An introduction to Information Theory

Adrish Banerjee

Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh
India

Aug. 22, 2016

Lecture #11: Differential Entropy

Outline of the lecture

- Differential entropy

Outline of the lecture

- Differential entropy
- Properties of differential entropy

Differential Entropy

- Let X be a random variable with cumulative distribution function F(x). If F(x) is continuous, the random variable is said to be continuous.

Differential Entropy

- Let X be a random variable with cumulative distribution function F(x). If F(x) is continuous, the random variable is said to be continuous.
- Let f(x)=F'(x) when the derivative is defined, and $\int_{-\infty}^{\infty} f(x) = 1$, then f(x) is called the probability density function of X.

Differential Entropy

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

Differential Entropy

- Let X be a random variable with cumulative distribution function F(x). If F(x) is continuous, the random variable is said to be continuous.
- Let f(x)=F'(x) when the derivative is defined, and $\int_{-\infty}^{\infty} f(x) = 1$, then f(x) is called the probability density function of 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 a density f(x) is defined as

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

where S is the support set of the random variable

Differential Entropy

- Let X is normal distributed random variable $N(0,\ \sigma^2)$, then differential entropy is given by

$$\begin{aligned} h(X) &= -\int \phi \log \phi \, dx \\ &= -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] dx \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2 \\ &= \frac{1}{2} \ln e + \frac{1}{2} \ln 2\pi\sigma^2 \\ &= \frac{1}{2} \ln 2\pi e\sigma^2 \ \text{ nats} \\ &= \frac{1}{2} \log 2\pi e\sigma^2 \ \text{ bits} \end{aligned}$$

Differential Entropy

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

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

Differential Entropy

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

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

- Proof follows from weak law of large numbers.

Differential Entropy

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

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

- Proof follows from weak law of large numbers.
- For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to f(x) as follows

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \cdots, x_n) \in S^n : \left|-\frac{1}{n}\log f(x_1, x_2, \cdots, x_n) - h(X)\right| \leq \epsilon\right\}$$

where $f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^n f(x_i)$

Differential Entropy

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

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

Typical set $A_\epsilon^{(n)}$ has the following properties

- $Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.

Proof:

- By the AEP, we have $-\frac{1}{n} f(x_1, x_2, \cdots, x_n) = -\frac{1}{n} \sum \log f(x_i) \rightarrow h(X)$ in probability, establishing $Pr(A_\epsilon^{(n)}) > 1 - \epsilon$.

Differential Entropy

- If $\mathbf{x} \in A_\epsilon^{(n)}$, then we have

$$2^{-n(h(X)+\epsilon)} < f(\mathbf{x}) < 2^{-n(h(X)-\epsilon)}$$

Proof:

- Follows from the definition of typical set.

Differential Entropy

- $Vol(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$ for all n.

Proof:

- We have

$$\begin{aligned} 1 &= \int_{S^n} f(x_1, x_2, \cdots, x_n)\, dx_1, dx_2, \cdots, dx_n \\ &\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n)\, dx_1, dx_2, \cdots, dx_n \\ &\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)}\, dx_1, dx_2, \cdots, dx_n \\ &= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1, dx_2, \cdots, dx_n \\ &= 2^{-n(h(X)+\epsilon)} Vol(A_\epsilon^{(n)}) \end{aligned}$$

Differential Entropy

- $Vol(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ for n sufficiently large.

Proof:

- For large n we have

$$\begin{aligned} 1-\epsilon &\leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n)\, dx_1, dx_2, \cdots, dx_n \\ &\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)}\, dx_1, dx_2, \cdots, dx_n \\ &= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1, dx_2, \cdots, dx_n \\ &= 2^{-n(h(X)-\epsilon)} Vol(A_\epsilon^{(n)}) \end{aligned}$$

Differential Entropy

- The entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Proof:

- Consider a random variable X with density $f(x)$.

