) log \frac{P_z(b)}{P_{xz}(a,b)} = H(X,Z) - H(Z) \tag{5}$$

The two variables are independent iff:

$$H(X|Z) = H(X) \tag{6}$$

Relative Entropy

Relative entropy, or Kullback Leibler (KL) divergence, is a measure of how different one probability distribution is from another reference probability distribution[2]. It measures the information gained by using an informed distribution instead of a naive distribution, or the information lost by using a naive distribution to approximate an informed distribution. KL divergence is defined as the expected information difference between two distributions where the expectation is taken with respect to the reference distribution. Taking P_x as the reference distribution, the KL divergence from P_x to P_y is:

$$D_{KL}(P_x||P_y) = \mathbb{E}_{P_x}\left(log\frac{P_x}{P_y}\right) = \sum_{a\in A} P_x(a) log \frac{P_x(a)}{P_y(a)} \tag{7}$$

The **mutual information** of two variables X and Z is the KL divergence from their joint distribution $P_{xz}(a,b)$ to the cartesian product of their marginal distributions P_x and P_z. The variables are independent iff their mutual information is null:

$$I(X;Z) = I(Z;X) = D_{KL}(P_{xz}(a,b)||P_x(a)\otimes P_z(b)) = \sum_{a\in A b\in B} P_{xz}(a,b) log \frac{P_{xz}(a,b)}{P_x(a)P_z(b)} \tag{8}$$

Differential Entropy

All of the concepts above can also be extended to continuous random variables with probability density functions. For a probability density function $P_d(x)$ over X, differential entropy measures the amount of information carried by P_d relative to a reference measure (e.g. the Lebesgue measure on $\mathbb{R}^N$):

$$H_d(P_d) = -\int_X P_d(x) log(P_d(x)) d\mu(x) \tag{9}$$

Where μ is the reference measure on the support X of P_d.

[2]KL divergence is not symmetric and is therefore not a true distance metric. This means that it matters which distribution is used as the reference. In the context of statistics, the reference distribution is almost always the more informed or "correct" probability distribution.