INFORMATION THEORY & CODING

Differential Entropy

Dr. Rui Wang

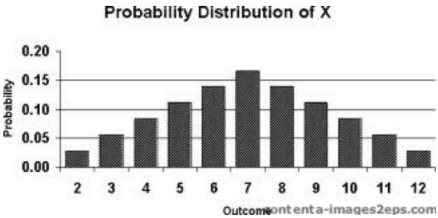
Department of Electrical and Electronic Engineering
Southern Univ. of Science and Technology (SUSTech)

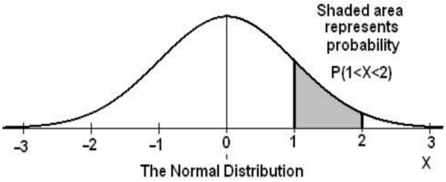
Email: wang.r@sustech.edu.cn

December 20, 2022

- Definitions
- AEP for Continuous Random Variables
- Relation of differential entropy to discrete entropy

From discrete to continuous variables





Differential Entropy

Definition

(累积分布函数)

Let X be a random variable with cumulative distribution function (CDF) $F(x) = \Pr(X \leq x)$. 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.

概率密度函数

Definition

代表 连续随机变量的entropy. H(X)代表离散随机变量的entropy.

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

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

可正可负

where $\mathcal{S}$ 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)

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)

Example: Normal distribution

$X \sim N(0, \sigma^2)$

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

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

Calculation:

$$h(\phi) = -\int \phi \log \phi \mathrm{d}x = -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] \mathrm{d}x$$

$$= \frac{\mathbb{E}(X^2)}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 = \frac{1}{2} \log e + \frac{1}{2} \log 2\pi\sigma^2$$

$$= \frac{1}{2} \log 2\pi e \sigma^2$$

Normal distribution

$$X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \sim N(0, \sigma^2)$$

$$h(X) = h(\phi) = -\int_{-\infty}^{+\infty} \phi(x) \left[-\log \sqrt{2\pi\sigma^2} - \frac{x^2}{2\sigma^2} \log e \right] dx$$

$$= \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \log e = \frac{1}{2} \log (2\pi\sigma^2 e)$$

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: 代表方差.

$$h(\phi) = -\int \phi \log \phi \mathrm{d}x = -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] \mathrm{d}x$$

$$= \frac{\mathbb{E}(X^2)}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 = \frac{1}{2} \log e + \frac{1}{2} \log 2\pi\sigma^2$$

$$= \frac{1}{2} \log 2\pi e \sigma^2$$

AEP for continuous random variables

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

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

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

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

Proof follows from the weak law of large numbers.

AEP for continuous random variables

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

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

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

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

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

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

Definition

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

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

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

Typical set

$X_1, X_2 \cdots\cdots X_n$ are i.i.d random variable

$-\frac{1}{n}\log f(X_1, X_2, \cdots X_n) = -\frac{1}{n}\log \prod_{i=1}^{n} f(X_i) = -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i)$

when $n\to\infty$, law of large number $-\frac{1}{n}\sum_{i=1}^{n}\log f(X_i) = -E[\log f(X)] = -\int_S [\log f(x)] f(x)\,dx = h(X)$

$A_\varepsilon^{(n)} = \{(x_1, x_2, \cdots x_n) : |-\frac{1}{n}\log f(x_1, x_2, \cdots x_n) - h(X)| \le \varepsilon\}$

$\text{Vol}(A_\varepsilon^{(n)}) = \int_{A_\varepsilon^{(n)}} 1 \cdot dx_1 dx_2 \cdots dx_n$

$1 = \int_S f(x_1, x_2, \cdots x_n)\, dx_1 dx_2 \cdots dx_n \ge \int_{A_\varepsilon^{(n)}} f(x_1, x_2, \cdots x_n)\, dx_1 dx_2 \cdots dx_n$

According to the definition of typical set. $2^{-n(h(X)+\varepsilon)} \le f(x_1, x_2, \cdots x_n) \le 2^{-n(h(X)-\varepsilon)}$

$\ge \int_{A_\varepsilon^{(n)}} 2^{-n(h(X)+\varepsilon)}\, dx_1 \cdots dx_n = 2^{-n(h(X)+\varepsilon)}\, \text{Vol}(A_\varepsilon^{(n)}) \to \text{Vol}(A_\varepsilon^{(n)}) \le 2^{n(h(X)+\varepsilon)}$

when $n\to\infty$, $\varepsilon\to 0$. $\text{Vol}(A_\varepsilon^{(n)}) \approx 2^{nh(X)}$

Typical set

Theorem

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

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

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ *for n sufficiently large.*

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

3. $\mathrm{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ *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. $\square$

Typical set

Theorem

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

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

Poof. 2.

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

$\square$

Typical set

Theorem

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

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

总结的.

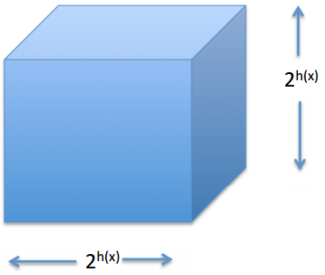
Proof. 3.

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

□

An interpretation

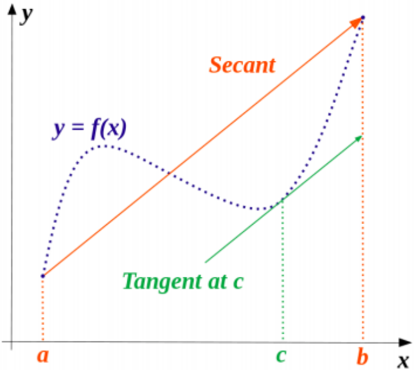
- The volume of the smallest set that contains most of the probability is approximately $2^{nh(X)}$.
- For an n-dim volume, this means that each dim has length $\left(2^{nh(X)}\right)^{\frac{1}{n}} = 2^{h(X)}$.



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

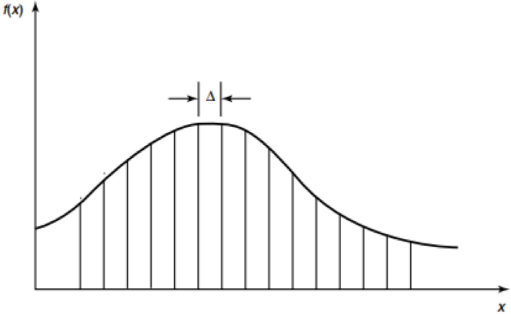
$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



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

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



Relation of differential entropy to discrete entropy

- 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 $\Delta \to 0$,

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

Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^{\Delta} = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ 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 $\Delta \to 0$,

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

Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^{\Delta} = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ 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 $\Delta \to 0$,

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