Differential Entropy

吴家麟教授

Definition

- Let X be a random variable with cumulative distribution function $\mathrm{F(x)} = \mathrm{P_r(X \le x)}$. If F(x) is continuous, the r.v. is said to be continuous. Let f(x) = F'(x) when the derivative is defined. If $\int_{-\infty}^{\infty} f(x)dx = 1$, then f(x) is called the pdf for X. The set where f(x) > 0 is called the support set of X.

Definition

- The differential entropy $h(X)$ of a continuous r.v. X with a density function f(x) is defined as

$$h(X) = -\int_s f(x) \log f(x)\, dx \qquad (1)$$

where S is the support of the r.v.

- Since $h(X)$ depends only on f(x), sometimes the differential entropy is written as $h(f)$ rather then $h(X)$.

EX.1: (Uniform distribution)

$$f(x)=\begin{cases}\frac{1}{a}, & 0 \le \mathrm{x} < \mathrm{a} \\ 0, & \text{otherwise}\end{cases}$$

$$h(X)=-\int_0^a \frac{1}{a}\log\frac{1}{a}\,dx=\log a$$

Note: For a<1, log a < 0, and $h(X)=\log a<0$.

However, $2^{h(X)}=2^{\log a}=a$ is the volume of the support set, which is always non-negative.

Ex. 2: (Normal distribution)

- Let $X \sim \Phi = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$, then

$$h(\Phi) = -\int \Phi \ln \Phi$$

$$= -\int \Phi(x)\left[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right]$$

$$= \frac{1}{2\sigma^2} E\left[X^{-2}\right] + \ln\left(2\pi\sigma^2\right)^{\frac{1}{2}}$$

$$= \frac{1}{2\sigma^2}\sigma^2 + \frac{1}{2}\ln 2\pi\sigma^2$$

$$= \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2$$

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

$$= \frac{1}{2}\ln 2\pi e\sigma^2 \qquad \text{nats}$$

Changing the base of the logarithm, we have

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

Theorem 1

- Let $X_1, X_2, \cdots, X_n$ be a sequence of rv's drawn i.i.d. according to the density f(x). Then

$$-\frac{1}{n}\log f\left(X_1, X_2, \cdots X_n\right) \rightarrow$$
$$E\left[-\log f\left(X\right)\right] = h\left(X\right) \quad \text{in probability}$$

proof: The proof follows directly from the weak law of the large numbers.

- Def: For ε>0 and any n, we define the typical set $A_\varepsilon^{(n)}$ w.r.t. f(x) as follows:

$$A_\varepsilon^{(n)} = \left\{ \left(X_1, X_2, \cdots X_n\right) \in S^n : \left| -\frac{1}{n} \log f\left(X_1, X_2, \cdots X_n\right) - h(X) \right| \le \varepsilon \right\},$$

$$\text{where } f\left(X_1, X_2, \cdots X_n\right) = \Pi_{i=1}^{n} f\left(X_i\right)$$

- Def: The volume Vol(A) of a set $A \in R^n$ is defined as

$$vol(A) = \int_A dx_1 dx_2 \cdots dx_n$$

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

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

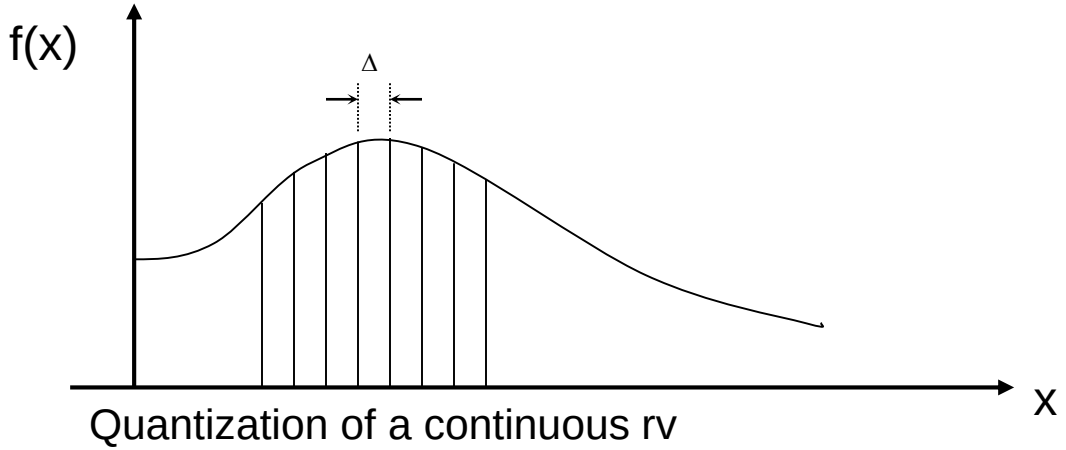
- Thm: The Set $A_{\epsilon}^{(n)}$ is the smallest volume set with probability $\geq 1-\epsilon$, to the first order in the exponent.

