

1 Information Theory

1.1 Information

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

The information of an event x will therefore depends on the probability distribution $p(x)$ of its random variable X .

1.2 Construct Information Formula

Let $h(\cdot)$ be a monotonic function of any distribution $p(x)$ that returns the information of $p(x)$. If x and y are unrelated events, we hope the information they take are also unrelated, so

$$h(x, y) = h(x) + h(y) \quad (1)$$

$$p(x, y) = p(x)p(y) \quad (2)$$

Note that we can interpret $h(p(x))$ as $h(x)$ and interpret $h(p(x, y))$ as $h(x, y)$. $\log_2(x)$ is a monotonic function that satisfied both (1) and (2), hence we can define

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

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

$h(x)$ is the amount of bits that being enough for representing $\frac{1}{p(x)}$ in binary.

When $p(x)$ is low, the probability is low, we need more bits to represent it.

1.3 Entropy

Let X be the random variable of 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. $p(x)$

$$\sum_{x \in X} p(x)h(x) = - \sum_{x \in X} p(x) \log_2 p(x)$$

This is called the *entropy* of the random variable X and denote it by $H(X)$ or $H(p)$ or $H(x)$, based on the context of the paragraph.

Since $\lim_{p \rightarrow 0} p \ln p = 0$, we just take $p(x) \ln p(x) = 0$ when we encounter $p(x) = 0$ for some x .

1.3.1 Nats

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

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

$$H(p) = - \sum_{x \in X} p(x) \ln p(x)$$

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

1.3.2 Entropy as Lower Bound

Entropy is a lower bound of the amount of bits that a random variable can transmits.

by the *Noiseless Coding Theorem*.

1.3.3 Noiseless Coding Theorem

N i.i.d. random variables each with entropy $H(X)$ can be compressed into more than $N H(X)$ bits with negligible risk of information loss, as $N \rightarrow \infty$; but conversely, if they are compressed into fewer than $N H(X)$ bits it is virtually certain that information will be lost.

(ChatGPT) In essence, the theorem states that for any given data source with a certain probability distribution of symbols, it is possible to encode the source in such a way that the average length of the encoded message per symbol is close to the entropy of the source.

1.4 Maximize Entropy in Discrete Case

TL;DR

The distribution that can carry the most average amount of information is the uniform.

Proof

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

$$\max \left(- \sum_{i=1}^M p(x_i) \ln p(x_i) \right)$$

with the normalization constraint on the probabilities

$$\sum_{i=1}^M p(x_i) = 1$$

The Lagrangian is

$$\mathcal{L} = - \sum_{i=1}^M p(x_i) \ln p(x_i) + \lambda \left(\sum_{i=1}^M p(x_i) - 1 \right)$$

From $\partial \mathcal{L} / \partial p(x_i) = -(\ln p(x_i) + 1) + \lambda = 0$, we have $\lambda = \ln p(x_i) + 1$, $\forall i = 1, \dots, M$. By $\sum_{i=1}^M p(x_i) = 1$ and $\lambda = \ln p(x_i) + 1$, it's easy to get $\lambda = \ln(1/M) + 1$, we then have

$$p(x_i) = \frac{1}{M}, \forall i$$

is the stationary point.

To verify the maximum, first compute the Hessian matrix

$$\frac{\partial^2 \mathcal{L}}{\partial p(x_i) \partial p(x_j)} = -I_{ij} \frac{1}{p(x_i)}$$

It's obvious that all the eigenvalues are negative (negative definite). So $p(x_i) = \frac{1}{M}$ actually attains a maximum, and the maximum entropy is $H(p) = \ln M$.

1.5 Differential Entropy

Let $X \subseteq \mathbb{R}$ be a continuous random variable and $p(x)$ 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 estimation

$$\begin{aligned} H_\Delta &= - \sum_i p(x_i)\Delta \ln(p(x_i)\Delta) \\ &= - \sum_i p(x_i)\Delta \ln(p(x_i)\Delta) + \sum_i p(x_i)\Delta \ln \Delta - \sum_i p(x_i)\Delta \ln \Delta \\ &= - \sum_i p(x_i)\Delta \ln \left(\frac{p(x_i)\Delta}{\Delta} \right) - \left(\sum_i p(x_i)\Delta \right) \ln \Delta \\ &= - \sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \end{aligned}$$

Note that $\sum_i p(x_i)\Delta = \int_{x \in X} p(x)dx = 1$.

The limit of the first term of right hand side is

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

This integral is called the *differential entropy*.

The difference term $\ln \Delta$ shows the fact that we need lots of bits to describe a continuous variable.

The differential entropy can have negative values when $\sigma^2 < 1/(2\pi e)$.

For multi dimension random variable, the differential entropy is similar

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

1.6 Conditional Entropy

Given a joint probability $p(\mathbf{x}, \mathbf{y})$. When \mathbf{x} is known, the additional information needed to specify the corresponding value of \mathbf{y} is given by $-\ln p(\mathbf{y}|\mathbf{x})$. We can compute the *conditional entropy*

$$H(\mathbf{y}|\mathbf{x}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})d\mathbf{y}d\mathbf{x}$$

Note that

$$H(\mathbf{x}, \mathbf{y}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y})d\mathbf{x}d\mathbf{y}$$

$$\begin{aligned}
&= - \iint p(\mathbf{x}, \mathbf{y}) \ln(p(\mathbf{y}|\mathbf{x})p(\mathbf{x})) d\mathbf{x}d\mathbf{y} \\
&= - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x}d\mathbf{y} - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}) d\mathbf{x}d\mathbf{y} \\
&= - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x}d\mathbf{y} - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} \\
&= H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x})
\end{aligned}$$

1.7 Cross Entropy

The cross entropy of the distribution $q(x)$ relative to a distribution $p(x)$ is

$$H(p, q) = -\mathbb{E}_p[\ln q] = - \sum_x p(x) \ln q(x)$$

In deep learning, $p(x)$ often refers to the ground truth label, and $q(x)$ refers to the output from a deep neural network model.

