ECE253/CSE208 Introduction to Information Theory

Lecture 13: Differential Entropy

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 8 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*.

Differential Entropy

Definition

The differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ and support $\mathcal{S}$ is defined as

$$\boxed{h(X) = \mathrm{E}(-\log f(X)) = -\int_{\mathcal{S}} f(x) \log f(x) dx}$$

Example (Differential entropy can be negative)

Consider $X \sim$ Uniform$[0, a]$, its differential entropy is

$$h(X) = -\int_0^a \frac{1}{a} \log\left(\frac{1}{a}\right) dx = \log a \implies h(X) < 0 \text{ for } 0 < a < 1$$

Example (Entropy of normal distribution)

Let $X \sim \mathcal{N}(0, \sigma^2)$ with the pdf $\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. Then, we have

$$h(X) = -\int \phi(x) \ln \phi(x)\ dx = \frac{\mathrm{E}X^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2) \text{ nats} = \frac{1}{2}\log(2\pi e\sigma^2) \text{ bits}$$

AEP for Continuous Random Variables

Theorem

Let i.i.d. $X^n \sim f(x)$. *Then* $-\frac{1}{n}\log f(X^n) \xrightarrow{\text{i.p.}} \mathrm{E}(-\log f(X)) = h(X)$.

Definition (Typical set)

$A_\epsilon^{(n)} = \left\{x^n \in S^n : |-\frac{1}{n}\log f(x^n) - h(X)| \le \epsilon\right\}$, where $f(x^n) = \prod_{i=1}^n f(x_i)$.

Properties of the typical set.

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,

where the volume of a set $A \subset \mathbb{R}^n$ is defined as $\mathrm{Vol}(A) = \int_A dx_1 dx_2 \cdots dx_n$.

Implications of Differential Entropy

Theorem (cf. section 3.3 in the book)

$A_\epsilon^{(n)}$ is the smallest volume set w.p. at least $1-\epsilon$, to first order in the exponent.

Implication.

1. The volume of the smallest set that contains most of the probability $\approx 2^{nh}$.
2. The corresponding side length is $(2^{nh})^{\frac{1}{n}} = 2^h \implies h(X)$ is the logarithm of the *equivalent side length* of the smallest set that contains most of the probability.
3. A random variable with low entropy is confined to a small effective volume, and widely dispersed if it has a high entropy.
4. $h(X)$ is related to $\mathrm{Vol}(A_\epsilon^{(n)})$. Fisher information is related to the surface area of $A_\epsilon^{(n)}$; see details in Sections 11.10 and 17.8.

Theorem (On average $h(X)+n$ bits are required to describe X to n-bit accuracy.)

Consider a continuous random variable $X \sim f(x)$ and its quantized version $X^\Delta = x_i$ for $i\Delta \le X < (i+1)\Delta$, where $f(x_i) = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx$. If X is Riemann integrable, then $H(X^\Delta) + \log\Delta \xrightarrow{\Delta\to 0} h(X)$.

Joint, Conditional, Relative Entropy, and Mutual Information

Definition

$$h(X^n) = -\int f(x^n) \log f(x^n)\, dx^n \tag{1}$$

$$h(X|Y) = -\int f(x,y) \log f(x|y)\, dxdy = h(X,Y) - h(Y) \tag{2}$$

$$D(f||g) = \int f \log \frac{f}{g} \tag{3}$$

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}}, [Y]_{\mathcal{Q}}) \tag{4}$$

$$= D\left(f(x,y)||f(x)f(y)\right) \tag{5}$$

$$= h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y) \tag{6}$$

The supremum is taken over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$. The quantization of X by $\mathcal{P}$ is defined as $\Pr([X]_{\mathcal{P}} = i) = \Pr(X \in P_i) = \int_{P_i} f(x)\, dx$, where the disjoint sets P_i's form a partition of the range of X such that $\cup_i P_i = \mathcal{X}$.

