Information Theory: Lecture Notes 8

zqy1018

2020. 4. 20

Contents

1 Differential Entropy

1.1 Basic Definitions

We first give some basic definitions.

Definition. Let $F(x)$ be the c.d.f. (Cumulative Distribution Function) and $f(x)$ be the p.d.f. (Probability Density Function) of X. If $F(x)$ is continuous, then X is **continuous**. Let $S = \text{supp}(F) = \{x : f(x) > 0\}$ be the support of X.

We now try to define entropy for continuous random variables.

Definition. Then the **differential entropy** $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

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

Note.

(1) Like the discrete case, $h(X)$ only depends on $f(x)$. So sometimes we use $h(f)$ instead of $h(X)$.

(2) Differential entropy does not serve as a measure of the average amount information contained in a continuous random variable. In fact, the information contained by a continuous random variable can be infinitive. Example: for $X \sim U[0, 1)$, write $X = 0.X_1X_2 \cdots$. Then X_i: i.i.d. and $H(X) = \sum H(X_i) = \infty$.

(3) The differential entropy of a discrete random variable can be considered to be $-\infty$, since $f(x)$ can be seen as a set of impulses.

Remark. As in every example involving an integral, or even a density, we should include the statement *if it exists*. But sometimes for convenience we will assume that it exists.

1.2 Basic Properties

Theorem 1. $\forall c, h(X + c) = h(X)$. So transition does not change the entropy.

Proof. By definition. □

Theorem 2. $h(aX) = h(X) + \log |a|$.

Proof. Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$. And then we can prove it by definition. □

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

2 Some Important Examples

2.1 Example: Uniform Distribution

For $X \sim U[0,a](a>0)$, $h(X) = \log a$. So we can see that *differential entropy may be negative!*

2.2 Example: Normal Distribution

For $X \sim N(\mu, \sigma^2)$, we have

$$\begin{aligned} h(f) &= -\int f(x) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \mathrm{dx} + \int f(x) \frac{(x-\mu)^2}{2\sigma^2} \mathrm{dx} \\ &= \frac{1}{2}\log 2\pi\sigma^2 + \frac{DX}{2\sigma^2} \\ &= \frac{1}{2}\log 2\pi e\sigma^2 \end{aligned}$$

Note. Here we assume that the log function takes e as base.

2.3 Example: Multi-variable Normal Distribution

For $X \sim N(\mu, K)$, where K is the covariance matrix, we have

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|K|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)\right]$$

Thus

$$\begin{aligned} h(f) &= -\int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)\right] \mathrm{d}\mathbf{x} + \frac{1}{2}\log(2\pi)^n|K| \\ &= \frac{1}{2} E\left[\sum_{i,j} (X_i - \mu_i)\left(K^{-1}\right)_{ij}(X_j - \mu_j)\right] + \frac{1}{2}\log(2\pi)^n|K| \\ &= \frac{1}{2}\sum_{i,j} E\left[(X_i - \mu_i)(X_j - \mu_j)\right]\left(K^{-1}\right)_{ij} + \frac{1}{2}\log(2\pi)^n|K| \\ &= \frac{1}{2}\sum_{i,j} K_{ji}\left(K^{-1}\right)_{ij} + \frac{1}{2}\log(2\pi)^n|K| \\ &= \frac{1}{2}n + \frac{1}{2}\log(2\pi)^n|K| \\ &= \frac{1}{2}\log(2\pi e)^n|K| \end{aligned}$$

3 AEP for Differential Entropy

We also have the AEP theorem for differential entropy.

Theorem 3. (AEP) $X_1, X_2, \cdots, X_n$: i.i.d. and obeys $f(x)$. Then

$$-\frac{1}{n}\log f(X_1, X_2, \cdots, X_n) \xrightarrow{p} h(X)$$

Proof. By weak LLN. □

And we also define the typical set.