In information theory, minimize cross entropy means

Minimizes the amount of information required to specify the value of x as a result of using $q(x)$.

1.8 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

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

This is known as the *KL divergence* from q to p . *KL divergence* is also called the *relative entropy*.

Note that $\text{KL}(p\|q) \neq \text{KL}(q\|p)$.

TL;DR

$$\text{KL}(p\|q) \geq 0, \text{KL}(p\|q) = 0 \iff p = q.$$

Convex function

For $0 \leq \lambda \leq 1$, a convex function satisfies

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$$

A convex function is called strictly convex if the equality holds only when $\lambda = 0$ or $\lambda = 1$.

Jensen's Inequality

Recall the *Jensen's inequality*, for a convex function f ,

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i)$$

where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.

When each λ_i becomes the probability $p(x_i)$, we have

$$f(E[x]) \leq E[f(x)].$$

For continuous random variable, we have

$$f\left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x}\right) \leq \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Proof

Observe that $-\ln x$ is a strictly convex function, so by Jensen's inequality

$$\begin{aligned} \text{KL}(p||q) &= - \int p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \\ &\geq - \ln \left(\int p(\mathbf{x}) \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \right) \\ &= - \ln \left(\int q(\mathbf{x}) d\mathbf{x} \right) = 0 \end{aligned}$$

When the equality holds,

$$\begin{aligned} p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} &= 0 \\ \Rightarrow \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} &= 0 \\ \Rightarrow \frac{q(\mathbf{x})}{p(\mathbf{x})} &= 1 \\ \Rightarrow q(\mathbf{x}) &= p(\mathbf{x}) \end{aligned}$$

1.9 Gaussian Maximizes Differential Entropy

We want to maximize

$$H(p) = - \int p(x) \ln p(x) dx$$

with three natural constraints

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

$$\begin{aligned}\int_{-\infty}^{\infty} xp(x)dx &= \mu \\ \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx &= \sigma^2\end{aligned}$$

The Lagrangian is

$$\begin{aligned}\mathcal{L}[p] = & - \int_{\mathbb{R}} p(x) \ln p(x) dx + \lambda_1 \left(\int_{\mathbb{R}} p(x) dx - 1 \right) + \\ & \lambda_2 \left(\int_{\mathbb{R}} xp(x) dx - \mu \right) + \lambda_3 \left(\int_{\mathbb{R}} (x - \mu)^2 p(x) dx - \sigma^2 \right)\end{aligned}$$

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

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

where $\phi(x)$ is a variation term.

Deriving from the definition

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

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

$$p(x) = e^{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2}$$

Substitute this result back to three constraints leading to

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

which is the Gaussian.

To verify the maximum, let $f(x)$ be any distribution has the variance σ^2 . Since differential entropy is translation invariant, we can also assume $f(x)$ has the same mean μ . Now consider the KL divergence

$$\begin{aligned}\text{KL}(f\|p) &= - \int f(x) \ln \left(\frac{p(x)}{f(x)} \right) dx \\ &= -H(f) - \int f(x) \ln \left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \\ &= -H(f) + \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \int f(x)(x - \mu)^2 dx \\ &= -H(f) + \frac{1}{2} \ln(2\pi\sigma^2) + \frac{\sigma^2}{2\sigma^2} \\ &= -H(f) + \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \\ &= -H(f) + H(p) \geq 0\end{aligned}$$

Hence

$$H(p) \geq H(f), \forall f$$

The corresponding maximum entropy is,

$$H(p) = \frac{1}{2}(1 + \ln(2\pi\sigma^2))$$

1.10 KL Divergence in Deep Learning

Let $p(\mathbf{x})$ be an unknown target distribution. We have N training data $\{x_i\}_{i=1}^N$ drawn from it. Let $q(\mathbf{x}|\theta)$ be a neural network tries to approximate $p(\mathbf{x})$ with the model weight θ . In theory, the training process should minimize the KL divergence

$$\text{KL}(p\|q) = -\mathbb{E}_p \left[\ln \frac{q}{p} \right] = -\int p(\mathbf{x}) \ln \frac{q(\mathbf{x}|\theta)}{p(\mathbf{x})} d\mathbf{x}.$$

Because $p(\mathbf{x})$ is unknown, we can't directly compute the KL divergence. However, we can use sample mean to estimate it

$$\text{KL}(p\|q) = -\mathbb{E}_p \left[\ln \frac{q}{p} \right] \approx -\frac{1}{N} \sum_{i=1}^N \ln \left(\frac{q(\mathbf{x}_i|\theta)}{p(\mathbf{x}_i)} \right)$$

During training, N is fixed, so

$$\text{KL}(p\|q) \approx \sum_{i=1}^N -\ln q(\mathbf{x}_i|\theta) + \ln p(\mathbf{x}_i).$$

$\ln p(\mathbf{x})$ is not related to the training, so minimizing $\sum_i -\ln q(\mathbf{x}_i|\theta)$ is equivalent to minimizing the KL divergence.

1.11 Mutual Information

The following KL divergence is called the *Mutual Information* $I[\mathbf{x}, \mathbf{y}]$,

$$I[\mathbf{x}, \mathbf{y}] = \text{KL}(p(\mathbf{x}, \mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) = -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x}$$