EE 231A Information Theory
Lecture 10
Differential Entropy

A. Introduction and uniform example
B. Conditional and joint differential entropy, continuous versions of relative entropy, mutual information
C. Normal and multivariate normal examples
D. Offset and scaling properties of differential entropy
E. Differential entropy as a limit of discrete entropy.

Part 10 A: Introduction and uniform example

$H = \infty$ for continuous RV's

- For discrete random variables:

$$H(X) = -\sum p(x) \log p(x)$$

- What if X is continuous?

- Then the above equation doesn't parse, but any continuous random variable will take an infinite number of bits to describe, so H should be infinite for such RVs.

Mutual information still makes sense.

- Mutual information still makes sense for continuous random variables.

- In fact, in many important communication channels we will find that $I(X;Y)$ can be finite even though $H(X)$ and $H(Y)$ are infinite.

- We would still like to use $I(X;Y) = H(Y) - H(Y \mid X)$.

- But there is that pesky problem that $H(Y) = \infty$ and $H(X) = \infty$.

Differential Entropy

- Surprisingly enough, there is a way to strip off a constant amount of ∞ so that the mutual information is computed correctly by a difference similar to $I(X;Y) = H(Y) - H(Y \mid X)$.
- For continuous random variables we use Differential Entropy:

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

- S is the support set where $f(x) > 0$. $f(x)$ is the probability density function (pdf) for X.

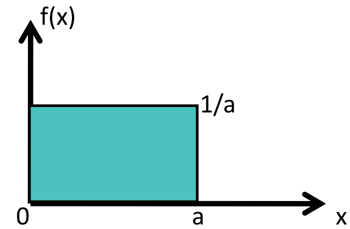
Entropy and Differential Entropy

$$H(X) = -E \log p(x)$$

$$h(X) = -E \log f(x)$$

Example: Uniform Distribution

- Uniform distribution



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

1. scaling changes h. (unlike H)
2. h can be negative. (unlike H)

Part 10 B:

Conditional and joint differential entropy, continuous versions of relative entropy, mutual information

Conditional and Joint h

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

$$\begin{aligned} h(X,Y) &= -\int f(x,y) \log f(x,y)\, dx\, dy \\ &= -\int f(x,y) \log\Big(f\big(y\big) f\big(x \mid y\big)\Big) dx\, dy \\ &= h(Y) + h(X \mid Y) \end{aligned}$$

Relative entropy

- Relative Entropy: $D(f \,\|\, g) = \int f(x) \log \frac{f(x)}{g(x)} dx$
  - $D(f \,\|\, g)$ is finite only when the support of f is contained in the support of g. That is $g(x)>0$ whenever $f(x)>0$.

Relative Entropy is Positive

- Jensen $E[f(x)] \geq f(EX)$ for $f(x)$ convex

$$D(f \,\|\, g) = \int_S f(x)\left(-\log \frac{g(x)}{f(x)}\right) dx \qquad \boxed{t = \frac{g(x)}{f(x)}}$$

$$= E_t[-\log t]$$

$$\geq -\log E_t t$$

$$= -\log \int_S f(x) \frac{g(x)}{f(x)} dx$$

$$= -\log 1$$

$$= 0$$

Mutual Information

- $I(X;Y) = D\big(f(x,y) \,\|\, f(x)f(y)\big)$
- $I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \, dx\, dy$
- $I(X;Y) = h(X) - h(X \mid Y)$
  $= h(Y) - h(Y \mid X)$

Some inequalities

- $D(f \| g) \geq 0$
  - by Jensen
- $I(X;Y) \geq 0$
  - Since $I=D$
- $h(X|Y) \leq h(X)$
  - Since $I\left(X;Y\right) = h(X) - h(X|Y)$

Chain Rule and a related inequality

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

Part 10C:
Normal and Multivariate Normal Examples

Example: Normal (Gaussian) Distribution

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

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

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

$$= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log e \cancelto{1}{\frac{E\left[X^2\right]}{\sigma^2}}$$

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

Joint differential entropy

- $h(X_1, X_2, ..., X_n) = -\int f(\bar{x}) \log f(\bar{x})\, dx_1\, dx_2 ... dx_n$

