1 Information Theory

1.1 Information

Core concept:

> "Highly improbable events bring more information to us while certain events bring no information."

The information of an event x will therefore depends on its probability distribution $p(x)$. Let $h(\cdot)$ be the monotonic function of $p(x)$ that returns information. If x and y are unrelated events, we hope the information they take are also unrelated, so

$$h(x, y) = h(x) + h(y)$$
$$p(x, y) = p(x)p(y)$$

Note that we can interpret $h(p(x))$ as $h(x)$ and $h(p(x, y))$ as $h(x, y)$. This means we can define

$$h(x) = -\log_2 p(x)$$

Then $h(x)$ satisfies $2^{h(x)} = 1/p(x)$. We can interpret this as

> "$h(x)$ is the amount of bits that being enough for representing $1/p(x)$ in binary."

1.2 Entropy

Let x be the state that transmitted from a sender to a receiver. Intuitively, the average amount of the information that x carries is obtained by taking the expectation of information $h(x)$ with respect to the p.d.f. $p(x)$

$$\sum_x p(x)h(x) = -\sum_x p(x) \log_2 p(x)$$

This is called the *entropy* of the random variable x and denote it by $\mathrm{H}[x]$. Since $\lim_{p\to 0} p \ln p = 0$, we just take $p(x) \ln p(x) = 0$ when we encounter $p(x) = 0$ for some x. The *noiseless coding theorem*(not actually understand what the hell is this) states that entropy is a lower bound of the amount of bits that a random variable can transmits.

In practice, we use $\ln p(x)$ instead of $\log_2 p(x)$. That is

$$h(x) = -\ln p(x)$$
$$\mathrm{H}[x] = -\sum_x p(x) \ln p(x)$$

In this situation, we said the information is measured in the units of 'nats'.

1.3 Maximize Entropy in Discrete Case

Let $X = \{x_i\}_{i=1}^M$ be a discrete random variable and let p be the distribution of X. For the problem

$$\max \left(-\sum_i p(x_i) \ln p(x_i) \right) \quad \textit{subject to} \quad \sum_i p(x_i) = 1$$

The Lagrangian function is

$$\tilde{\mathrm{H}} = -\sum_i p(x_i)\ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right)$$

From $\partial\tilde{H}/\partial p(x_k) = -(\ln p(x_k)+1)+\lambda = 0$, we have $\lambda = \ln p(x_k)+1$, $\forall\ k = 1,\ldots,M$. Since $\sum_k p(x_k) = M\cdot e^{\lambda-1} = 1$, $\lambda = \ln(1/M)+1$. We have

$$p(x_k) = 1/M,\ \forall\ k$$

The entropy becomes $\mathrm{H}[x] = \ln M$.

1.4 Differential Entropy(Entrpoy in Continuous Case)

Let X be a continuous random variable and p be the distribution of X. By M.V.T, we know there exists some x_i such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta$$

where Δ is the length of one partition of X. Now for any $x\in[i\Delta,(i+1)\Delta]$, we can use $p(x_i)\Delta$ to estimate its probability as long as Δ small enough. Here comes an entropy

$$\begin{aligned}
\mathrm{H}_\Delta &= -\sum_i p(x_i)\Delta\ln(p(x_i)\Delta)\\
&= -\sum_i p(x_i)\Delta\ln(p(x_i)\Delta) + \sum_i p(x_i)\Delta\ln\Delta - \sum_i p(x_i)\Delta\ln\Delta\\
&= -\sum_i p(x_i)\Delta\ln p(x_i) - \ln\Delta
\end{aligned}$$

Note that $\sum_i p(x_i)\Delta = 1$. Take out the first term of right hand side,

$$\lim_{\Delta\to 0} -\sum_i p(x_i)\Delta\ln p(x_i) = -\int p(x)\ln p(x)dx$$

This integral is called the *differential entropy*. The equation between H_Δ and the differential entropy shows the fact the we need lots of bits to describe a continuous variable.

When it comes to multiple dimension we also have

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x}$$

1.5 Gaussian Maximize Differential Entropy

We want to maximize

$$\mathrm{H}[x] = -\int p(x)\ln p(x)dx$$

with three constraints

$$\int_{-\infty}^{\infty} p(x)dx = 1$$
$$\int_{-\infty}^{\infty} xp(x)dx = \mu$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2$$

Here comes the Lagrangian function

$$-\int_{\mathbb{R}} p(x) \ln p(x) dx + \lambda_1 \left(\int_{\mathbb{R}} p(x) dx - 1 \right) + \lambda_2 \left(\int_{\mathbb{R}} x p(x) dx - \mu \right) + \lambda_3 \left(\int_{\mathbb{R}} (x-\mu)^2 p(x) dx - \sigma^2 \right)$$

This is a functional $F[p]$. The derivative of a functional is denoted by $\frac{\delta F}{\delta p}$ and is defined to satisfy

$$\int \frac{\delta F[p]}{\delta p} \phi(x) dx = \lim_{\epsilon \to 0} \frac{F[p+\epsilon\phi] - F[p]}{\epsilon} = \left[\frac{d}{d\epsilon} F[p+\epsilon\phi] \right]_{\epsilon=0}$$

where $\phi(x)$ is a variation term. Followed by the definition

$$\int \frac{\delta F[p]}{\delta p} \phi(x) dx = \int \left(-(\ln p(x) + 1) + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 \right) \phi(x) dx$$

We have the actual form of $\frac{\delta F[p]}{\delta p}$ and can let it be zero then get

$$p(x) = e^{-1+\lambda_1+\lambda_2 x + \lambda(x-\mu)^2}$$

Substitute this result back to three constraints leading to

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Furthermore,

$$\mathrm{H}[x] = \frac{1}{2}\{1 + \ln(2\pi\sigma^2)\}$$

1.6 Kullback-Leibler Divergence

Let $p(x)$ be an unknown distribution and we use $q(x)$ to approximate it. This will cause additional amount of information when transmitting

$$\begin{aligned} \mathrm{KL}(p\|q) &= \left(-\int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} \right) - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x} \end{aligned}$$

This is known as the *relative entropy* or *Kullback-Leiber divergence*, or *KL divergence* between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. Note that $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$.