Information Theory

Quan Wen

17 Dec 2019

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 $x \bmod 32 \geq 16$?

- is $x \bmod 16 \geq 8$?

- is $x \bmod 8 \geq 4$?

- is $x \bmod 4 \geq 2$?

- is $x \bmod 2 \geq 1$?

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

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

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 $P = Q$.

Mutual information of Gaussian variable

Consider a gaussian random variable x drawn from the distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \langle x \rangle)^2}{2\sigma^2}\right] \tag{10}$$

Let us compute its entropy

$$S = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) \tag{11}$$

To compute it explicitly, we have

$$S = \frac{1}{\ln 2} \left[\ln \sqrt{2\pi\sigma^2} + \int p(x) \frac{(x - <x>)^2}{2\sigma^2} \right] \tag{12}$$

$$= \frac{1}{\ln 2} \left[\frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \right] \tag{13}$$

$$= \frac{1}{\ln 2} \frac{1}{2} \left[\ln(2\pi\sigma^2) + 1 \right] \tag{14}$$

$$= \frac{1}{\ln 2} \frac{1}{2} \ln(2\pi e \sigma^2) \tag{15}$$

$$= \frac{1}{2} \log_2(2\pi e \sigma^2) \tag{16}$$

Now let us consider a very simple linear model

$$y = wx + \xi \tag{17}$$

Here ξ is noise with mean $\langle x \rangle$ zero and variance σ_ξ^2.

$$I(x, y) = S(y) - \langle S(y|x) \rangle_x \tag{18}$$

$$= S(y) - \langle S(wx + \xi|x) \rangle_x \tag{19}$$

$$= S(y) - \langle S(\xi|x) \rangle_x \tag{20}$$

$$= S(y) - S(\xi) \tag{21}$$

$$= \frac{1}{2} \log_2 \frac{\sigma_y^2}{\sigma_\xi^2} \tag{22}$$

$$= \log_2 \frac{\sigma_y}{\sigma_\xi} \tag{23}$$

In addition, we know that

$$\sigma_y^2 = w^2 \sigma_x^2 + \sigma_\xi^2 \tag{24}$$

Therefore,

$$I(x, y) = \frac{1}{2} \log_2 \left(1 + \frac{w^2 \sigma_x^2}{\sigma_\xi^2} \right) \tag{25}$$

Since we can rewrite

$$y = w(x + \xi_{eff}) \tag{26}$$

where $\xi_{eff} = \xi/w$. We thus have $\langle \xi_{eff}^2 \rangle = \sigma_\xi^2/w^2$. Define the signal to noise ratio as

$$\mathbf{SNR} = \frac{\langle x^2 \rangle}{\langle \xi_{eff}^2 \rangle} \tag{27}$$

4

The mutual information

$$I(x, y) = \frac{1}{2} \log_2(1 + \mathbf{SNR}) \tag{28}$$

We can also generalize the above case to multivariate Gaussian distribution. Consider the model

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \xi \tag{29}$$

Let us define the covariance matrices for the output and the noise as

$$\mathbf{C_Y} = \langle \mathbf{y}\mathbf{y}^T \rangle \tag{30}$$
$$\mathbf{C}_\xi = \langle \xi \xi^T \rangle \tag{31}$$

The mutual information is now given by

$$I(x, y) = \frac{1}{2} \log_2 \frac{|\mathbf{C_Y}|}{|\mathbf{C}_\xi|}, \tag{32}$$

where $|\mathbf{C}|$ is the determinant of the covariance matrix.

Sensory encoding in multiple neurons

With all the above preparations, we could now consider a very interesting feedforward model. The output response in neuron y_i is given by

$$y_i = \sum_j W_{ij}(x_j + z_j) + \eta_i \tag{33}$$

Here x may be viewed as the voltage signal from the photoreceptor, and we are measuring the response in the biopolar cell in the retina. $\mathbf{x} = [x_1, x_2, ..., x_N]^T$ with zero means drawn from a multivariate gaussian distribution

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x}\right) \tag{34}$$

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 $\langle \eta_i \eta_j \rangle = \sigma_\eta^2 \delta_{ij}$. z_j is the input noise, and $\langle z_i z_j \rangle = \sigma_z^2 \delta_{ij}$.

