Differential Entropy

Definition

Let X be a random variable with cumulative distribution function (CDF) $F(x) = \Pr(X \le x)$. If $F(x)$ is continuous, the random variable is continuous. Let $f(x) = F'(X)$ when the derivative is defined. If $\int_{-\infty}^{+\infty} f(x) = 1$, $f(x)$ is called the probability density function (pdf) for X. The set of x where $f(x) > 0$ is called the support set of the X.

Definition

The differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

$$h(X) = -\int_S f(x) \log f(x) \mathrm{d}x = h(f),$$

where S is the support set of the random variable.

Example: Uniform distribution

- $f(x) = \frac{1}{a}, x \in [0, a]$
- The differential entropy is:

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \mathrm{d}x = \log a \text{ bits}$$

- for $a < 1$, $h(X) = \log a < 0$, differential entropy can be negative! (unlike discrete entropy)

微分熵可以为负

Example: Normal distribution

- $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-x^2}{2\sigma^2})$, $x \in \mathbb{R}$
- Differential entropy:

$$h(\phi) = \frac{1}{2} \log 2\pi e\sigma^2 \text{ bits}$$

Calculation:

$$\begin{aligned} h(\phi) &= -\int \phi \log \phi \mathrm{d}x = -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] \mathrm{d}x \\ &= \frac{\mathbb{E}(X^2)}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 = \frac{1}{2} \log e + \frac{1}{2} \log 2\pi\sigma^2 \\ &= \frac{1}{2} \log 2\pi e\sigma^2 \end{aligned}$$

AEP for continuous random variables

- Discrete world: for a sequence of i.i.d. random variables

$$\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \to H(X).$$

- Continuous world: for a sequence of i.i.d. random variables

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \to \mathbb{E}[-\log f(X)] = h(X) \quad \text{in probability}$$

Proof follows from the weak law of large numbers.

Typical set

- Discrete case: number of typical sequences

$$\left|A_\epsilon^{(n)}\right| \approx 2^{nH(X)}$$

- Continuous case: The volume of the typical set

$$\mathrm{Vol}(A) = \int_A \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n, \ A \subset \mathbb{R}^n.$$

Definition

For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \dots, x_n) \in \mathcal{S}^n : \left| -\frac{1}{n} \log f(x_1, x_2, \dots, x_n) - h(X) \right| \le \epsilon \right\},$$

where $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$.

Theorem

The typical set $A_\epsilon^{(n)}$ has the following properties:

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ *for n sufficiently large.*
2. $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X)+\epsilon)}$ *for all n.*
3. $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\epsilon) 2^{n(h(X)-\epsilon)}$ *for n sufficiently large.*

Proof. 1.

Similar to the discrete case.
By definition, $-\frac{1}{n} \log f(X^n) = -\frac{1}{n} \sum \log f(X_i) \to h(X)$ in probability. □

Theorem

The typical set $A_\epsilon^{(n)}$ has the following properties:

2. $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X)+\epsilon)}$ *for all n.*

Poof. 2.

$$\begin{aligned} 1 &= \int_{\mathcal{S}^n} f(x_1, x_2, \dots, x_n) \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n \\ &\ge \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n \\ &\ge \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)} \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n = 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n \\ &= 2^{-n(h(X)+\epsilon)} \mathrm{Vol}(A_\epsilon^{(n)}). \end{aligned}$$

□

Theorem

The typical set $A_\epsilon^{(n)}$ has the following properties:

3. $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\epsilon) 2^{n(h(X)-\epsilon)}$ *for n sufficiently large.*

Proof. 3.

$$\begin{aligned} 1 - \epsilon &\le \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n \\ &\le \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)} \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n \\ &= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_n \\ &= 2^{-n(h(X)-\epsilon)} \mathrm{Vol}(A_\epsilon^{(n)}). \end{aligned}$$