  $\Rightarrow$ The volume of the smallest set that contains most of the Prob. Is approximately 2^{nh}. This is an n-D volume, so the corresponding side length is $(2^{nh})^{1/n}=2^{h}$.

  $\Rightarrow$ The differential entropy is the logarithm of the equivalent side length of the smallest set that contains most of the Prob.

  $\Rightarrow$ low entropy implies that the rv is confined to a small effective volume and high entropy indicates that the rv is widely dispersed.

Relation of Differential Entropy to Discrete Entropy



Quantization of a continuous rv

S'pose we divide the range of X into bins of length Δ. Let's assume that the density is continuous within the bins.

By the mean value theorem, there is a value x_i within each bin such that

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

Consider the quantized rv X^Δ , which is defined by

$$X^\Delta = x_i, \quad \text{if} \quad i\Delta \le X < (i+1)\Delta$$

Then the prob. that $X^\Delta = x_i$ is

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

The entropy of the quantized version is

$$
\begin{aligned}
H\left(X^{\Delta}\right) &= -\sum_{-\infty}^{\infty} P_i \log P_i \\
&= -\sum_{-\infty}^{\infty} f\left(x_i\right)\Delta \log\left(f\left(x_i\right)\Delta\right) \\
&= -\sum \Delta f\left(x_i\right)\log f\left(x_i\right) - \sum f\left(x_i\right)\Delta \log \Delta \\
&= -\sum \Delta f\left(x_i\right)\log f\left(x_i\right) - \log \Delta
\end{aligned}
$$

Since

$$\sum f\left(x_i\right)\Delta = \int f\left(x\right) = 1$$

If f(x)log f(x) is Riemann integrable, then

$$-\sum \Delta f(x_i)\log f(x_i) \;\rightarrow\; -\int f(x)\log f(x)\,\mathrm{dx}\,,\ \text{as}\ \Delta \rightarrow 0$$

This proves the following

Thm: If the density f(x) of the rv X is Riemann integrable, then

$$H(X^{\Delta}) + \log\Delta \rightarrow h(f) = h(X), \quad \text{as}\ \Delta \rightarrow 0$$

Thus the entropy of an n-bit quantization of a continuous rv X is approximately $h(X)+n$

Since $\Delta = 2^{-n}$ for a n - bit uniform quantizer

Joint and Conditional Differential Entropy:

Riemann Integrable : a condition to ensure that the limit is well defined.

- $h(X_1, X_2, \cdots, X_n) = -\int f(x^n) \log f(x^n) dx^n$
- $h(X \mid Y) = -\int f(x, y) \log f(x \mid y) dx dy$
- $h(X \mid Y) = h(X, Y) - h(Y)$

Theorem (Entropy of a multivariate normal distribution)

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

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

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

$$\text{pf}: (X_1, X_2, \cdots, X_n) \sim N_n(\mu, K) \Rightarrow f(\tilde{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\tilde{x}-\mu)^T K^{-1}(\tilde{x}-\mu)}$$

Then

$$h(f) = -\int f(x)\left[-\frac{1}{2}(\tilde{x}-\mu)^T K^{-1}(\tilde{x}-\mu) - \ln\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}}\right] dx$$

$$= \frac{1}{2} E\left[\sum_{ij}(X_i - \mu_i)\left(K^{-1}\right)_{ij}(X_j - \mu_j)\right] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2} E\left[\sum_{ij}(X_i - \mu_i)(X_j - \mu_j)\left(K^{-1}\right)_{ij}\right] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2}\sum_{ij} E\left[\left(X_j - \mu_j\right)\left(X_i - \mu_i\right)\right]\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n\left|K\right|$$

$$= \frac{1}{2}\sum_{j}\sum_{i} K_{ji}\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n\left|K\right|$$

$$= \frac{1}{2}\sum_{j}\left(KK^{-1}\right)_{jj} + \frac{1}{2}\ln(2\pi)^n\left|K\right|$$

$$= \frac{1}{2}\sum_{j} I_{jj} + \frac{1}{2}\ln(2\pi)^n\left|K\right|$$

$$= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n\left|K\right|$$

$$= \frac{1}{2}\ln(2\pi e)^n\left|K\right| \quad \text{nats}$$

$$= \frac{1}{2}\log(2\pi e)^n\left|K\right| \quad \text{bits}$$

Relative Entropy and Mutual Information

- $D(f \,//\, g) = \int f \log \frac{f}{g}$
- $I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)\,f(y)}\, dxdy$

- $I(X;Y) = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X) = h(X) + h(Y) - h(X,Y)$
- $I(X;Y) = D\big(f(x,y) \,//\, f(x)\,f(y)\big)$

Remark:

The mutual information between two continuous r.v's is the limit of the mutual information between their quantized versions.

$$I(X^{\Delta};Y^{\Delta}) = H(X^{\Delta}) - H(X^{\Delta}\,|\,Y^{\Delta})$$
$$\approx h(X) - \log\Delta - \left(h(X\,|\,Y) - \log\Delta\right) = I(X;Y)$$