By defining a new effective noise

$$\xi_i = \sum_j W_{ij} z_j + \eta_i, \tag{35}$$

we can now use the formula above, which gives us

$$I = \frac{1}{2} \log_2 \frac{|\mathbf{C_Y}|}{|\mathbf{C}_\xi|}. \tag{36}$$

Maximizing the mutual information may be achieved by boosting the signal. However, there is an energy cost. Here let us set a constraint on the total output signal, that is

$$\sum_i \langle y_i^2 \rangle = \text{Tr}\langle \mathbf{y}\mathbf{y}^T \rangle \tag{37}$$

Thus our job is to maximize the following objective function

$$F = \frac{1}{2} \log_2 \frac{|\mathbf{C_Y}|}{|\mathbf{C_\xi}|} - \lambda \text{Tr}\langle \mathbf{y}\mathbf{y}^T \rangle \tag{38}$$

Writing it down explicitly, we have

$$\mathbf{C_Y} = \langle (\mathbf{W}(\mathbf{x}+\mathbf{z})+\eta)(\mathbf{W}(\mathbf{x}+\mathbf{z})+\eta)^T \rangle \tag{39}$$
$$= \langle \mathbf{W}\mathbf{x}\mathbf{x}^T\mathbf{W}^T \rangle + \langle \mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T \rangle + \langle \eta\eta^T \rangle \tag{40}$$

Likewise,

$$\mathbf{C_\xi} = \langle \mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T \rangle + \langle \eta\eta^T \rangle \tag{41}$$

Now let us consider choosing an orthogonal matrix $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$ that can diagonalize the input and output covariance matrix

$$\mathbf{U}\langle \mathbf{x}\mathbf{x}^T \rangle \mathbf{U}^T = \begin{bmatrix} \searrow & & \\ & D_i^2 & \\ & & \searrow \end{bmatrix} \tag{42}$$

Since both determinant and trace are self-invariant under a similarity transformation, we can take

$$\mathbf{U}\mathbf{C_Y}\mathbf{U}^T = \hat{\mathbf{W}} \begin{bmatrix} \searrow & & \\ & D_i^2 & \\ & & \searrow \end{bmatrix} \hat{\mathbf{W}}^T + \hat{\mathbf{W}} \begin{bmatrix} \searrow & & \\ & \sigma_z^2 & \\ & & \searrow \end{bmatrix} \hat{\mathbf{W}}^T + \mathbf{U} \begin{bmatrix} \searrow & & \\ & \sigma_\eta^2 & \\ & & \searrow \end{bmatrix} \mathbf{U}^T, \tag{43}$$

where

$$\hat{\mathbf{W}} = \mathbf{U}\mathbf{W}\mathbf{U}^T = \begin{bmatrix} \searrow & & \\ & \psi_i & \\ & & \searrow \end{bmatrix} \tag{44}$$

The objective function can now be expressed as

$$F = \frac{1}{2} \sum_i \left[\log_2 \frac{D_i^2\psi_i^2 + \sigma_z^2\psi_i^2 + \sigma_\eta^2}{\sigma_z^2\psi_i^2 + \sigma_\eta^2} - 2\lambda(D_i^2\psi_i^2 + \sigma_z^2\psi_i^2 + \sigma_\eta^2) \right] \tag{45}$$

By defining $d_i^2 = \frac{D_i^2}{\sigma_z^2}$, $\hat{\psi}_i^2 = \frac{\psi_i^2\sigma_z^2}{\sigma_\eta^2}$, $\hat{\lambda} = \lambda\sigma_\eta^2$, the objective function can be rewritten as

$$F = \frac{1}{2} \sum_i \left[\log_2 \frac{d_i^2\hat{\psi}_i^2 + \hat{\psi}_i^2 + 1}{\hat{\psi}_i^2 + 1} - 2\hat{\lambda}(d_i^2\hat{\psi}_i^2 + \hat{\psi}_i^2 + 1) \right] \tag{46}$$

When the signal to noise level is small $d_i^2 \ll 1$, F is maximized when

$$\frac{d_i^2}{(\hat{\psi}_i^2 + 1)^2} = \text{const} \tag{47}$$

When the signal to noise level is very large $d_i^2 \gg 1$, F is maximized when

$$\hat{\psi}_i^2 \sim \frac{1}{d_i^2} \tag{48}$$