□

② $1 = \int_{S^n} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$

$\ge \int_{A_\varepsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$

因为 $A_\varepsilon^{(n)} = \{(x_1, \dots, x_n) \mid |-\frac{1}{n}\log f(x_1, \dots, x_n) - h(X)| \le \varepsilon\}$

所以 $2^{-n[h(X)+\varepsilon]} \le f(x_1, \dots, x_n) \le 2^{-n[h(X)-\varepsilon]}$

故 $\ge \int_{A_\varepsilon^{(n)}} 2^{-n[h(X)+\varepsilon]} dx_1 dx_2 \dots dx_n$

$\Rightarrow 2^{n[h(X)+\varepsilon]} \ge \int_{A_\varepsilon^{(n)}} dx_1 \dots dx_n$

即 $\mathrm{Vol}(A_\varepsilon^{(n)}) \le 2^{n[h(X)+\varepsilon]}$

③ $\Pr[A_\varepsilon^{(n)}] \ge 1-\varepsilon$ for sufficiently large n.

$\Pr[A_\varepsilon^{(n)}] = \int_{A_\varepsilon^{(n)}} f(x_1 \dots x_n) dx_1 \dots dx_n$

$\le \int_{A_\varepsilon^{(n)}} 2^{-n[h(X)-\varepsilon]} dx_1 \dots dx_n$

$= 2^{-n[h(X)-\varepsilon]} \mathrm{Vol}(A_\varepsilon^{(n)})$

$\Rightarrow \mathrm{Vol}(A_\varepsilon^{(n)}) \ge (1-\varepsilon) 2^{n[h(X)-\varepsilon]}$

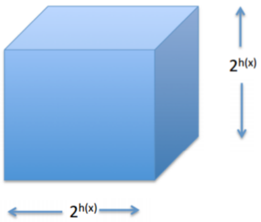
for sufficiently large n

ε是无穷小，由夹逼得 $\mathrm{Vol}(A_\varepsilon^{(n)}) \approx 2^{nh(X)}$

An interpretation

$A_\epsilon^{(n)}$可想象为一个容器，其体积为$2^{nh(X)}$，若该容器是n维的，则其边长为$2^{h(X)}$

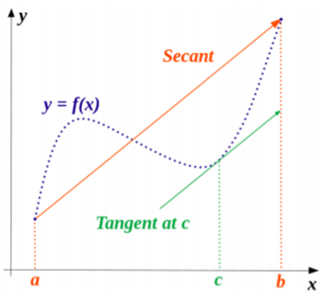
- The volume of the smallest set that contains most of the probability is approximately $2^{nh(X)}$.
- For an n-dim volume, this means that each dim has length $\left(2^{nh(X)}\right)^{\frac{1}{n}} = 2^{h(X)}$.



Mean value theorem (MVT)

If a function f is continuous on the closed interval $[a, b]$, and differentiable on (a, b), then there exists a point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



$X \sim f$ (PDF)

$$\begin{aligned} h(X) &= h(f) \\ &= -\int_S f(x) \log f(x) dx \\ &= -E[\log f(x)] \end{aligned}$$

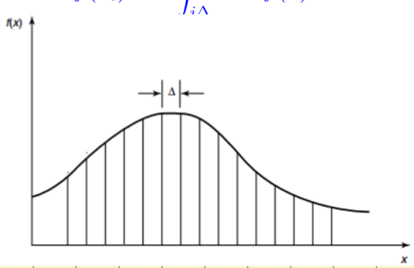
$X \sim p$ (PMF)

$$\begin{aligned} H(X) &= H(p) \\ &= -\sum_S p \log p \\ &= -E[\log p] \end{aligned}$$

Relation of differential entropy to discrete entropy

- Consider a random variable X with pdf $f(x)$. We divide the range of X into bins of length Δ.
- MVT: there exists a value $x_i \in (i\Delta, (i+1)\Delta)$ within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x) \mathrm{d}x.$$