Differential Entropy

- The entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Proof:

- Consider a random variable X with density $f(x)$.
- We divide the range of X into bins of length Δ. Also, there exists a value x_i with each bin such that

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

Differential Entropy

- The entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Proof:

- Consider a random variable X with density $f(x)$.
- We divide the range of X into bins of length Δ. Also, there exists a value x_i with 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 as

$$X^\Delta = x_i, \quad i\Delta \leq X < (i+1)\Delta$$

Differential Entropy

- The probability that $X^\Delta = x$ is given by

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

Differential Entropy

- The probability that $X^{\Delta} = x$ is given by

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

- The entropy of the quantized version is

$$\begin{aligned} H(X^{\Delta}) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\ &= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log\left(f(x_i)\Delta\right) \\ &= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\ &= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta \\ &= h(X) - \log \Delta \quad (as\ \Delta \to 0) \end{aligned}$$

Differential Entropy

- If X has a uniform distribution on $[0, 1]$, and let $\Delta = 2^{-n}$, then we have $h(X) = 0, H(X^{\Delta}) = n$ implying n bits suffice to describe X to n-bit accuracy.

Differential Entropy

- If X has a uniform distribution on $[0,1]$, and let $\Delta = 2^{-n}$, then we have $h(X) = 0, H(X^{\Delta}) = n$ implying n bits suffice to describe X to n-bit accuracy.
- The differential entropy of a set $X_1, X_2, \cdots, X_n$ of a random variables with density $f(x_1, x_2, \cdots, x_n)$ is defined as

$$h(X_1, X_2, \cdots, X_n) = -\int f(x_1, x_2, \cdots, x_n) \log f(x_1, x_2, \cdots, x_n)\, dx_1 dx_2 \cdots dx_n$$

Differential Entropy

- If X has a uniform distribution on $[0,1]$, and let $\Delta = 2^{-n}$, then we have $h(X) = 0, H(X^{\Delta}) = n$ implying n bits suffice to describe X to n-bit accuracy.
- The differential entropy of a set $X_1, X_2, \cdots, X_n$ of a random variables with density $f(x_1, x_2, \cdots, x_n)$ is defined as

$$h(X_1, X_2, \cdots, X_n) = -\int f(x_1, x_2, \cdots, x_n) \log f(x_1, x_2, \cdots, x_n)\, dx_1 dx_2 \cdots dx_n$$

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

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

Differential Entropy

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

$$\begin{aligned} I(X;Y) &= \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \\ &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \end{aligned}$$

Differential Entropy

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

$$\begin{aligned} I(X;Y) &= \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \\ &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \end{aligned}$$

- The relative entropy $D(f||g)$ between two densities f and g is defined by

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

Differential Entropy

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

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

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

Proof:

- Probability density function of $X_1, \cdots, X_n$ is given by

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

Differential Entropy

- Then we have

$$\begin{aligned} h(f) &= -\int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln\left(\sqrt{2\pi}\right)^n |K|^{1/2} \right] d\mathbf{x} \\ &= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(K^{-1})_{ij}(x_j - \mu_j) \right] + \frac{1}{2}\ln(2\pi)^n |K| \\ &= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(x_j - \mu_j)(K^{-1})_{ij} \right] + \frac{1}{2}\ln(2\pi)^n |K| \\ &= \frac{1}{2} \sum_{i,j} E\left[(x_i - \mu_i)(x_j - \mu_j)\right](K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n |K| \end{aligned}$$

Differential Entropy

$$\begin{aligned} h(f) &= \frac{1}{2}\sum_j \sum_i K_{ji} K_{ij}^{-1} + \frac{1}{2}\ln(2\pi)^n |K| \\ &= \frac{1}{2}\sum_j (KK^{-1})_{jj} + \frac{1}{2}\ln(2\pi)^n |K| \\ &= \frac{1}{2}\sum_j I_{jj} + \frac{1}{2}\ln(2\pi)^n |K| \\ &= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n |K| \\ &= \frac{1}{2}\ln(2\pi e)^n |K| \text{ nats} \\ &= \frac{1}{2}\log(2\pi e)^n |K| \text{ bits} \end{aligned}$$

Mutual Information

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

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

Mutual Information

- Example: 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

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

Mutual Information

- Example: 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.