Example: Multivariate Normal

$$E\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \mu \quad E\left[\begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_n-\mu_n \end{bmatrix}\begin{bmatrix} x_1-\mu_1 & \cdots & x_n-\mu_n \end{bmatrix}\right] = K$$

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

$$h(f) = \frac{1}{2}\log(2\pi)^n |K| + \frac{1}{2}\log e E[(X-\mu)^T K^{-1}(X-\mu)]$$

$$h(f) = \frac{1}{2}\log(2\pi)^n |K| + \frac{1}{2}\log e E[\underbrace{\mathrm{trace}((X-\mu)^T K^{-1}(X-\mu))}_{\text{a scalar}}]$$

Multivariate Normal Differential Entropy

$$\boxed{\text{trace } AB = \text{trace } BA}$$

$$\frac{1}{2}\log eE\left[\text{trace}\left(\underbrace{(X-\mu)^T}_{A}\underbrace{K^{-1}(X-\mu)}_{B}\right)\right]$$

$$= \frac{1}{2}\log eE\left[\text{trace}\left(K^{-1}(x-\mu)(x-\mu)^T\right)\right]$$

$$= \frac{1}{2}\log e\ \text{trace}\left(K^{-1}E\left[(x-\mu)(x-\mu)^T\right]\right)$$

$$= \frac{1}{2}\log e\ \text{trace}(K^{-1}K)$$

$$= \frac{1}{2}\log e\ \text{trace}(I)$$

$$= \frac{n}{2}\log e$$

Multivariate Normal Conclusion

$$h(f) = \frac{1}{2}\log(2\pi)^n\left|K\right| + \frac{n}{2}\log e$$

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

Part 10 D:
The offset and scaling properties of differential entropy.

Offset and Scaling Properties

- $h(X+c)=h(X)$ where c is a constant
- $h(aX)=h(X)+\log|a|$
  - Proof: We consider $Y=aX$ for $a<0,\ a>0$.

$$Y = aX \qquad a>0$$

$$F_Y(y) = P(Y \le y)$$
$$= P(aX \le y)$$
$$= P\left(X \le \frac{y}{a}\right)$$
$$= F_X\left(\frac{y}{a}\right)$$

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$
$$= \frac{d}{dy} F_X\left(\frac{y}{a}\right)$$
$$= \frac{dF_X\left(y/a\right)}{dx} \frac{d\left(y/a\right)}{dy}$$
$$= f_X\left(y/a\right)\frac{1}{a}$$

$$y = ax$$
$$dy = adx$$

$$h\left(y\right) = -\int f\left(y\right)\log\left(f\left(y\right)\right)dy$$
$$= -\int f\left(y\right)\log\left(\frac{1}{a} f_X\left(y/a\right)\right)dy$$
$$= -\int f\left(y\right)\log\left(\frac{1}{a}\right)dy - \int f\left(y\right)\log\left(f_X\left(y/a\right)\right)dy$$
$$= \log a - \int f\left(y\right)\log\left(f_X\left(y/a\right)\right)dy$$
$$= \log a - \int \frac{1}{a} f_X\left(y/a\right)\log\left(f_X\left(y/a\right)\right)dy$$
$$= \log a - \int f_X\left(y/a\right)\log\left(f_X\left(y/a\right)\right)\frac{dy}{a}$$
$$= \log a - \int f_X\left(x\right)\log\left(f_X\left(x\right)\right)dx \quad \boxed{= \log\left|a\right| + h\left(x\right)}$$

$$Y = aX \qquad a<0$$

$$F_Y(y) = P(Y \le y)$$
$$= P(aX \le y)$$
$$= 1 - P\left(X \le \frac{y}{a}\right)$$
$$= 1 - F_X\left(\frac{y}{a}\right)$$

