Information Theory

Quan Wen

7 Nov 2018

Shannon Information

Shannon information can be formulated intuitively by playing the game "sixty-three". If you have a number in your hand, and this integer is randomly picked from 0 and 63 with equal probability. What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

Intuitively, Six questions suffice. One reasonable strategy ask the following questions:

- $x \geq 32$?
- is x mod $32 \geq 16$?
- is x mod $16 \geq 8$?
- is x mod $8 \geq 4$?
- is x mod $4 \geq 2$?
- is x mod $2 \geq 1$?

What are the Shannon information contents of the outcomes in this example? If we assume that all values of x are equally likely, then the answers to the questions are independent and each has Shannon information content $\log_2(1/0.5) = 1$bit; the total Shannon information gained is always $\log_2 64 = 6$ bits. Furthermore, the number x that we learn from these questions is a six-bit binary number. Our questioning strategy defines a way of encoding the random variable x as a binary file.

Now let us put everything in the context of sensory encoding. Given a sensory stimulus, such as flash light with given intensity, neurons in the retina, such as biopolar cells, can faithfully respond to the intensity change. Let us denote the response amplitude r, which could be membrane potential, or the firing rate of the neuron. Just by measuring the response of the neuron, how much information we could learn about the stimulus. The answer is just given by

$$h = -\log_2 P(r)$$

where $P(r)$ is the probability of a given response amplitude.

When we write down the above formula, we are assuming that all responses have equal probability. When responses have unequal probabilities, we modify the above formula, by computing the average over all different responses. This leads to the definition of Shannon information

$$I = -\sum_r P(r) \log_2 P(r) \tag{1}$$

When we treat the response of a neuron as a continuous variable, one should then define the probability density function $p(r)$, and the Shannon information becomes

$$I = -\sum_r p(r)\Delta r \log_2 p(r)\Delta r \tag{2}$$

Furthermore,

$$\lim_{\Delta r \to 0}(I + \log_2 \Delta r) = -\int dr p(r) \log_2 p(r) \tag{3}$$

We now ask, what is the input-output relationship of a neuron, $r = f(s)$, that would maximize the information (or entropy) of the response? Let's consider the simple case, by assuming that there is a maximum response r_{max} a neuron could achieve. This could be the maximum firing rate. And the minimum response is zero. It is now your homework to show that under this condition, different responses with equal probability would maximize the entropy. Now use the normalization condition

$$\int_0^{r_{max}} p(r)dr = 1,$$

we found that $p(r) = \frac{1}{r_{max}}$. Given the probability density distribution of inputs $p(s)$, One must have $p(s)ds = p(r)dr$. As a result, $\frac{dr}{ds} = r_{max}p(s)$, and the input-output function should be proportional to the cumulative probability distribution of stimulus

$$f(s) = \int_0^s r_{max}p(s')ds' \tag{4}$$

Mutual Information

As we discussed in previous lectures, the entropy of the neural response is given by

$$H_r = -\sum_r P(r) \log_2 P(r) \tag{5}$$

In early sensory processing, what we are really interested is how much useful information has been transmitted from the input to the output. Are all the bits

in H_r useful? Not really, some of the bits may reflect noise. Given the stimulus value s, the response can be variable, as defined by the conditional probability $P(r|s)$. Thus the noise entropy is given by

$$H_{noise} = -\sum_r P(r|s) \log_2 P(r|s)$$

When we averaged this over different stimulus, one has

$$\langle H_{noise} \rangle = -\sum_s \sum_r P(s)P(r|s) \log_2 P(r|s) \tag{6}$$

The useful information is the subtraction of noise entropy from the response entropy, called mutual information

$$\begin{aligned} I &= H_r - H_{noise} \\ &= -\sum_r P(r) \log_2 P(r) + \sum_s \sum_r P(s)P(r|s) \log_2 P(r|s) \\ &= -\sum_r \sum_s P(r|s)P(s) \log_2 P(r) + \sum_s \sum_r P(s)P(r|s) \log_2 P(r|s) \\ &= \sum_r \sum_s P(r|s)P(s) \log_2 \frac{P(r|s)}{P(r)} \end{aligned} \tag{7}$$

Recall that joint probability

$$P(r, s) = P(r|s)P(s),$$

the mutual information can then be expressed in a symmetric form between stimulus and response,

$$I = \sum_r \sum_s P(r, s) \log_2 \frac{P(r, s)}{P(r)P(s)} \tag{8}$$

Mutual information also provides a distance measure between two distributions, called K-L divergence, formally defined as

$$D_{KL}(P, Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \tag{9}$$

It is your homework to show that $D_{KL} \geq 0$ and $D_{KL} = 0$ only when $P = Q$.

