Lecture 11 Differential Entropy

Textbook 8.1-8.6

November 17 and 19, 2024

Outline

1. Definitions
2. Relation of Differential Entropy to Discrete Entropy
3. Joint and conditional differential entropy
4. Relative entropy and mutual information
5. Properties
6. Maximum entropy distributions
7. Examples

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

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

where S is the support of the random variable.

Example

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

$$X = .X_1X_2X_3\cdots ,$$

the dyadic expansion of X, where X_1, X_2, X_3, $\cdots$ is a sequence of pair bits. Then

$$H(X) = H(X_1, X_2, X_3, \cdots) = \sum_{i=1}^{\infty} H(X_i) = \sum_{i=1}^{\infty} 1 = \infty.$$

Uniform distribution

Consider a random variable distributed uniformly from 0 to a so that its density is $1/a$ from 0 to a and 0 elsewhere. Then

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

Normal distribution

Let $X \sim \phi(x) = (1/\sqrt{2\pi\sigma^2})e^{-x^2/2\sigma^2}$. Then calculating the differential entropy in nats, we obtain

$$\begin{aligned} h(\phi) &= -\int \phi \ln \phi \\ &= -\int \phi(x)[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}] \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2}\ln 2\pi e\sigma^2 \text{ nats.} \end{aligned}$$

Changing the base of the logarithm, we have

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

Outline

1. Definitions
2. Relation of Differential Entropy to Discrete Entropy
3. Joint and conditional differential entropy
4. Relative entropy and mutual information
5. Properties
6. Maximum entropy distributions
7. Examples

- Consider a random variable X with density $f(x)$.
- Suppose that we divide the range of X into bins of length Δ.
- Let us assume that the density is continuous within the bins. Then, 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

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

- The entropy of the quantized version is

$$\begin{aligned} H(X^\Delta) &= -\sum_{i=-\infty}^{\infty} p_i \log p_i \\ &= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta, \end{aligned}$$

since $\sum f(x_i)\Delta = \int f(x) = 1$.

- If $f(x)\log f(x)$ is Riemann integrable, the first term approaches the integral of $-f(x)\log f(x)$ as $\Delta \to 0$ by definition of Riemann integrability.

Theorem

If the density $f(x)$ of the random variable X is Riemman integrable, then

$$H(X^{\Delta}) + \log \Delta \to h(f) = h(X),\ as \Delta \to 0.$$

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

Example

1. *If X has a uniform distribution on $[0,1]$ and we let $\Delta = 2^{-n}$, then $h = 0$, $H(x^{\Delta}) = n$, and n bits suffice to describe X to n bit accuracy.*
2. *If X is uniformly distributed on $[0, \frac{1}{8}]$, the first 3 bits to the right of the decimal point must be 0. To describe X to n bit accurracy requires only $n-3$ bits, which agrees with $h(X) = -3$ bits.*
3. *If $X \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 100$, describing X to n bit accuracy would require on the average $n + \frac{1}{2}\log(2\pi e\sigma^2) = n + 5.37$ bits.*

Outline

1. Definitions
2. Relation of Differential Entropy to Discrete Entropy
3. Joint and conditional differential entropy
4. Relative entropy and mutual information
5. Properties
6. Maximum entropy distributions
7. Examples

Definition

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

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

Definition

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

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

Since in general $f(x|y) = f(x,y)/f(y)$, we can also write

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

But we must be careful if any of the differential entropies are infinite.

Entropy of a multivariate normal distribution

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

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

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

Proof.

The joint probability density function is

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

Proof.

Then

$$
\begin{aligned}
h(f) &= -\int f(x)[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}}]d\mathbf{x}. \\
&= \frac{1}{2}E[\sum_{i,j}(X_i-\mu_i)(K^{-1})_{ij}(X_j-\mu_j)] + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}E[\sum_{i,j}(X_i-\mu_i)(X_j-\mu_j)(K^{-1})_{ij}] + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}\sum_{i,j}E[(X_i-\mu_i)(X_j-\mu_j)](K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}\sum_{i,j}K_{ji}(K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}\sum_{j}I_{jj} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}\ln(2\pi e)^n|K| \text{ nats} \\
&= \frac{1}{2}\log(2\pi e)^n|K| \text{ bits.}
\end{aligned}
$$

□

Definition

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

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

Note that $D(f\|g)$ 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 density $f(x,y)$ is defined as

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

From the 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

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

Example

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

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

Then we have that $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$ *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*

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

Outline

1. Definitions
2. Relation of Differential Entropy to Discrete Entropy
3. Joint and conditional differential entropy
4. Relative entropy and mutual information
5. Properties
6. Maximum entropy distributions
7. Examples

Theorem

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

with equality if and only if $f = g$ *almost everywhere.*

Proof.

Let S be the support set of f. Then

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

We have equality if and only if we have equality in Jensen's inequality, which occurs if and only if $f = g$ a.e. □

Corollary

$I(X;Y) \geq 0$ with equality if and only if X and Y are independent.

Corollary

$h(X|Y) \leq h(X)$ with equality if and only if X and Y are independent.

Chain rule for differential entropy

Theorem

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

Corollary

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

with equality if and only if $X_1, X_2, \ldots, X_n$ *are independent.*

Application: Hadamard's 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

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

Theorem

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

Theorem

For $a \neq 0$*,*

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

Proof.