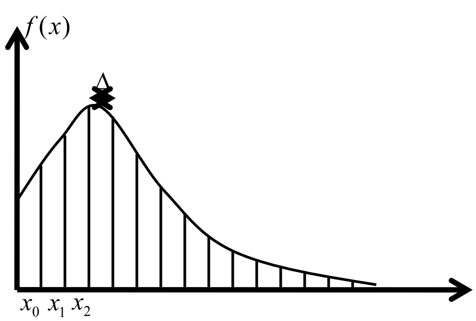
$$f_Y(y) = \frac{d}{dy} F_Y(y)$$
$$= \frac{d}{dy}\left(1 - F_X\left(\frac{y}{a}\right)\right)$$
$$= -\frac{dF_X(y/a)}{dx}\frac{d(y/a)}{dy}$$
$$= -f_X(y/a)\frac{1}{a}$$

$$y = ax$$
$$dy = adx$$

$$h(y) = -\int_{y=-\infty}^{\infty} f(y)\log\big(f(y)\big)dy$$
$$= -\int_{y=-\infty}^{\infty} f(y)\log\left(\frac{-1}{a} f_X(y/a)\right)dy$$
$$= \log(-a) - \int_{y=-\infty}^{\infty} f(y)\log\big(f_X(y/a)\big)dy$$
$$= \log(-a) - \int_{y=-\infty}^{\infty} -\frac{1}{a} f_X(y/a)\log\big(f_X(y/a)\big)dy$$
$$= \log(-a) - \int_{y=-\infty}^{\infty} -f_X(y/a)\log\big(f_X(y/a)\big)\frac{dy}{a}$$
$$= \log(-a) - \int_{x=\infty}^{-\infty} -f_X(x)\log\big(f_X(x)\big)dx \quad \boxed{= \log|a| + h(x)}$$

Part 10 E:
Differential entropy as a limit of discrete entropy

Differential entropy as a limit of discrete entropy



- Choose x_i so that $f(x_i)\Delta$ = area of that chunk.
- Define $x^{\Delta} = x_i \quad \text{if} \quad i\Delta \le x \le (i+1)\Delta$

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

Limit of discrete entropy is h minus log Δ

$$H(x^{\Delta}) = -\sum_{-\infty}^{\infty} f(x_i)\Delta\log(f(x_i)\Delta)$$

$$= \underbrace{-\sum_{-\infty}^{\infty} f(x_i)\Delta\log f(x_i)}_{\to\ -\int_{-\infty}^{\infty} f(x)\log f(x)dx} - \underbrace{\sum_{-\infty}^{\infty} f(x_i)\Delta\log\Delta}_{\log\Delta}$$

$$H(x^{\Delta}) + \log\Delta \to h(X)$$

- For large n, $H(x^{\Delta}) \approx h(X) - \log\Delta$

Limit of discrete mutual information

$$I(X^{\Delta};Y^{\Delta}) = H\left(Y^{\Delta}\right) - H\left(Y^{\Delta} \mid X^{\Delta}\right)$$

$$\approx h\left(Y\right) - \log\Delta - h\left(Y \mid X\right) + \log\Delta$$

$$= h\left(Y\right) - h\left(Y \mid X\right)$$

PMFs, PDFs, and Mass Points

- A probability mass function is comprised entirely of mass points. That is, individual values that have probability.
- A probability density function has no mass points. No individual point has positive probability. To have positive probability you have to integrate the density over a region.
- What happens when you have a random variable that is a mixture of density and mass points?

The differential entropy *h(x)* is -∞ whenever there is a mass point.

- To find the contribution of the mass point to *h(x)*, take a limit of a rectangular pdf as width goes to zero and height goes to infinity.
- Consider a mass point at zero with probability of ½....

$$
\begin{aligned}
\lim_{a\to 0}\left(-\int_{-a/2}^{a/2}\frac{1}{2a}\log\frac{1}{2a}df\right) &= \lim_{a\to 0}\left(-\frac{a}{2a}\log\frac{1}{2a}\right)\\
&= \lim_{a\to 0}\left(-\frac{1}{2}\log\frac{1}{2a}\right)\\
&= \lim_{a\to 0}\left(\frac{1}{2}+\log a\right)\\
&= \frac{1}{2}+\lim_{a\to 0}\log a\\
&= -\infty
\end{aligned}
$$