Entropy of Gaussian Distribution

Theorem (Entropy of multivariate normal distribution)

$$\boxed{h(\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})) = \frac{1}{2}\log(2\pi e)^n |\mathbf{K}| \text{ bits}}$$

Proof: $\phi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$

$$\begin{aligned}
h(\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})) &= -\int \phi(\mathbf{x}) \ln \phi(\mathbf{x})\, d\mathbf{x} \\
&= -\int \phi(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \ln(2\pi)^{n/2}|\mathbf{K}|^{1/2}\right] d\mathbf{x} \\
&= \frac{1}{2} \times \mathrm{E}\left[\mathrm{Tr}\left(\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right)\right] + \frac{1}{2}\ln(2\pi)^n|\mathbf{K}| \\
&= \frac{1}{2} \times \mathrm{Tr}\left(\mathrm{E}\left[\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\right) + \frac{1}{2}\ln(2\pi)^n|\mathbf{K}| \\
&= \frac{1}{2} \times \mathrm{Tr}\left(\mathbf{K}^{-1}\mathbf{K}\right) + \frac{1}{2}\ln(2\pi)^n|\mathbf{K}| \\
&= \frac{1}{2}\ln(2\pi e)^n|\mathbf{K}| \text{ nats} \\
&= \frac{1}{2}\log(2\pi e)^n|\mathbf{K}| \text{ bits}
\end{aligned}$$

Entropy of Gaussian Distribution (Cont'd)

Example

Let $(X, Y) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \implies h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$

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

$$\begin{cases} X \perp\!\!\!\perp Y \ (\rho = 0) & \Leftrightarrow I(X;Y) = 0 \\ X \parallel Y \ (\rho = \pm 1) & \Leftrightarrow I(X;Y) = \infty \end{cases}$$

Properties of Differential Entropy and KL Divergence

Similar to the discrete case, we have

- $D(f||g) \geq 0 \implies I(X;Y) \geq 0$, $h(X|Y) \leq h(X)$.
- $h(X^n) = \sum_{i=1}^{n} h(X_i|X^{i-1}) \leq \sum_i h(X_i)$.
- $h(X+c) = h(X)$ for any constant c.

Different from the discrete case, for a continuous random vector $\mathbf{x} \in \mathbb{R}^n$, we have

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

For one dimension: $h(aX) = h(X) + \log|a|$, which can be proved by the property $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$ for $Y = aX$.

Maximum Entropy

Theorem (Normal distribution maximizes the entropy for a given covariance)

Let the random vector $\mathbf{X} \in \mathbb{R}^n$ have zero mean and covariance $\mathbf{K}$. Then,

$$\max_{\mathrm{E}(\mathbf{X}\mathbf{X}^\top)=\mathbf{K}} h(\mathbf{X}) = \frac{1}{2}\log(2\pi e)^n|\mathbf{K}|,$$

where the maximum is attained iif $\mathbf{X} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{K})$.

Proof: Let $g(\mathbf{x})$ be any density function with covariance $\mathbf{K}$, and $\phi(\mathbf{x}) \sim \mathcal{N}_n(\mathbf{0}, \mathbf{K})$. Then, we have

$$0 \leq D(g||\phi) = -h(g) - \int g \log \phi = -h(g) - \int \phi \log \phi = -h(g) + h(\phi),$$

where the second equality is due to the fact g and ϕ have the same covariance matrix $\mathbf{K}$.

Minimum Estimation Error

Theorem (Estimation error)

For any random variable X and estimator $\hat{X}$, we have $\mathrm{E}(X-\hat{X})^2 \geq \frac{1}{2\pi e}e^{2h(X)}$ with equality iif X is Gaussian with mean $\hat{X}$.

Proof: $\mathrm{E}(X-\hat{X})^2 \geq \min_{\hat{X}} \mathrm{E}(X-\hat{X})^2 = \mathrm{E}(X-\mathrm{E}(X))^2 = \mathrm{Var}(X) \geq \frac{1}{2\pi e}e^{2h(X)}$.

Corollary (Estimation error with side information)

Given side information Y and estimator $\hat{X}(Y)$, it follows that

$$\mathrm{E}(X-\hat{X}(Y))^2 \geq \frac{1}{2\pi e}e^{2h(X|Y)}.$$

Thank You!

Email: <zhangy@ucsc.edu>

Homepage: https://people.ucsc.edu/~yzhan419/