Definition. The **typical set** $A_n^{(\epsilon)}$ is the set of typical sequences, i.e.

$$A_n^{(\epsilon)} = \left\{ x^n \in S^n : \left| -\frac{1}{n}\log f(x^n) - h(X) \right| \leq \epsilon \right\}$$

And the properties of a typical set in the discrete case are also preserved. We first define the volume of a set, as *an analogy to the cardinality in the discrete case.*

Definition. The **volume** of a set $A \subset \mathbb{R}^n$ is

$$Vol(A) = \int_A \mathrm{dx_1 dx_2 \cdots dx_n}$$

Theorem 4.

1. $p(A_n^{(\epsilon)}) > 1 - \epsilon$ when n is sufficiently large.
2. $Vol(A_n^{(\epsilon)}) \leq 2^{n(h(X)+\epsilon)}$.
3. $Vol(A_n^{(\epsilon)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ when n is sufficiently large.

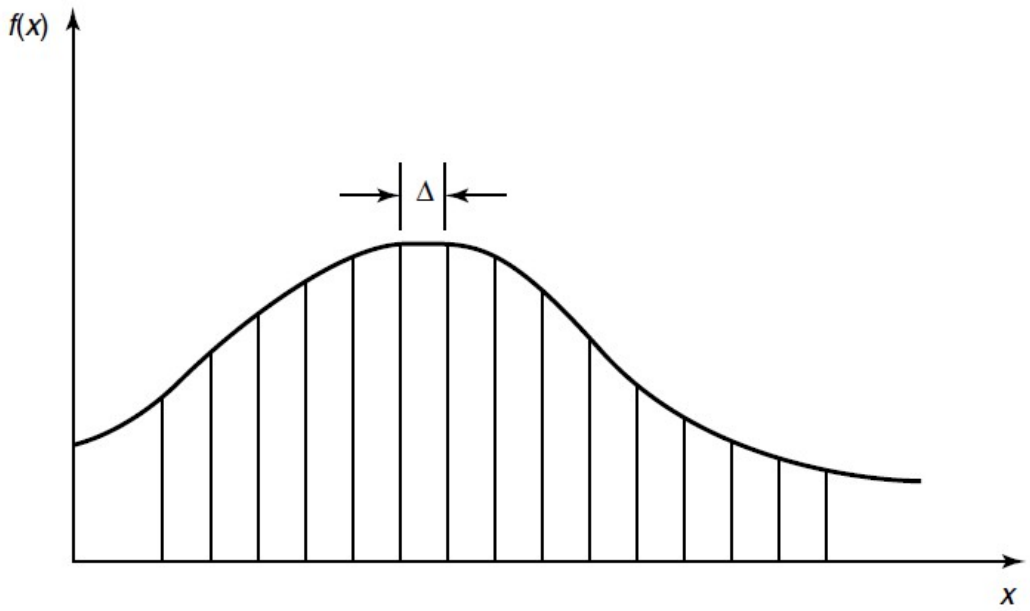
Proof. Almost the same as that in the discrete case. □

4 Discretization

Can we relate the differential entropy to the discrete entropy?

Consider a random variable X with density $f(x)$ illustrated as follows. Suppose that we divide the range of X into bins of length Δ. Let us assume that the density is continuous within the bins. Then, by the mean value theorem, there exists a value x_i within each bin such that

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



Consider the quantized random variable X^Δ, which is defined by

$$X^\Delta = x_i, \text{for } X \in [i\Delta, (i+1)\Delta)$$

Then

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

which leads to

$$H(X^\Delta) = -\sum \Delta f(x_i) \log f(x_i) - \log \Delta$$

Theorem 5. If $f(x)$ is Riemann integrable, then

$$\lim_{\Delta \to 0} H(X^\Delta) + \log \Delta = h(f)$$

Note. Thus, the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$. And in general, $h(X) + n$ is the number of bits on the average required to describe X to n-bit accuracy. So the more precise we quantize X, the more bits we will need.

