CS258: Information Theory

Fan Cheng
Shanghai Jiao Tong University

http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

Outline

- Differential Entropy
- AEP for Continuous Random Variable
- Relative Entropy and Mutual Information
- Property of Differential Information Measures
- Information inequalities and applications

Differential Entropy

- Let X be a random variable with **cumulative distribution function**
$$F(x) = \Pr(X \leq x)$$
- If $F(x)$ is continuous, the random variable is said to be **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** for X.
- The set where $f(x) > 0$ is called the **support set** of X.

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

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

where S is the support set of the random variable.
The differential entropy is sometimes written as $h(f)$ rather than $h(X)$

- $h(X+c) = h(X)$ (Translation does not change the differential entropy)

$$\boldsymbol{p(x) \Rightarrow f(x)}$$
$$\sum \Rightarrow \int$$
$$\boldsymbol{H(X) \Rightarrow h(X)}$$
$H(X)$ is always non-negative. $h(X)$ may be negative

Differential Entropy: Example

- Consider a random variable distributed uniformly from 0 to a, then $\boldsymbol{h(X) = \log a}$
- Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\boldsymbol{h(X) = \frac{1}{2}\log 2\pi e\sigma^2}$

- When X is uniformly distributed in $[0, a]$,

$$f(x) = 1/a$$

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

- When X is Gaussian $\mathcal{N}(\mu, \sigma^2)$, then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$h(f(x)) = -\int f(x)\log f(x)\,dx$$

$$= -\int f(x)\log\frac{1}{\sqrt{2\pi\sigma^2}} + f(x)\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx$$

$$\int f(x)dx = 1 \quad \text{and} \quad \mathrm{Var}\,(X) = \int (x-\mu)^2 f(x)dx = \sigma^2$$

Mean and Variance

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

$h(X)$: Infinite Information

- Differential entropy **does not serve as a measure of the average amount of information** contained in a continuous random variable.
- In fact, **a continuous random variable generally contains an infinite amount of information**

Let X be uniformly distributed on $[0,1)$. Then we can write

$$X = 0.X_1X_2, \dots .$$

The dyadic expansion of X, where $X_i's$ is a sequence of i.i.d bits.

Then

$$\begin{aligned} H(X) &= H(X_1, X_2, \dots) \\ &= \sum_{i=1}^{\infty} H(X_i) \\ &= \sum_{i=1}^{\infty} 1 \\ &= \infty \end{aligned}$$

Differential entropy does not serve as a measure of the average amount of information contained in X

--Ch. 10, R. W. Yeung, Information theory and Network Coding

$h(aX)$

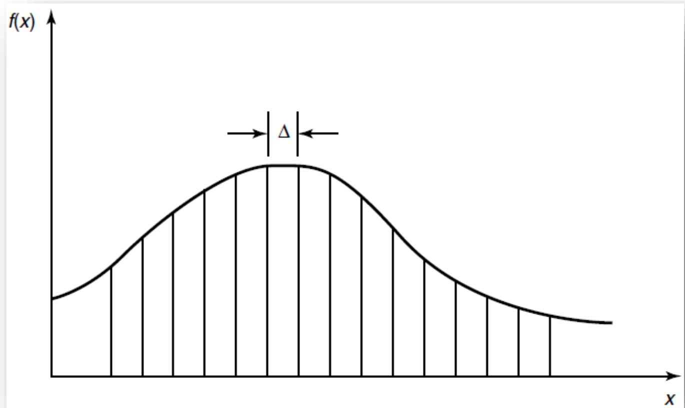
$$\boldsymbol{h(aX) = h(X) + \log|a|}$$
$$\boldsymbol{h(AX) = h(X) + \log|\det A|}$$

Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and

$$h(aX) = -\int f_Y(y) \log f_Y(y)\, dy$$
$$= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy$$
$$= -\int f_X(x) \log f_X(x)\, dx + \log|a|$$
$$= h(X) + \log|a|$$

$$\boldsymbol{h(AX)} = \boldsymbol{h(X)} + \log|\det(\boldsymbol{A})|$$

Differential and Discrete Entropy



- Suppose that we divide the range of X into bins of length $\boldsymbol{\Delta}$.
- By the mean value theorem, there exists a value x_i within each bin such that

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

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

$$X^\Delta = x_i \text{ if } i\Delta \leq x < (i+1)\Delta$$

- Then the probability that $X^\Delta = x_i$ is

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

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

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

AEP For Continuous Random Variable

- AEP for continuous random variables:
  Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to the density $f(x)$. Then
  $$-\frac{1}{n}\log f(X_1, X_2, \ldots, X_n) \rightarrow E(-\log f(X)) = h(f)$$
  in probability
- For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:
  $$A_\epsilon^{(n)} = \left\{(x_1, x_2, \ldots, x_n) \in S^n: \left|-\frac{1}{n}\log f(x_1, x_2, \ldots, x_n) - h(X)\right| \le \epsilon\right\}$$
  where $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.

- The volume of a set $A \subset \mathcal{R}^n$ is defined as
  $$\mathrm{Vol}(A) = \int_A dx_1 dx_2 \ldots dx_n.$$
- The typical set $A_\epsilon^{(n)}$ has the following properties:
  - 1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ 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.

$2^{nh(X)}$ is the volume

$h(X_1, X_2, \dots, X_n)$ and $h(X|Y)$

- The differential entropy of a set $X_1, X_2, \dots, X_n$ of random variables with density $f(x_1, x_2, \dots, x_n)$ is defined as

$$h(X_1, X_2, \dots, X_n) = -\int f(x^n) \log f(x^n) dx^n$$

- If X, Y have a joint density function $f(x, y)$, we can define the conditional differential entropy $h(X|Y)$ as

$$h(X|Y) = -\int f(x,y) \log f(x|y)\, dx\, dy.$$

$$h(X|Y) = h(X,Y) - h(Y)$$

- $h(X|Y) \le h(X)$
  with equality iff X and Y are independent.
- (Chain rule for differential entropy)

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

- $h(X_1, X_2, \dots, X_n) \le \sum_{i=1}^{n} h(X_i)$
  with equality iff $X_1, X_2, \dots, X_n$ are independent.

Covariance Matrix

- The **covariance** between two random variables X and Y is defined as

$$\mathbf{cov}(X;Y) = E(X - EX)(Y - EY) = E(XY) - (EX)(EY)$$

- For a random vector $X = [X_1, X_2, \ldots, X_n]^T$, the **covariance matrix** is defined as

$$K_X = E(X - EX)(X - EX)^T = [\mathbf{cov}(X_i; X_j)]$$

  and the **correlation matrix** is defined as

$$\tilde{K}_X = EXX^T = [EX_iX_j]$$

- $K_X = EXX^T - (EX)(EX^T) = \tilde{K}_X - (EX)(EX^T)$

- A covariance matrix is both symmetric and positive semidefinite.
  - The eigenvalues of a positive semidefinite matrix are non-negative.
- Let $Y = AX$, where X and Y are column vectors of n random variables and A is an $n \times n$ matrix. Then

