Entropy and Mutual Information

Zachary Yedidia

Introduction

We are interested in analyzing sensory processing in the brain using Efficient Coding Theory.

Entropy

Consider a discrete random variable X with sample space Ω of k events x

$$\Omega = \{x\}, \quad |\Omega| = k$$

where the probability an event x happens is $\Pr\{X = x\} = p_X(x)$. Here p_X is the probability mass function of X, and from now on $p_X(x) = p(x)$ for brevity. This means $p(x)$ and $p(y)$ represent the probability mass functions of two separate random variables.

Suppose we want to measure how surprising a certain event x is. If there is a high probability of x happening, then we are not very surprised if it occurs, and we don't receive much information as a result. Conversely, if $X = x$ has a small probability, the occurrence of the event gives us a lot of information since a surprising event has happened. The quantitative description of the *degree of surprise* is given by Shannon as

$$h(x) = \log\left(\frac{1}{p(x)}\right) = -\log p(x).$$

Notice that this exhibits the properties we want: when $p(x)$ is small, $h(x)$ is large and when $p(x)$ is large, $h(x)$ is small. Additionally, because of the logarithm, if X is a joint probability distribution made up of independent random variables such that $p(x) = p(y)p(z)$, the entropies add to give the entropy of the overall distribution: $h(x) = h(y) + h(z)$, which is a desirable outcome.

The entropy of $H(X)$ of a discrete random variable X is the degree of surprise averaged over the entire ensemble:

$$H(X) = \langle -\log p(x) \rangle = -\sum_{x \in \Omega} p(x) \log p(x).$$

Note that in computing this sum we use the convention that $0 \log 0 = 0$, which is justified by the limit $\lim_{x \to 0} x \log x = 0$.

The entropy of a probability distribution can be interpreted as a measure of the uncertainty of the distribution, or equivalently the amount of information needed to store X. When the base logarithm is 2, the information is measured in *bits*. When the base of the logarithm is e it is measured in *nats*. When quantifying information it makes most sense to use bits, so we will use base 2 for the most part. In physics entropy is defined in the same way using base e so certain definitions may instead use the natural logarithm.

Entropy of a Bernoulli random variable

Let X be a random variable with two possible events

$$\Pr\{X = 1\} = p$$
$$\Pr\{X = 0\} = 1 - p$$

Then the entropy of X is

$$H(X) = -p \log p - (1 - p) \log(1 - p)$$

If we plot the entropy as a function of p we can see some basic properties of the entropy function.

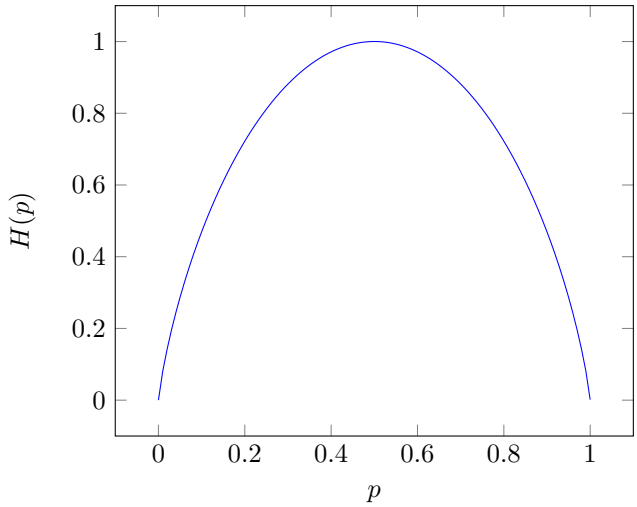

Figure 1: Entropy of a Bernoulli RV with parameter p

Notice that H is a concave function and if p is 0 or 1 we don't need any bits to store the result. This makes sense because the outcome is deterministically guaranteed. If $p = 0.5$ we need 1 bit to store the information (consider a fair coin toss).

Joint entropy and conditional entropy

Having defined the entropy of a single random variable, we can now examine the entropy of a pair of random variables (X, Y) each defined for the sample spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. This is not changing anything because a single random variable can be thought of as a vector of multiple random variables. Nonetheless, the *joint entropy* $H(X, Y)$ of a pair of discrete random variables X and Y with join distribution function $p(x, y)$ is

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

We can also define the specific conditional entropy of a random variable X given that Y is assigned a specific value y as

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} p(x|y) \log p(x|y).$$

Then we define the *conditional entropy* $H(X|Y)$ as averaging $H(X|Y=y)$ over all possible values that y may take:

$$\begin{aligned} H(X|Y) &= \langle H(X|Y=y) \rangle \\ &= \sum_{y \in \mathcal{Y}} p(y) H(X|Y=y) \\ &= -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y) \\ &= -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log p(x|y). \end{aligned}$$

This value is a quantitative measure of the uncertainty of X given that Y is known. It then follows intuitively that $H(X,Y) = H(X) + H(Y|X)$. This is called the *chain rule* and is proved below.

$$\begin{aligned} H(X,Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x) p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

Relative entropy and mutual information

The entropy of a random variable is a measure of the uncertainty in its distribution. The relative entropy $D(p\|q)$ between two distributions is a way of measuring the inefficiency of assuming the distribution is q when the real distribution is p. For example to represent a random variable with probability mass function $p(x)$ we would on average need $H(p)$ bits. If instead we used used the probability mass function $q(x)$, we would need $H(p) + D(p\|q)$ bits on average. The *relative entropy* (also known as the *Kullback Leibler distance*) between two probability mass functions $p(x)$ and $q(x)$ is defined as

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

This definition uses the convention that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. This is justified as before by continuity arguments.

We should note that the relative entropy is not a true measure of distance between two distributions because it is not symmetric, but it is often useful to still think of it as some sort of "distance."

Next we define mutual information, which is a measure of the amount of information one gains about a random variable Y by knowing another random variable X. It is the average reduction in the entropy of Y by an observation of X. If X and Y are random variables with joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$ then the *mutual information* $I(X;Y)$ is the relative entropy between the joint distribution and the distribution $p(x)p(y)$:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

or equivalently

$$I(X;Y) = D(p(x,y)\|p(x)p(y)).$$

The mutual information (unlike relative entropy) is a symmetric quantity:

$$I(X;Y) = I(Y;X).$$

By rewriting the definition of mutual information we can also reach the following conclusions (and their symmetric counterparts):

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X). \end{aligned}$$

The mutual information measures the information that X and Y share, so if the two random variables are independent then their mutual information will be 0, because knowing one does not give any information about the other.

Differential entropy

The concept of entropy can be extended for a continuous random variable X with probability density function $p(x)$. This means if we consider an interval $[a,b]$, the probability that an event x falls in this interval is $\int_a^b p(x)dx$. Suppose X has support $\mathcal{X}$, we now define the *differential entropy* of X as

$$H(X) = -\int_{\mathcal{X}} p(x)\log p(x)dx.$$

We assume that $p(x)$ exists and is normalized such that $\int_{-\infty}^{\infty} p(x)dx = 1$. Notice that like the discrete case the entropy on depends on the probability density function of the random variable, so it is sometimes written as $H(p)$ rather than $H(X)$.

Entropy of a Gaussian

To calculate the differential entropy of a Gaussian distribution $X \sim \mathcal{N}(\mu,\sigma)$ in nats, we compute

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \\ &= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2}\ln(2\pi e\sigma^2). \end{aligned}$$

It turns out that the Gaussian distribution is the maximum entropy distribution for a continuous random variable $X \in (-\infty,\infty)$ with fixed mean and variance. This result can be obtained by the use of Lagrange multipliers (not shown here).