- Define the quantized random variable as $X^{\Delta} = x_i$ if $i\Delta \le X \le (i+1)\Delta$ with pmf

$$p_i = \Pr[X^{\Delta} = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)\mathrm{d}x = f(x_i)\Delta.$$

- The entropy of X^{Δ} is

$$H(X^{\Delta}) = -\sum_{-\infty}^{+\infty} p_i \log p_i = -\sum \Delta f(x_i)\log f(x_i) - \log\Delta.$$

- If $f(x)$ is is Riemann integrable, as $\Delta \to 0$,

$$H(X^{\Delta}) + \log\Delta \to h(f) = h(X)$$

$$X^{\Delta} \xrightarrow{\Delta\to 0} X$$

entropy, differential entropy

$$\underbrace{H(X^{\Delta})}_{\text{entropy}} = \underbrace{h(X)}_{\text{differential entropy}} \underbrace{-\log\Delta}_{+\infty}$$

连续随机变量的信息量是无穷大的
因此要比较用相对量 differential entropy
(这也是为什么 differential entropy 可以为负)

Joint and conditional entropy

Definition

The joint differential entropy of $X_1, X_2, ..., X_n$ with pdf $f(x_1, x_2, \ldots, x_n)$ is

$$h(X_1, X_2, \ldots, X_n) = -\int f(x^n)\log f(x^n)\mathrm{d}x^n.$$

Definition

If X, Y have a joint pdf $f(x, y)$, the conditional differential entropy $h(X|Y)$ is

$$h(X|Y) = -\int f(x,y)\log f(x|y)\mathrm{d}x\mathrm{d}y = h(X,Y) - h(Y).$$

Entropy of a multivariate Gaussian

Definition (Multivariate Gaussian Distribution)

If the joint pdf of $X_1, X_2, \ldots, X_n$ satisfies

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)\right),$$

then $X_1, X_2, \ldots, X_n$ are multivariate/joint Gaussian/normal distributed with mean μ and covariance matrix K. Denote as $(X_1, X_2, \ldots, X_n) \sim \mathcal{N}_n(\mu, K)$.

$$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mu = E\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

$$K = E[(\vec{X}-\mu)(\vec{X}-\mu)^T]$$

Theorem (Entropy of a multivariate normal distribution)

Let $X_1, X_2, \ldots, X_n$ have multivariate normal distribution with mean μ and covariance matrix K. Then

$$h(X_1, X_2, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2}\log(2\pi e)^n|K| \text{ bits},$$

where $|K|$ denotes the determinant of K.

$\vec{X} \sim N_n(\mu, K)$

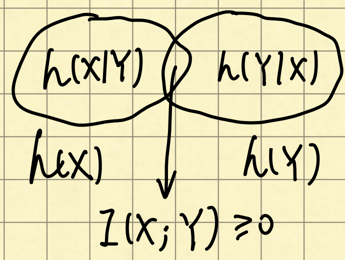
$$\begin{aligned}
h(\vec{X}) &= h(X_1, X_2, \ldots, X_n) \\
&= -\int f(x^n)\log f(x^n)\,dx^n \\
&= -\int f(x^n)\left[-\log\left((\sqrt{2\pi})^n|K|^{\frac{1}{2}}\right) - (\log e)\tfrac{1}{2}(\vec{x}-\mu)^T K^{-1}(\vec{x}-\mu)\right]dx^n \\
&= \log\left[(2\pi)^{\frac{n}{2}}|K|^{\frac{1}{2}}\right] + \tfrac{1}{2}\log e\,\underbrace{\int f(x^n)(\vec{x}-\mu)^T K^{-1}(\vec{x}-\mu)\,dx^n}_{E[(\vec{X}-\mu)^T K^{-1}(\vec{X}-\mu)]}
\end{aligned}$$