$$K_Y = AK_XA^T$$

and

$$\tilde{K}_Y = A\tilde{K}_YA^T$$

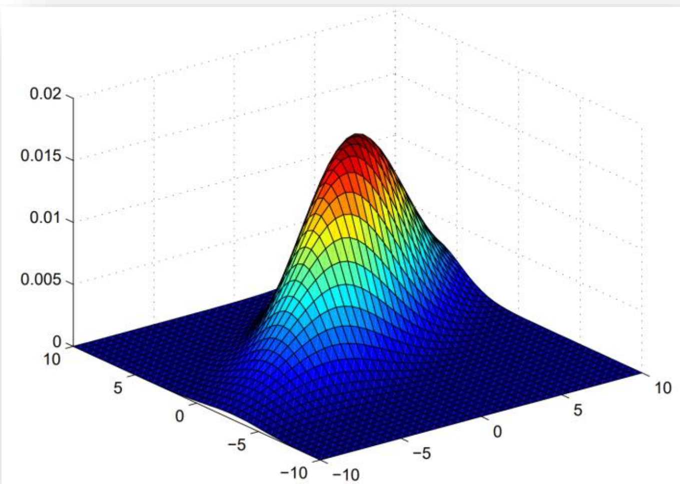
A set of correlated random variables can be regarded as an orthogonal transformation of a set of uncorrelated random variables.
--Ref: Ch. 10.1 Yeung, Information theory and network coding

Multivariate Normal Distribution

- In probability theory and statistics, the **multivariate normal distribution, multivariate Gaussian distribution, or joint normal distribution** is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions.
- More generally, let $\mathcal{N}(\mu, K)$ denote the multivariate Gaussian distribution with mean μ and covariance matrix K, i.e., the joint pdf of the distribution is given by

$$\boldsymbol{f}(\mathrm{x}) = \frac{\mathbf{1}}{\left(\sqrt{\mathbf{2}\boldsymbol{\pi}}\right)^{\boldsymbol{n}} |\boldsymbol{K}|^{\mathbf{1/2}}} \boldsymbol{e}^{-\frac{\mathbf{1}}{\mathbf{2}}(\mathrm{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{K}^{-\mathbf{1}}(\mathrm{x}-\boldsymbol{\mu})}$$

- One definition is that a random vector is said to be k-variate normally distributed if every linear combination of its k components has a univariate normal distribution.



- In general, random variables may be uncorrelated but statistically dependent.
- But if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent.
- This implies that any two or more of its components that are pairwise independent are independent.

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Entropy of Multivariate Normal Distribution

(Entropy of a multivariate normal distribution) Let $X_1, X_2, \dots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K

$$\boldsymbol{h(X_1, X_2, \dots, X_n) = h(\mathcal{N}(\mu, K)) = \frac{1}{2}\log(2\pi e)^n |K|}$$

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

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

$h(AX) = h(X) + \log|\det(A)|$
Ref: Ch. 10.3 Yeung

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

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

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

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

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

$$= \frac{1}{2}\sum_j I_{jj} + \frac{1}{2}\ln(2\pi)^n|K|$$

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

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

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

Relative Entropy

- The **relative entropy (or Kullback–Leibler distance) $D(f||g)$** between two densities f and g is defined by

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

- The **mutual information $I(X;\ Y)$** between two random variables with joint density $f(x,y)$ is defined as

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

- $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y)$

$$I(X;Y) = D(f(x,y)||f(x)f(y))$$

- $D(f||g) \geq 0$

  with equality iff $f = g$ almost everywhere (a.e.).
- $I(X;Y) \geq 0$

  with equality iff X and Y are independent.

Mutual Information: Master Definition

The mutual information between two random variables is the limit of the mutual information between their quantized versions

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta | Y^\Delta) \\ &\approx h(X) - \log\Delta - (h(x|y) - \log\Delta) \\ &= I(X;Y) \end{aligned}$$

Definition. The mutual information between **two random variables X and Y** is given by

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

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

- Let $\mathcal{X}$ be the range of a random variable X. A partition $\mathcal{P}$ of $\mathcal{X}$ is a finite collection of disjoint sets P_i such that $\cup_i P_i = \mathcal{X}$. The quantization of X by $\mathcal{P}$ (denoted $[X]_{\mathcal{P}}$) is the discrete random variable defined by

$$\Pr([X]_P = i) = \Pr(X \in P_i) = \int_{P_i} dF(x)$$

- 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

This is the master definition of mutual information that always applies, even to joint distributions with atoms, densities, and singular parts.

Summary

Cover: 8.1, 8.2, 8.3, 8.4, 8.5, 8.6
Yeung: Ch. 10.1, 10.2