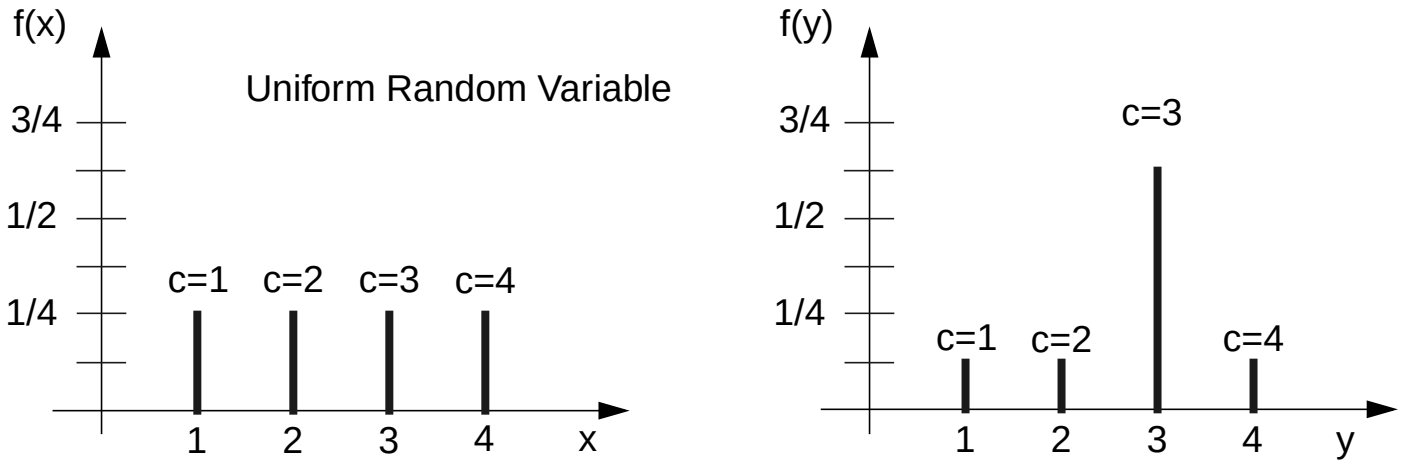
What Is Information? (When not to bet at a casino...)

Consider two distributions of discrete random variables:



Which variable is more unpredictable?

Now, consider sampling random numbers from a random number generator whose statistics are not known. The more numbers we draw, the more we discover about the underlying distribution. Assuming the underlying distribution is from one of the above distributions, how much more information do we receive with each new number we draw?

The answer lies in the shape of the distributions. For the random variable **x**, each class is equally likely. Each new number we draw provides the maximum amount of information, because, on the average, it will be from a different class (so we discover a new class with every number). On the other hand, for **y**, chances are, **c=3** will occur 5 times more often than the other classes, so each new sample will not provide as much information.

We can define the information associated with each class, or outcome, as:

$$I(c = \hat{c}) \equiv \log_2 \frac{1}{P(c = \hat{c})} = -\log_2 P(c = \hat{c})$$

Since $0 \le P(x) \le 1$, information is a positive quantity. A base 2 logarithm is used so that K discrete outcomes can be measured in $2^M = K$ bits. For the distributions above,

$$I_x(c = 1) = (-1)\log_2 (1/4) = 2\ bits \qquad I_y(c = 1) = (-1)\log_2\left(\frac{1}{8}\right) = 3\ bits$$

Huh??? Does this make sense?

What is Entropy?

Entropy is the expected (average) information across all outcomes:

$$H(c) \equiv E[I(c)] = -\sum_{k=1}^{K} P(c = c_k) \log_2 P(c = c_k)$$

Entropy using $\log_2$ is also measured in bits, since it is an average of information.

For example,

$$H_x = -\left[\sum_{k=1}^{4} \left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right)\right] = 2.0 \text{ bits} \qquad H_y = -\left[\sum_{k=1}^{3} \left(\frac{1}{8}\right) \log_2 \left(\frac{1}{8}\right) + \left(\frac{5}{8}\right) \log_2 \left(\frac{5}{8}\right)\right] = 0.8 \text{ bits}$$

We can generalize this to a joint outcome of N random vectors from the same distribution, which we refer to as the *joint entropy*:

$$H[\bar{x}(1),\ldots,\bar{x}(N)] = -\sum_{l_1=1}^{N} \cdots \sum_{l_N}^{N} P[(\bar{x}(1) = \bar{x}_{l_1}), \ldots, (\bar{x}(N) = \bar{x}_{l_N})] \times \log_2 P[(\bar{x}(1) = \bar{x}_{l_1}), \ldots, (\bar{x}(N) = \bar{x}_{l_N})]$$

If the random vectors are statistically independent:

$$H[\bar{x}(1),\ldots,\bar{x}(N)] = \sum_{n=1}^{N} H[\bar{x}(n)]$$

If the random vectors are independent and indentically distributed:

$$H[\bar{x}(1),\ldots,\bar{x}(N)] = NH[\bar{x}(1)]$$

We can also define conditional entropy as:

$$H(\bar{x}|\bar{y}) = \sum_{k=1}^{K} \sum_{l=1}^{L} P((\bar{x} = \bar{x}_l)\big|(\bar{y} = \bar{y}_k)) \log_2 [P((\bar{x} = \bar{x}_l)\big|(\bar{y} = \bar{y}_k))]$$

For continuous distributions, we can define an analogous quantity for entropy:

$$H = -\int_{-\infty}^{\infty} f(x) \log_2 f(x)\, dx \quad \text{(bits)}$$

A zero-mean Gaussian random variable has maximum entropy $(\frac{1}{2}\log_2(2\pi e\sigma^2)$.

Why?

Mutual Information

The pairing of random vectors produces less information than the events taken individually. Stated formally:

$$I(\bar{x}, \bar{y}) \leq I(\bar{x}) + I(\bar{y})$$

The shared information between these events is called the mutual information, and is defined as:

$$M(\bar{x}, \bar{y}) \equiv [I(\bar{x}) + I(\bar{y})] - I(\bar{x}, \bar{y})$$

From this definition, we note:

$$\begin{aligned} M(\bar{x}, \bar{y}) &= \log_2 \left[\frac{P(\bar{x}, \bar{y})}{P(\bar{x})P(\bar{y})} \right] \\ &= \log_2 \left[\frac{1}{P(\bar{x})} \right] + \log_2 \left[\frac{1}{\frac{P(\bar{x}, \bar{y})}{P(\bar{y})}} \right] = \log_2 \left[\frac{1}{P(\bar{x})} \right] - \log_2 \left[\frac{1}{P(\bar{x}|\bar{y})} \right] \\ &= I(\bar{x}) - I(\bar{x}|\bar{y}) \\ &= I(\bar{y}) - I(\bar{y}|\bar{x}) \end{aligned}$$

This emphasizes the idea that this is information shared between these two random variables.

We can define the *average mutual information* as the expectation of the mutual information:

$$\begin{aligned} \overline{M}(\bar{x}, \bar{y}) &= E\left[\log_2 \left(\frac{P(\bar{x}, \bar{y})}{P(\bar{x})P(\bar{y})} \right) \right] \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L} P((\bar{x} = \bar{x}_l), (\bar{y} = \bar{y}_k)) \log_2 \left[\frac{P((\bar{x} = \bar{x}_l), (\bar{y} = \bar{y}_k))}{P(\bar{x} = \bar{x}_l)P(\bar{y} = \bar{y}_k)} \right] \end{aligned}$$

Note that:

$$\overline{M}(\bar{x}, \bar{y}) = H(\bar{x}) - H(\bar{x}|\bar{y}) = H(\bar{y}|\bar{x}) - H(\bar{y})$$

Also note that if $\bar{x}$ and $\bar{y}$ are independent, then there is no mutual information between them.

Note that to compute mutual information between two random variables, we need a joint probability density function.

Entropy in Pattern Recognition

Generalized entropy measures are used to assess the effectiveness of a set of features at pattern classification. The conditional entropy is one such measure:

$$H(c|\bar{x}) = \sum_{c=1}^{K} \sum_{l=1}^{L} P((c = \hat{c})|(\bar{x} = \bar{x}_l))\log_2[P((c = \hat{c})|(\bar{x} = \bar{x}_l))]$$

This is sometimes referred to as the equivocation. We want this measure to be small, meaning the feature vector $\bar{x}$ greatly reduces the uncertainty about the class identity.

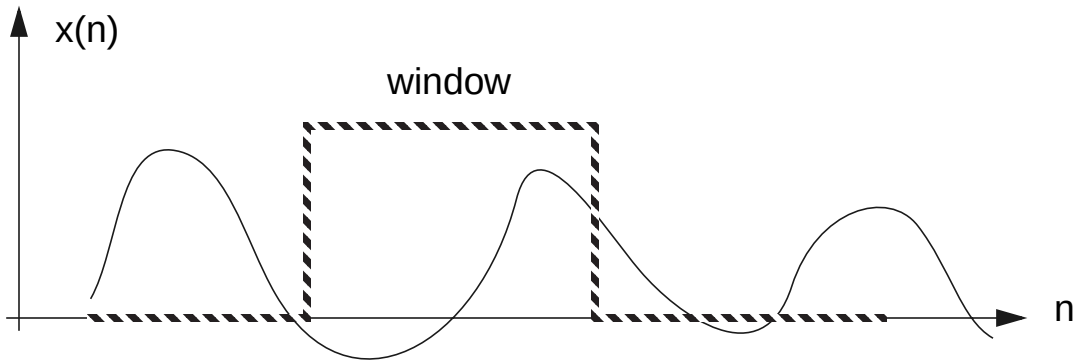
Another way to assess the usefulness of a feature is the average mutual information:

$$\overline{M}(c, \bar{x}) = \sum_{c=1}^{K} \sum_{l=1}^{L} P((c = \hat{c}), (\bar{x} = \bar{x}_l))\log_2\left[\frac{P((c = \hat{c}), (\bar{x} = \bar{x}_l))}{P(c = \hat{c})P(\bar{x} = \bar{x}_l)}\right]$$

If this measure is large, a given feature contains a significant information about the class outcome.

How Does Entropy Relate To DSP?

Consider a window of a signal:



What does the sampled z-transform assume about the signal outside the window?

What does the DFT assume about the signal outside the window?

How do these influence the resulting spectrum that is computed?

What other assumptions could we make about the signal outside the window? How many valid signals are there?

How about finding the spectrum that corresponds to the signal that matches the measured signal within the window, and has maximum entropy?

What does this imply about the signal outside the window?

This is known as the principle of maximum entropy spectral estimation. Later we will see how this relates to minimizing the mean-square error.