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

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

after a change of variables in the integral. □

Corollary

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

Theorem

Let the random vector $\mathbf{X} \in \mathbb{R}^n$ *have zero mean and covariance* $K = E\mathbf{X}\mathbf{X}^t$. *Then* $h(\mathbf{X}) \leq \frac{1}{2}\log(2\pi e)^n|K|$, *with equality if and only if* $\mathbf{X} \sim \mathcal{N}(0, K)$.

Proof.

Let $g(\mathbf{x})$ be any density satisfying $\int g(\mathbf{x})x_i x_j d\mathbf{x} = K_{ij}$ for all i, j. Let ϕ_K be the density of a $\mathcal{N}(0, K)$ vector, where we set $\mu = 0$. Note that $\log \phi_K(\mathbf{x})$ is a quadratic form and $\int x_i x_j \phi_K(\mathbf{x}) d\mathbf{x} = K_{ij}$. Then

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

where the substitution $\int g \log \phi_K = \int \phi_K \log \phi_K$ follows from the fact that g and ϕ_K yields the same moments of the quadratic form $\log \phi_K(\mathbf{x})$. □

Outline

1. Definitions
2. Relation of Differential Entropy to Discrete Entropy
3. Joint and conditional differential entropy
4. Relative entropy and mutual information
5. Properties
6. Maximum entropy distributions
7. Examples

Consider the following problem: Maximize the entropy $h(f)$ over all probability densities f satisfying

1. $f(x) \geq 0$, with equality outside the support set S
2. $\int_S f(x)dx = 1$
3. $\int_S f(x)r_i(x)dx = \alpha_i$ for $1 \leq i \leq m$.

Thus, f is a density on support set S meeting certain moment constraints α_1, α_2, $\cdots$, α_m.

Approach 1

The differential entropy $h(f)$ is a concave function over a convex set. We form the functional

$$J(f) = -\int f \ln f + \lambda_0 \int f + \sum_{i=1}^{m} \lambda_i \int f r_i$$

and "differentiate" with respect to $f(x)$, the xth component of f, to obtain

$$\frac{\partial J}{\partial f(x)} = -\ln f(x) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x).$$

Setting this equal to zero, we obtain the form of maximizing density

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}, \ x \in S,$$

where $\lambda_0, \lambda_1, \cdots, \lambda_m$ are chosen so that f satisfies the constraints.

The approach using calculus only suggests the form of the density that maximizes the entropy. To prove that this is indeed the maximum, we can take the second variation. It is simpler to use the information inequality $D(g\|f) \geq 0$.

Approach 2

If g satisfies the conditions and f^* is of the form

$$f^*(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}, \ x \in S,$$

then

$$0 \leq D(g \| f^*) = -h(g) + h(f^*).$$

Thus $h(g) \leq h(f^*)$ for all g satisfying the constraints. We prove this in the following theorem.

Theorem

Let $f^* = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)}$, $x \in S$, *where* $\lambda_0, \cdots, \lambda_m$ *are chosen so that* f^* *satisfies*

1. $f(x) \geq 0$, *with equality outside the support set* S
2. $\int_S f(x)dx = 1$
3. $\int_S f(x)r_i(x)dx = \alpha_i$ *for* $1 \leq i \leq m$.

Then f^* *uniquely maximizes* $h(f)$ *over all probability densities* f *satisfying the above constraints.*

Proof.

Let g satisfy the constraint. Then

$$\begin{aligned}
h(g) &= -\int_S g \ln g \\
&= -\int_S g \ln(\frac{g}{f^*} f^*) \\
&= -D(g\|f^*) - \int_S g \ln f^* \\
&\le -\int_S g \ln f^* \\
&= -\int_S g(\lambda_0 + \sum \lambda_i r_i) \\
&= -\int_S f^*(\lambda_0 + \sum \lambda_i r_i) \\
&= -\int_S f^* \ln f^* \\
&= h(f^*).
\end{aligned}$$

Note the equality holds if and only if $g(x) = f^*(x)$ for all x, except for a set of measure 0, thus proving uniqueness.

Outline

1. Definitions
2. Relation of Differential Entropy to Discrete Entropy
3. Joint and conditional differential entropy
4. Relative entropy and mutual information
5. Properties
6. Maximum entropy distributions
7. Examples

Example

Let the constraints be $EX = 0$ *and* $EX^2 = \sigma^2$. *Then the form of the maximizing distribution is*

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}.$$

To find the appropriate constants, we first recognize that this distribution. Hence, the density that satisfies the constraints and also maximizes the entropy is the $\mathcal{N}(0, \sigma^2)$ *distribution:*

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

Example

Let $S = [a, b]$, with no other constraints. Then the maximum entropy distribution is the uniform distribution over this range.

Example

$S = [0, \infty)$ *and* $EX = \mu$. *Then the entropy-maximizing distribution is*

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \; x \geq 0.$$

Example

$S = (-\infty, \infty)$, and $EX = \mu$. Here the maximum entropy is infinite, and there is no maximum entropy distribution. (Consider normal distributions with larger variances.)

Example

$S = (-\infty, \infty)$, $EX = \alpha_1$, and $EX^2 = \alpha_2$. The maximum entropy distribution is $\mathcal{N}(\alpha_1, \alpha_2 - \alpha_1^2)$.

Example

Find the word lengths of the optimal binary encoding of $p = (\frac{1}{100}, \frac{1}{100}, \cdots, \frac{1}{100})$.

Example

The Z-channel has binary input and output alphabets and translation probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}, \; x, y \in \{0, 1\}.$$

Find the capacity of the Z-channel and the maximizing input probability distribution.