Sensory encoding in multiple neurons

Let us now consider a two layer neural network. The output response in neuron y_i is given by

$$y_i = \sum_j W_{ij} x_j + \eta_i \tag{10}$$

Here x are random variables $\mathbf{x} = [x_1, x_2, ..., x_N]^T$ with zero means drawn from a multivariate gaussian distribution

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C^{-1}}\mathbf{x}\right) \tag{11}$$

where Z is the normalization factor. $\mathbf{C}$ is the covariance matrix of $\mathbf{x}$, satisfying $\mathbf{C} = \langle \mathbf{x}\mathbf{x^T} \rangle$. η_i is noise drawn also from a gaussian distribution, whose covariance matrix is given by $\langle \eta_i \eta_j \rangle = \sigma^2 \delta_{ij}$.

In this model, one can show (as your homework) that the mutual information has an analytical expression, which is given by

$$I(\mathbf{y}, \mathbf{x}) = \frac{1}{2}\mathrm{Tr}\big(\mathbf{I} + \frac{1}{\sigma^2}\mathbf{W}\mathbf{C}\mathbf{W^T}\big) \tag{12}$$

Now if we transform to a basis where $\mathbf{WCW^T}$ is diagonal, whose eigenvalues are given by λ_μ, then the mutual information reduces to

$$I(\mathbf{y}, \mathbf{x}) = \frac{1}{2}\sum_\mu \log_2\left(1 + \frac{\lambda_\mu}{\sigma^2}\right) \tag{13}$$

Clearly, larger D can overcome noise and increase mutual information. However, there is an energy cost. Here, let us set a constraint on the total output signal, that is

$$\begin{aligned}
\sum_i \langle y_i^2 \rangle &= \sum_i \sum_{jk} W_{ij} \langle x_j x_k \rangle W_{ik} + N\sigma^2 \\
&= \sum_i \sum_{jk} W_{ij} \langle x_j x_k \rangle W_{ki}^T + N\sigma^2 \\
&= \mathrm{Tr}(\mathbf{WCW^T}) + N\sigma^2 \\
&= \sum_\mu \lambda_\mu + N\sigma^2 \\
&= \mathrm{const}
\end{aligned} \tag{14}$$

One can therefore show that mutual information is maximized when all eigenvalues are equal. This indicates that $\mathbf{WCW^T} \propto \mathbf{I}$! As a result, mutual information is maximized by removing all correlations between inputs, and by making all outputs independent from each other, i.e., $\langle y_i y_j \rangle \propto \delta_{ij}$.

In the following, let us make additional assumptions that the transformation of inputs to outputs are spatially invariant. Consider inputs are photoreceptors lying on a two dimensional lattice. Receptor i is at position r_i. There is a linear mapping between the object positions and the receptor positions on the retina. The element of input covariance matrix only depends on the relative distance between the two neurons, $C_{ij} = f(\mathbf{r_i} - \mathbf{r_j})$, so do the synaptic weights

$W_{ij} = D(\mathbf{r_i} - \mathbf{r_j})$. This is like a two-dimensional lattice model. In this case, the condition of whitening reduced to

$$\sum_{km} D(\mathbf{r_k} - \mathbf{r_i}) f(\mathbf{r_k} - \mathbf{r_m}) D(\mathbf{r_m} - \mathbf{r_j}) \propto \delta_{ij} \tag{15}$$

In the continuous limit, we would like to impose the following condition

$$\int d^2\mathbf{r}' \int d^2\mathbf{r}'' D(\mathbf{r}'' - \mathbf{r_i}) f(\mathbf{r}'' - \mathbf{r}') D(\mathbf{r}' - \mathbf{r_j}) \propto \delta(\mathbf{r_i} - \mathbf{r_j}) \tag{16}$$

Using Fourier transform,

$$f(\mathbf{r_k} - \mathbf{r_i}) = \int d^2\mathbf{k} e^{-i\mathbf{k}\cdot(\mathbf{r_k} - \mathbf{r_i})} \tilde{f}(\mathbf{k}) \tag{17}$$

$$D(\mathbf{r} - \mathbf{r_i}) = \int d^2\mathbf{k} e^{-i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r_i})} \tilde{D}(\mathbf{k}) \tag{18}$$

$$D(\mathbf{r} - \mathbf{r_j}) = \int d^2\mathbf{k} e^{-i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r_j})} \tilde{D}(\mathbf{k}) \tag{19}$$

and plugging the above equations into the whitening conditions, it is your homework to show that

$$\int d^2\mathbf{k} |\tilde{D}(\mathbf{k})|^2 \tilde{f}(\mathbf{k}) e^{-i\mathbf{k}\cdot(\mathbf{r_i} - \mathbf{r_j})} \propto \delta(\mathbf{r_i} - \mathbf{r_j}) \tag{20}$$

This implies that

$$|\tilde{D}(\mathbf{k})| \propto \frac{1}{\sqrt{\tilde{f}(\mathbf{k})}} \tag{21}$$