$$\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$$

$$\begin{aligned}
&E[(\vec{X}-\mu)^T K^{-1}(\vec{X}-\mu)] \\
&= E[\mathrm{tr}((\vec{X}-\mu)^T K^{-1}(\vec{X}-\mu))] \\
&= E[\mathrm{tr}((\vec{X}-\mu)(\vec{X}-\mu)^T K^{-1})] \\
&= \mathrm{tr}(E[(\vec{X}-\mu)(\vec{X}-\mu)^T]K^{-1}) \\
&= \mathrm{tr}(KK^{-1}) = \mathrm{tr}(I_n) = n
\end{aligned}$$

$$\begin{aligned}
h(\vec{X}) &= \log\left[(2\pi)^{\frac{n}{2}}|K|^{\frac{1}{2}}\right] + \tfrac{1}{2}\log e^n \\
&= \tfrac{1}{2}\log\left[(2\pi e)^n|K|\right]
\end{aligned}$$

Relative entropy and mutual information



Definition

The relative entropy $D(f||g)$ between two pdfs f and g is

$$D(f||g) = \int f \log \frac{f}{g}.$$

Note: $D(f||g)$ is finite only if the support set of f is contained in the support set of g.

Definition

The mutual information $I(X;Y)$ between two random variables with joint pdf $f(x,y)$ is

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy.$$

By definition, it is clear that

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y).$$

and

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

Mutual information between correlated Gaussian r.v.s

- Let $(X,Y) \sim \mathcal{N}(0,K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

$$K = E\left[\begin{pmatrix} X \\ Y \end{pmatrix}(X\ Y)\right] = E\begin{bmatrix} X^2 & XY \\ YX & Y^2 \end{bmatrix}$$

- $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$
- $h(X,Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2)$
- $I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log(1-\rho^2)$

  if $\rho = 0$, X and Y are independent, the mutual information is 0.

  if $\rho \pm 1$, X and Y are perfectly correlated, the mutual information is infinite.

Theorem

$D(f||g) \geq 0$ with equality iff $f = g$ almost everywhere.

Proof.

Let $\mathcal{S}$ be the support set of f. Then

$$\begin{aligned} -D(f||g) &= \int_{\mathcal{S}} f \log \frac{g}{f} \\ &\leq \log \int_{\mathcal{S}} f \frac{g}{f} \quad \text{(by Jensen's inequality)} \\ &= \log \int_{\mathcal{S}} g \\ &\leq \log 1 = 0 \end{aligned}$$

□

Properties of differential entropy

- $I(X;Y) \geq 0$ with equality iff X and Y are independent.
- $h(X|Y) \leq h(X)$ with equality iff X and Y are independent.

Theorem (Chain rule for differential entropy)

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}).$$

- $h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)$, with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Theorem (Translation does not change the differential entropy)

$$h(X+c) = h(X).$$

Theorem

$$h(aX) = h(X) + \log|a|.$$

Proof.

Let $Y = aX$, Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and we have

$$h(aX) = -\int f_Y(y)\log f_Y(y)\mathrm{d}y = -\int \frac{1}{|a|} f_X(\frac{y}{a}) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right)\mathrm{d}y$$
$$= -\int f_X(x)\log f_X(x)\mathrm{d}x + \log|a| = h(X) + \log|a|$$

□

Corollary.

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log|\det(A)|.$$

Multivariate Gaussian maximizes the entropy

Theorem

Let the random vector $\mathbf{X} \in \mathbb{R}^n$ have zero mean and covariance $K = \mathbb{E}\mathbf{X}\mathbf{X}^t$ (i.e., $K_{ij} = \mathbb{E}X_iX_j$, $1 \le i, j \le n$). Then

$$h(\mathbf{X}) \le \frac{1}{2}\log(2\pi e)^n|K|$$

with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.