5 Other Definitions

5.1 Straightforward Ones

We can also define joint differential entropy and conditional differential entropy.

Definition. The **joint differential entropy** of $X_1, X_2, \cdots, X_n$ is

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

The **conditional differential entropy** of X, Y with a joint density function $f(x, y)$ is

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

The **relative entropy** (or **Kullback–Leibler distance**) of f, g is

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

The **mutual information** of X, Y is

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

Note.

(1) Since in general $f(x|y) = \frac{f(x,y)}{f(y)}$, we have $h(X|Y) = h(X, Y) - h(Y)$. But we must be careful if any of the differential entropies are infinite.

(2) $D(f\|g)$ is finite only if the support set of f is contained in the support set of g. (Motivated by continuity, we set $0 \log 0 = 0$.)

(3) The information diagram still applies. So $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y)$. But now, since $h(X)$ may be negative, we should check the sign of an area carefully.

They also preserves many properties in the discrete case.

Theorem 6.

1. $h(X|Y) \leq h(X)$, or $I(X;Y) \geq 0$. The equality holds iff X, Y are independent.
2. $D(f\|g) \geq 0$. The equality holds iff $f = g$ almost everywhere. That is, $\{x \in S_f : f(x) \neq g(x)\}$ has a zero measure.
3. (Chain Rule) $h(X_1, \cdots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, \cdots, X_{i-1})$
4. (Independence Bound) $h(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} h(X_i)$.

Note. The chain rule and the independence bound holds mainly because of the linearity of the expected value.

Here is an example picked from [Cover]:

Example *(Mutual information between correlated Gaussian random variables with correlation ρ)* Let $(X, Y) \sim \mathcal{N}(0, K)$, where

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

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

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log(1-\rho^2).$$

If $\rho = 0$, X and Y are independent and the mutual information is 0. If $\rho = \pm 1$, X and Y are perfectly correlated and the mutual information is infinite.

5.2 Generalized Mutual Information

We can generalize the definition of mutual information. We first define the quantization of a random variable.

Definition. Let $\mathcal{P} = \{P_i\}$ be a partition over the domain of X. The **quantization** of X by $\mathcal{P}$ (denoted $[X]_\mathcal{P}$) is the discrete random variable defined by

$$p([X]_\mathcal{P} = i) = p(X \in P_i) = \int_{P_i} \mathrm{dF(x)}$$

We can see that X^Δ is one kind of quantization of X.

Since $[X]_\mathcal{P}$ is a discrete random variable, for two random variables X and Y with partitions $\mathcal{P}$ and $\mathcal{Q}$, we can calculate the mutual information between the quantized versions of X and Y.

Definition. (Master Definition) The mutual information between two random variables X and Y is given by

$$I(X; Y) = \sup_{\mathcal{P},\mathcal{Q}} I\left([X]_\mathcal{P}; [Y]_\mathcal{Q}\right)$$

where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

Note. We can prove that this definition is equivalent with the former definitions, for both differential entropies and discrete entropies.

Remark. This definition is called "master definition" because it can also apply to many other cases, such as joint distributions with atoms, densities, and singular parts. And it allows us to calculate the mutual information between a continuous random variable and a discrete random variable.

6 Compared To Discrete Entropy

6.1 Differences

1. $h(X)$ can be negative. Example: for $X \sim U[0, a], h(X) = \log a$.
2. $h(X)$ does not serve as a measure of the average amount information contained in a continuous random variable.
3. $I(X;Y)$ is generalized.

6.2 Common Properties

1. The chain rule holds.
2. The relative entropy and the mutual information are non-negative. The conditions for equality are almost the same.
3. The conditional entropy is less than the entropy.

In a word, they are almost the same in the definitions other than entropy itself.

Acknowledgment

The contents are mainly based on the course materials of CS258, 2020 Spring, Shanghai Jiao Tong University and *Elements of Information Theory* by Thomas M. Cover and Joy A. Thomas.