Mutual Information

- Example: 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.

Properties of differential entropy

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

Properties of differential entropy

- $D(f||g) \geq 0$.
- Proof: Let S be the support set of f, then we have

$$\begin{aligned} -D(f||g) &= \int_S f \log \frac{g}{f} \\ &\leq \log \int_S f \frac{g}{f} \\ &= \log \int_S g = \log 1 = 0 \end{aligned}$$

Properties of differential entropy

- $D(f||g) \geq 0$.
- Proof: Let S be the support set of f, then we have

$$\begin{aligned} -D(f||g) &= \int_S f \log \frac{g}{f} \\ &\leq \log \int_S f \frac{g}{f} \\ &= \log \int_S g = \log 1 = 0 \end{aligned}$$

- Chain rule for differential entropy

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

Chain rule for differential entropy

- Corollary:

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Chain rule for differential entropy

- Corollary:

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

- Proof: Follows directly from chain rule and the fact that $h(X|Y) \leq h(X)$ with equality iff X and Y are independent.

Chain rule for differential entropy

- Corollary:

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)$$

  with equality iff $X_1, X_2, \ldots, X_n$ are independent.
- Proof: Follows directly from chain rule and the fact that $h(X|Y) \leq h(X)$ with equality iff X and Y are independent.
- Application: *(Hadamards inequality:)* If we let $X \sim \mathcal{N}(0, K)$ be a multivariate normal random variable, calculating the entropy in the above inequality gives us

$$|K| \leq \prod_{i=1}^{n} K_{ii}$$

  which is Hadamards inequality

Properties of differential entropy

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

Properties of differential entropy

- $h(X + c) = h(X)$
- Proof: Let $Y = X + c$. Then $f_Y(y) = f_X(y - c)$ and $S_Y = \{x - c : x \in S_X\}$. Letting $x = y - c$, we have

$$\begin{aligned} h(X) &= -\int f_X(x) \log f_X(x)\, dx \\ &= -\int f_X(y-c) \log f_X(y-c)\, dy \\ &= -\int f_Y(y) \log f_Y(y)\, dy \\ &= h(Y) = h(X + c) \end{aligned}$$

Properties of differential entropy

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

Properties of differential entropy

- $h(aX) = h(X) + \log|a|$
- Proof: Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$.

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

Properties of differential entropy

- Let the random vector $X \in R^n$ have zero mean and covariance $K = EXX^t$. Then

$$h(X) \leq \frac{1}{2} \log(2\pi e)^n |K|$$

with equality iff $X \sim N(0, K)$.

Properties of differential entropy

- Let the random vector $X \in R^n$ have zero mean and covariance $K = EXX^t$. Then

$$h(X) \leq \frac{1}{2} \log(2\pi e)^n |K|$$

with equality iff $X \sim N(0, K)$.

- Proof: Let $g(\mathbf{X})$ be any density satisfying $\int g(\mathbf{x})\mathbf{x_i}\mathbf{x_j}\,\mathbf{dx} = \mathbf{K_{ij}}$ for all i, j. Let ϕ_K denotes the density of zero mean, multivariate distribution N(0,K) and $\int x_i x_j \phi_K(\mathbf{x})\,\mathbf{dx} = \mathbf{K_{ij}}$. Then we have

$$\begin{aligned} 0 &\leq D(g||\phi_K) = \int g \log(g/\phi_K) \\ &= -h(g) - \int g \log \phi_K \\ &= -h(g) - \int \phi_K \log \phi_K \\ &= -h(g) + h(\phi_K) \end{aligned}$$