令 $\phi_K \sim N_n(0,K)$，$\vec{x} \sim g$ 的均值为0，协方差矩阵为 K.

$$\phi_K(x^n) = \frac{1}{(\sqrt{2\pi})^n|K|^{1/2}}\exp\left(\tfrac{1}{2}\vec{x}^T K \vec{x}\right)$$

$$0 \le D(g\|\phi_K) = \int g(x^n)\log\frac{g(x^n)}{\phi_K(x^n)}dx^n$$
$$= \int g(x^n)\log g(x^n)dx^n - \int g(x^n)\log\phi_K(x^n)dx^n$$
$$= -h(g) - \int \underline{g(x^n)}\log\phi_K(x^n)dx^n$$

换成 $\phi_K(x^n)$ 后积分值不变

$$= -h(g) + h(\phi_K)$$
$$\Rightarrow h(g) = h(\vec{x}) \le h(\phi_K) = \tfrac{1}{2}\log(2\pi e)^n|K|$$

$$\int g(x^n)\log\phi_K(x^n)dx^n$$
$$= \int g(x^n)\log\frac{1}{(\sqrt{2\pi})^n|K|^{1/2}}dx^n + \int g(x^n)(-\tfrac{1}{2})\vec{x}^T K^{-1}\vec{x}\log e\, dx^n$$

$\int g(x^n)dx^n = 1$

$$= \log\frac{1}{(2\pi)^n|K|^{1/2}} - \tfrac{1}{2}\log e \underbrace{\int g(x^n)\vec{x}^T K^{-1}\vec{x}\,dx^n}_{E_g(\vec{x}^T K^{-1}\vec{x})}$$

$$E_g(\vec{x}^T K^{-1}\vec{x}) = E_g\,\mathrm{tr}(\vec{x}^T K^{-1}\vec{x}) = E_g\,\mathrm{tr}(\vec{x}\,\vec{x}^T K^{-1})$$
$$= \mathrm{tr}[\underbrace{E_g(\vec{x}^T\vec{x})}_{K}K^{-1}] = \mathrm{tr}\, I_n = n$$

Random variable X, estimator $\hat{X}$. The expected prediction error $\mathbf{E}(X-\hat{X})^2$.

Theorem (Estimation error and differential entropy)

For any random variable X and estimator $\hat{X}$,

$$\mathbb{E}(X-\hat{X})^2 \geq \frac{1}{2\pi e}\exp\Big(2h(X)\Big),$$

with equality iff X is Gaussian and $\hat{X}$ is the mean of X.

Proof.

We have

$$\begin{aligned}
\mathbb{E}(X-\hat{X})^2 &\geq \min_{\hat{X}} \mathbb{E}(X-\hat{X})^2 \\
&= \mathbb{E}(X-\mathbb{E}(X))^2 && \text{mean is the best estimator} \\
&= \mathrm{Var}(X) \\
&\geq \frac{1}{2\pi e}\exp\Big(2h(X)\Big). && \text{The Gaussian has maximum entropy}
\end{aligned}$$

□

$$\begin{aligned}
E(X-\hat{X})^2 &\geq \min_a E(X-a)^2 \\
&= \min_a E(X^2-2aX+a^2) \\
&= \min_a EX^2-2aEX+a^2
\end{aligned}$$

$-2EX+2a=0 \Rightarrow a=EX$

$$\begin{aligned}
&= E(X-EX)^2 \\
&= Var(X)
\end{aligned}$$

由 $h(\vec{X}) \leq \frac{1}{2}\log(2\pi e)^n|K|$ 可得，对一维随机变量 x，则有 $h(X) \leq \frac{1}{2}\log[2\pi e\, Var(X)] \Rightarrow Var(X) \geq \frac{1}{2\pi e}\exp(2h(X))$

故 $E(X-\hat{X})^2 \geq \frac{1}{2\pi e}\exp(2h(X))$

Summary

- Discrete r.v. $\Rightarrow$ continuous r.v.
- entropy $\Rightarrow$ differential entropy.
- Many things similar: mutual information, relative entropy, AEP, chain rule, ...
  Some things different: $h(X)$ can be negative, maximum entropy distribution is Gaussian