Properties of $h(x), \mathrm{D}(\mathrm{p} /\!/ \mathrm{q})$, $\mathrm{I}(\mathrm{X};\mathrm{Y})$

- $\mathrm{D}(\mathrm{f} /\!/ \mathrm{g}) \geq 0$

  $\mathrm{pf}: -\mathrm{D}(\mathrm{f} /\!/ \mathrm{g}) = \int_s \mathrm{f} \log \frac{\mathrm{g}}{\mathrm{f}} \leq \log \int_s \mathrm{f} \frac{\mathrm{g}}{\mathrm{f}}$ (Jenser's inequality)

  $= \log \int_s g \leq \log 1 = 0.$
- $\mathrm{I}(\mathrm{X};\mathrm{Y}) \geq 0$
- $\mathrm{h}(\mathrm{X} \mid \mathrm{Y}) \leq \mathrm{h}(\mathrm{X})$

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

with equality iff X1, X2, ..., Xn are independent!

Theorems
- $h(X + c) = h(x)$: translation does not change the differential entropy
- $h(aX) = h(x) + \log|a|$

$$pf: \text{let } \mathrm{Y} = \mathrm{aX}. \text{ Then }, \mathrm{f_Y}(y) = \frac{1}{|a|} f_x(\frac{y}{a}), and$$

$$\mathrm{h(aX)} = -\int \mathrm{f_Y}(y) \log \mathrm{f_Y(y)}\, d_y$$

$$= -\int \frac{1}{|a|} f_x(\frac{y}{a}) \log (\frac{1}{|\mathrm{a}|} \mathrm{f_x}(\frac{\mathrm{y}}{\mathrm{a}}))\mathrm{dy}$$

$$= -\int \mathrm{f_x}(x) \log_x(x) dx + \log|a| = h(x) + \log|a|$$

$$\text{Corollary}: \mathrm{h(AX)} = \mathrm{h(X)} + \log|\det(\mathrm{A})|$$

Theorem : The multivariate normal distribution maximizes the entropy over all distributions with the same variance.

Let the random vector $X \in R^n$ have zero mean and convariance $K = E\,XX^T$ (i.e., $K_{ij} = EX_iX_j$, $1 \le i, j \le n$) , Then

$$h(x) \le \frac{1}{2} \log (2\Pi e)^n |K|, \text{with equality iff } X \sim N_n(0, k)$$

Pf:

Let $\mathrm{g}(\tilde{\mathrm{x}})$ be any density satisfying $\int \mathrm{g}(\tilde{x}) x_i x_j \, d\tilde{x} = K_{ij}$ for all i, j.

Let ϕ_{k} be the density of a $\mathrm{N}(0, \mathrm{K})$ vector.

Note that $\log \phi_{\mathrm{k}}(\tilde{\mathrm{x}})$ is a quadratic form and $\int x_i x_j \phi_{\mathrm{k}}(\tilde{\mathrm{x}}) d\tilde{x} = K_{ij}$.

$$0 \le D(g \,/\!/\, \phi_{\mathrm{k}}) = \int g \log(g / \phi_k)$$

$$= -h(g) - \int g \log \phi_{\mathrm{k}} = -h(g) - \int \phi_{\mathrm{k}} \log \phi_{\mathrm{k}}$$

$$= -h(g) + h(\phi_{\mathrm{k}})$$

where the substitution $\int g \log \phi_k$ follows from the fact

that g and ϕ_k yield the same moments of the quadraic form $\log \phi_k(x)$

$\Rightarrow$ the Gaussian distribution maximizes the entropy over all distributions with the same variance.

Let X be a random variable with differential entropy h(x)

Let $\hat{X}$ be an estimate of X and let . $E(X - \hat{X})^2$ be the expected prediction error. Let h(x) be in nats

Theorem : For any r.v. X and estimator $\hat{X}$

$$E(X - \hat{X})^2 \geq \frac{1}{2\Pi e} e^{2h(x)}$$

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

pf : Let $\hat{X}$ be any estimator of X then

$$E(X - \hat{X})^2 \geq \min_{\hat{x}} E(X - \hat{X})^2 \qquad (1)$$

$$= E(X - E(X))^2 \quad [\text{the mean of } X \text{ is the best estimator for } X]$$

$$= \text{var}(X) \geq \frac{1}{2\Pi e} e^{2h(x)} \qquad (2)$$

[Gaussian distribution has the maximum entropy for a given varance]

$$\text{i.e., } h(x) \leq \frac{1}{2} \ln 2\Pi e \sigma^2$$

We have equality, only in (1), only if $\hat{x}$ is the best estimator

(i.e. , $\hat{x}$ is the mean of X) and equality in (2) only if X is Gaussian.

Gorollary : Given side information Y and estimator $\hat{X}(Y)$

it follows that

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\Pi e} e^{2h(X \mid Y)}$$

$\rightarrow$ Fano's inequality