INFORMATION THEORY & CODING

Differential Entropy 2

Dr. Rui Wang

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

Email: wang.r@sustech.edu.cn

December 27, 2022

Differential Entropy - 2

- Definitions
- AEP for Continuous Random Variables
- Relation of differential entropy to discrete entropy
- Joint and Conditional Differential Entropy
- Relative Entropy and Mutual Information
- Estimation Counterpart of Fano's Inequality

Joint and conditional differential 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 $(X_1, X_2, \ldots, X_n) \sim \mathcal{N}_n(\mu, K)$.

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.

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

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

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

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

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

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

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

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.

Mutual information between correlated Gaussian r.v.s

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

$$K = \left[\begin{array}{cc} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{array} \right].$$

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

  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

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

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)

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

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

Properties of differential entropy

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

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

□

Properties of differential entropy

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

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

□

Properties of differential entropy

Theorem (Translation does not change the differential entropy)

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

Theorem

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

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 $\mathbf{X} \sim \mathcal{N}(0, K)$*.*

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.

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

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

□

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

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY