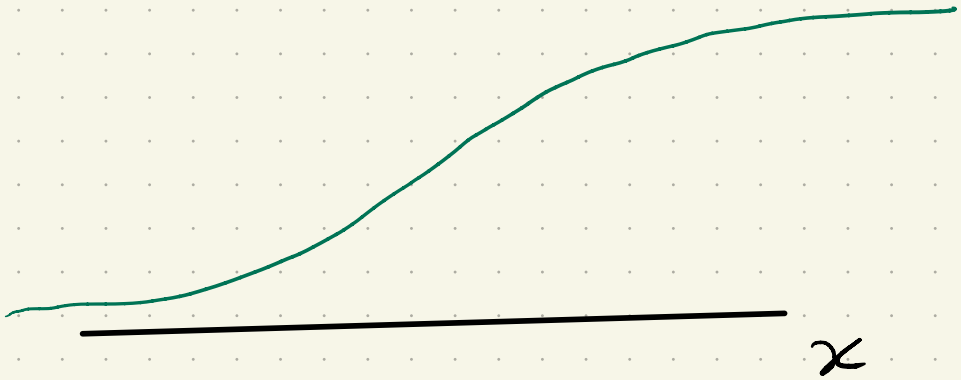
Introducción a la Teoría de la Información

Entropía diferencial.

Facultad de Ingeniería, UdelaR

Año ~~2020~~ 2024

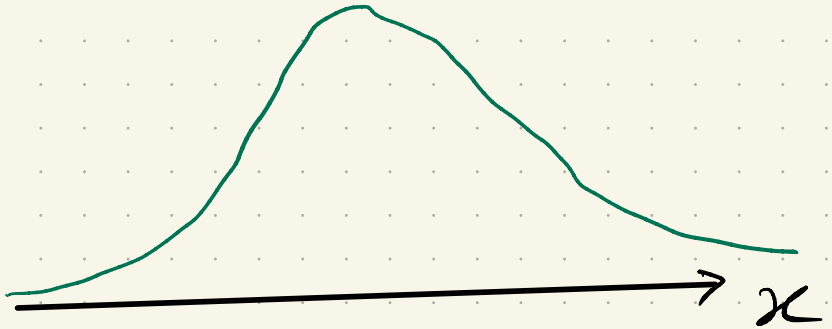
Distribución



$F(x) = P(X \leq x)$

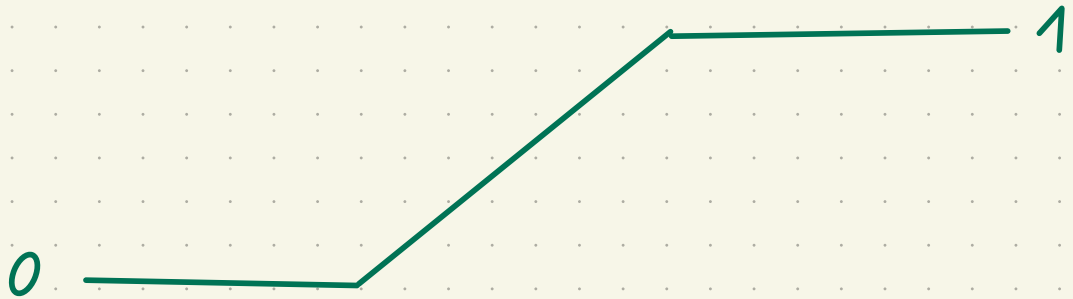
$F(x)$ creciente

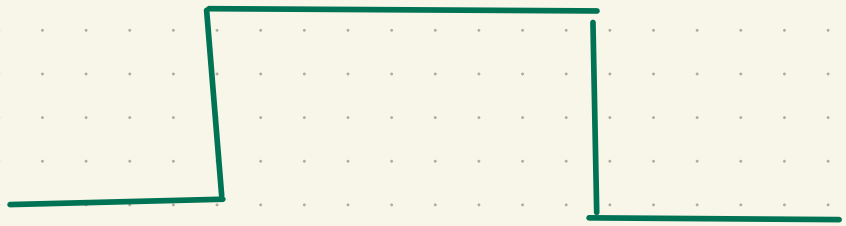
Densidad



$f(x) = F'(x)$

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$





Ej uniforme

Definición

$\sum p_i \log p_i$

Definición (Entropía diferencial)

Para una variable aleatoria X con densidad de probabilidad $f(x)$ de soporte S, se define como:

$$h(X) = -\int_S f(x) \log f(x)\, dx \qquad E_f[-\log f(x)]$$

- Para variables aleatorias continuas
- Similar a la entropía discreta
- Algunas diferencias importantes, hay que usarla con cuidado.
- Puede *no existir* para ciertas $f(x)$.

Ejemplos

Ejemplo (Distribución uniforme en $(0, a)$)

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

- Para $a < 1$, $h(X) = \log a$ es *negativa*

Ejemplos

una variable

Ejemplo (Distribución Normal $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$)

$$h(X) = -\int \phi(x) \log \phi(x) = \frac{1}{2} \log(2\pi e \sigma^2) \text{bits}$$

media $\mu = 0$

varianza σ

$\mu = E_f(X)$

$\sigma^2 = E_f[(X-\mu)^2]$

$\int x f(x) dx$

mejor usar ln.

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

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

$\log a^n = n \log a$

$E[x^2] = \sigma^2$

$$-\int_{-\infty}^{+\infty} f(x) \ln f(x)\, dx = \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} =$$

$\frac{1}{2} \ln e$

$$= \frac{1}{2} \ln(2\pi e \sigma^2) \text{ nats}$$

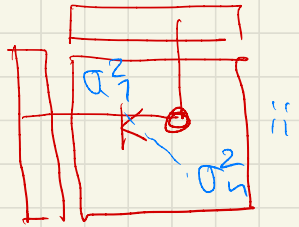
$$= \frac{1}{2} \log(2\pi e \sigma^2) \text{ bits.}$$

(cambio de base)

Multidimensional $$f(x) = \frac{1}{\sqrt{2\pi}\,|k|}\, e^{-x^t k^{-1} x}$$

Covariance $k = E_f[x x^t]$ $\quad k_{ij} = E[x_i x_j]$

$$h(\vec{x}) = \frac{1}{2} \log\left((2\pi e)^n |k|\right)$$



$\vec{X} = (X_1, X_2 \cdots X_n)$ gaussianas

$E_f(\vec{x}) = \vec{\mu}$

AEP

si son i.i.d, $f(x_1, x_2 \cdots x_n) = f(x_1) f(x_2) \cdots f(x_n)$

Teorema (AEP)

$X_1 \ldots X_n$ variables *i.i.d.*, $X_i \sim f(x)$

$$-\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \to E\left[-\log f(X)\right] = h(X)$$

Definición (Conjunto típico)

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

$-\frac{1}{n} \log f(x_1, x_2 \cdots x_n) = -\frac{1}{n} \sum_i \log f(x_i) \to E_f[-\log f(x)] =$

$= h(x)$

Propiedades de $A_\epsilon^{(n)}$

Propiedades

- $P\left(A_\epsilon^{(n)}\right) > 1-\epsilon$ para n suf. grande
- $\text{Vol}\left(A_\epsilon^{(n)}\right) \leq 2^{n(h(X)+\epsilon)}$ para **todo** n
- $\text{Vol}\left(A_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ para n suf. grande

Las pruebas son análogas a las demostradas para el conjunto típico de una variable aleatoria discreta.

$$h(x) - \varepsilon \leq -\frac{1}{n} \log f(x_1 \dots x_n) \leq h(x) + \varepsilon$$

$$-h(x) + \varepsilon \geq \frac{1}{n} \log f(x_1 \dots x_n) \geq -h(x) - \varepsilon$$

$$2^{-n(h(x)} \geq f(x_1 \dots x_n) \geq 2^{-n(h(x)+\varepsilon)}$$

V.A. discreta – cardinal V.A continua – Volumen

Def: $\text{Vol } A = \iiint_A dx_1 dx_2 \dots dx_n$

$$1 = \iint_S f(x_1 \dots x_n)\, dx_1 \dots dx_n \geq \iint_{A_\varepsilon^{(n)}} f(x_1 \dots x_n)\, dx_1 \dots dx_n \geq$$

$$\iint_{A_\varepsilon^{(n)}} 2^{-n(h(x)+\varepsilon)}\, dx \dots dx_n = 2^{-n[h(x)+\varepsilon]} \text{Vol } A_\varepsilon^{(n)}$$

$$\text{Vol } A_\varepsilon^{(n)} \leq 2^{n[h(x)+\varepsilon]}$$

Algunas observaciones

?

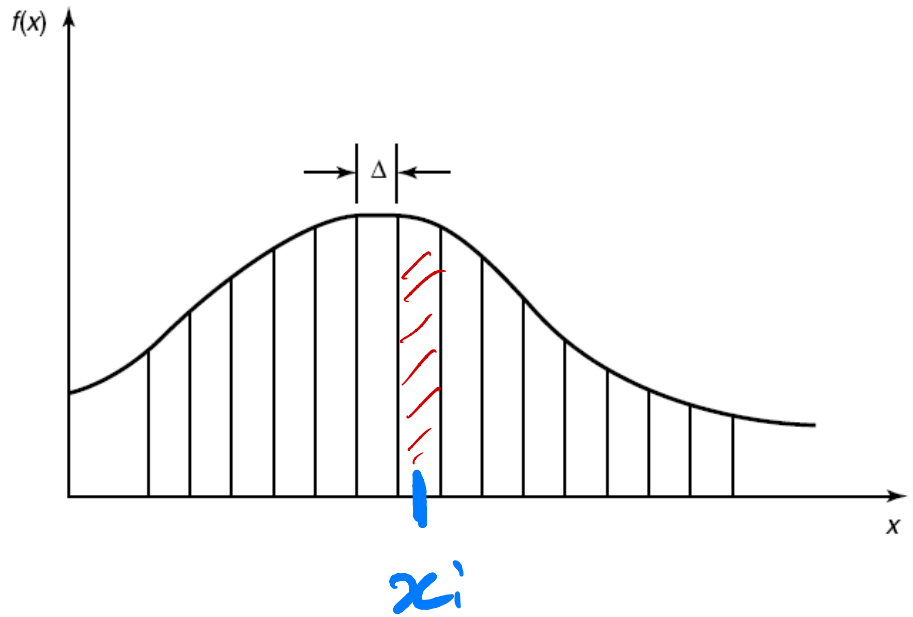
- El volumen del conjunto típico es aproximadamente 2^{nh}, o sea el volumen de un cubo n-dimensional de lado 2^h.
- Esto da una interpretación de $h(X)$ como el logaritmo del lado de un cubo, que tiene el mismo volumen que un subconjunto del soporte de X que concentra la mayoría de la probabilidad.

Entropía diferencial y Entropía discreta (1)

- Dividimos el soporte S de $X \sim f(x)$ en intervalos de ancho Δ
- Para f continua, por Teorema del Valor Medio podemos tomar en cada intervalo un punto x_i tal que

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

- Definimos $X^{\Delta} = x_i \Leftrightarrow i\Delta \leq X < (i+1)\Delta$



Entropía diferencial y Entropía discreta (2)

- $\Pr\left\{X^{\Delta} = x_i\right\} = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \Delta f(x_i)$
- X^{Δ} es una variable aleatoria discreta con entropía dada por:

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

Entropía diferencial y Entropía discreta (3)

Teorema ($h(X)$ vs $H(X^\Delta)$)

Si $f(x) \log f(x)$ es integrable Riemann,

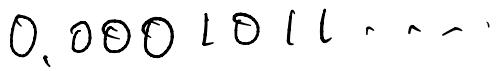
$$H(X^\Delta) + \log \Delta \to h(f) = h(X) \text{ cuando } \Delta \to 0$$

- Si tomamos $\Delta = 2^{-n}$, vemos que $h(X) + n$ es aproximadamente la cantidad de bits promedio necesarios parar codificar X con n bits de precisión.

Entropía diferencial y Entropía discreta (4)

Ejemplo ($X \sim U(0,a)$, $h(X) = \log a$)

- Si $X \sim U(0,1)$, $h(X) = 0$ entonces $h(X) + n = n$ coincide con la cantidad necesaria de bits para describir X con n bits de precisión.
- $X \sim U(0, \frac{1}{8})$ Representando X en binario, los 3 bits a la derecha del punto quedan en 0, o sea que describir X con n bits de precisión requiere sólo $n-3$ bits, que coincide con $h(X) + n = \log 1/8 + n$.



Entropía diferencial conjunta y condicional

Definición (Entropía conjunta)

$$h(x,y) = -\int_S f(x,y)\log f(x,y)dxdy$$

soporte de las dos

Definición (Entropía condicional)

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

Regla de la cadena

De forma similar al caso discreto, obtenemos

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

$H(Y) + h(X|Y)$

Entropía relativa e Información mutua

Definición (Distancia de Kullback-Leibler)

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

Notar que D es finita sólo si el soporte de f está contenido en el de g.

Definición (Información mutua)

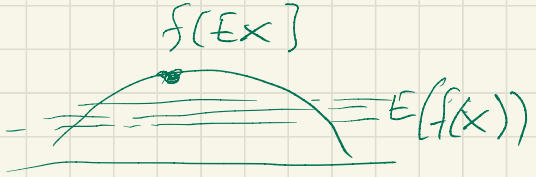
$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy$$

Por definición se obtiene, al igual que para el caso discreto, que

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

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

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



$$-D(f\|g) = \int f \log \frac{g}{f} = E_f \log \frac{g}{f} \leq \text{jensen}$$

↑ concava

$$\leq \log E_f\left(\frac{g}{f}\right) = \log \int f \frac{g}{f} = \log \int g = \log 1 = 0$$

La divergencia o distancia KL es positiva

Definición general de información mutua

Definición (Información mutua)

Sean $\mathcal{X}$, $\mathcal{Y}$ los soportes de X, Y, respectivamente.

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

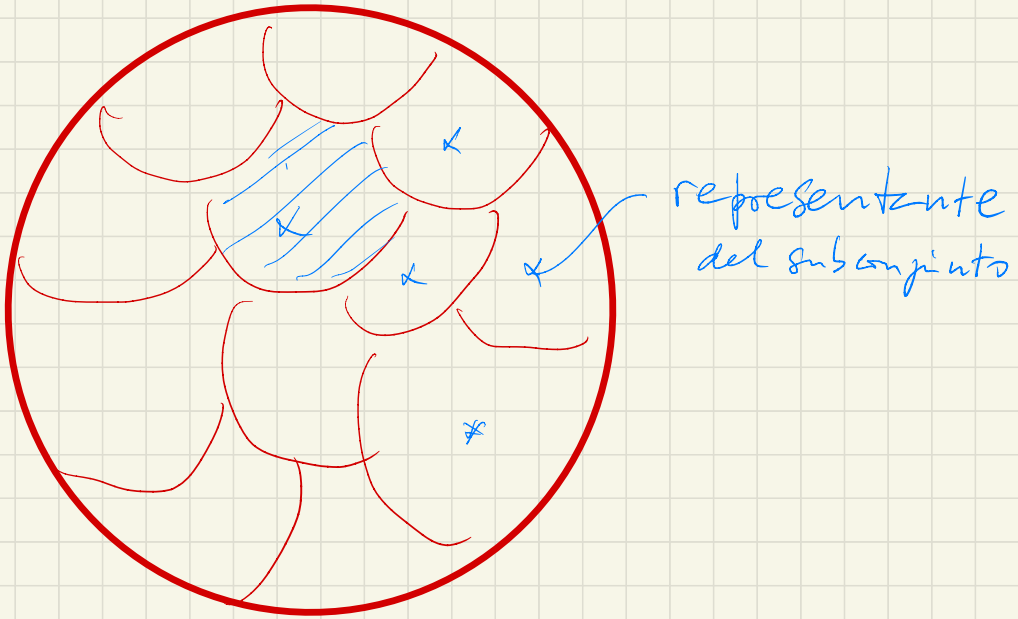
Donde $\mathcal{P} = \{P_i\}_{i=1\ldots N_p}$ y $\mathcal{Q} = \{Q_i\}_{i=1\ldots N_q}$ son particiones finitas de $\mathcal{X}$, $\mathcal{Y}$, respectivamente, y se cumple

$$\Pr\{[X]_{\mathcal{P}} = i\} = \Pr\{X \in P_i\}\ ; \quad \Pr\{[Y]_{\mathcal{Q}} = i\} = \Pr\{Y \in Q_i\}\ .$$

- Esta definición aplica a variables discretas, continuas, y mezclas de uno y otro tipo.
- Si $\mathcal{P}'$ refina $\mathcal{P}$, $[X]_{\mathcal{P}}$ es función de $[X]_{\mathcal{P}'}$, por lo tanto podemos escribir $[Y]_{\mathcal{Q}} \to [X]_{\mathcal{P}'} \to [X]_{\mathcal{P}}$ y por la desigualdad de procesamiento de datos

$$I([X]_{\mathcal{P}'};[Y]_{\mathcal{Q}}) \geq I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$$

$\mathcal{F}'$ refina $\mathcal{F}$ si $\mathcal{F} \subset \mathcal{F}'$



Definición general de información mutua

- Cuando X, Y son discretas de alfabeto finito, para particiones $\mathcal{P}$ y $\mathcal{Q}$ suficientemente finas se cumple que

$$\begin{aligned} H([X]_{\mathcal{P}}) &= H(X) \\ H([Y]_{\mathcal{Q}}) &= H(Y) \\ H([X]_{\mathcal{P}}\ ,\ [Y]_{\mathcal{Q}}) &= H(X,Y) \end{aligned}$$

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

Definición general de información mutua

Variable y su discretización

- Cuando X continua, Y discreta de alfabeto finito, consideremos particiones $\mathcal{P}$ y $\mathcal{Q}$ a intervalos regulares de ancho Δ. Entonces,

$$
\begin{aligned}
I(X^\Delta;Y^\Delta) &= H(X^\Delta) + \log\Delta - (H(X^\Delta|Y^\Delta) + \log\Delta) \\
&= H(X^\Delta) + \log\Delta - \\
&\quad - \sum p(Y^\Delta = y_i)\Big(H(X^\Delta|Y^\Delta = y_i) + \log\Delta\Big) \\
&\quad \text{Para } \Delta \text{ suficientemente chico,} \\
&= H(X^\Delta) + \log\Delta - \\
&\quad - \sum p(Y = y_i)\Big(H(X^\Delta|Y = y_i) + \log\Delta\Big) \\
\to\ & h(X) - \sum p(Y = y_i)\Big(h(X|Y = y_i)\Big) \\
&= h(X) - h(X|Y)
\end{aligned}
$$

y_i

ruido de cuantificación

Definición general de información mutua

X e Y continuas

- Cuando X, Y continuas, también puede verse que $I(X^{\Delta}; Y^{\Delta}) \to h(X) + h(Y) - h(X, Y) = I(X; Y)$

Propiedades

1. $D(f||g) \geq 0$ con igualdad sii $f = g$ en casi todo punto
2. $I(X;Y) \geq 0$ con igualdad sii X, Y independientes
3. $I(X;Y) = h(X) - h(X|Y)$
4. $h(X) \geq h(X|Y)$ con igualdad sii X, Y independientes
5. $h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, \ldots, X_{i-1})$
6. $h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)$
7. Si $X \to Y \to Z$, $I(X;Y) \geq I(X;Z)$
8. $h(X + c) = h(X)$
9. $h(aX) = \log|a| + h(X)$
10. Para un vector $\mathbf{X}$, $h(A\mathbf{X}) = \log|\det A| + h(\mathbf{X})$

La distribución normal maximiza h para varianza dada

Teorema

Sea $\mathbf{X} \in \mathbb{R}^n$ con media cero y covarianza $\mathbf{K} = E\left[\mathbf{X}\mathbf{X}^t\right]$. Entonces $h(\mathbf{X}) \leq \frac{1}{2}\log\left[(2\pi e)^n |\mathbf{K}|\right]$ con igualdad sii $\mathbf{X}$ tiene distribución normal.

Demostración.

Sea g densidad de probabilidad con media cero y covarianza $\mathbf{K}$.
Sea $\phi(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{|\mathbf{K}|}} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{K}^{-1}\mathbf{x}}$

$$\begin{aligned} D(g||\phi) &= \int g \log g/\phi = -h(g) - \int g \log \phi \\ &= -h(g) - E_g\left[\log \phi(\mathbf{X})\right] \\ &= -h(g) - E_\phi\left[\log \phi(\mathbf{X})\right] \text{ porque } E_\phi\left[X_i X_j\right] = E_g\left[X_i X_j\right] \\ &= -h(g) + h(\phi) \end{aligned}$$

La distribución normal maximiza h para varianza dada

$E X^2 = P$

Teorema

$K = \sigma^2$

Sea $\mathbf{X} \in \mathbb{R}^n$ con media cero y covarianza $\mathbf{K} = E\left[\mathbf{X}\mathbf{X}^t\right]$. Entonces $h(\mathbf{X}) \leq \frac{1}{2}\log\left[(2\pi e)^n |\mathbf{K}|\right]$ con igualdad sii $\mathbf{X}$ tiene distribución normal.

Demostración.

Sea g densidad de probabilidad con media cero y covarianza $\mathbf{K}$.

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

$\int g \log g \, dx = E_g(\log g) = -h(g)$

$$
\begin{aligned}
D(g||\phi) &= \int g \log g/\phi = -h(g) - \int g \log \phi \\
&= -h(g) - E_g\left[\log \phi(\mathbf{X})\right] \\
&= -h(g) - E_\phi\left[\log \phi(\mathbf{X})\right] \text{ porque } E_\phi\left[X_i X_j\right] = E_g\left[X_i X_j\right] \\
&= -h(g) + h(\phi) \geq 0
\end{aligned}
$$

$\log \phi$ contiene $K^{-1}_{ij} X_i X_j$ } ctes

$\log e^A = \log e \ln e^A = \underline{\text{cte}} \cdot A$

Las páginas que siguen desarrollan y explican la deducción de la entropía de la normal y la demostración de que la normal maximiza la entropía para una varianza dada.

Varianza y entropía

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

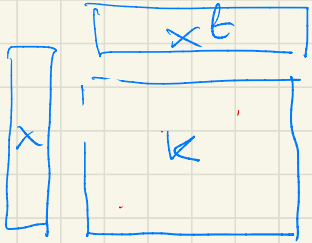
$X \sim N(0,\sigma)$ unidimensional

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

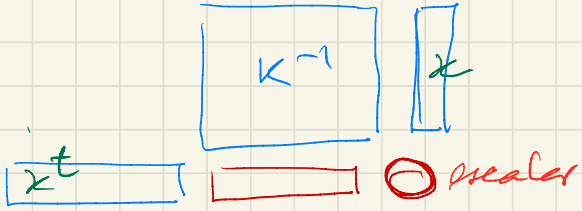
$\bar{X} \sim N_n(\bar{0}, K)$ multidimensional

covarianza $K = E[xx^t]$

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

$K = E$ 

$K_{ij} = E[x_i x_j]$



$$h(\bar{x}) = \frac{1}{2} \log\left[(2\pi e)^n |K|\right]$$

$$K^{-1}_{ij} x_i x_j$$

Nota: K es simétrica

1 dim $\rightarrow$ x $\quad f(x) \quad E_f(x) \quad E_f(x^2) \quad \sigma^2 = Ex^2 - (Ex)^2$

n dim $\vec{x}$ $\quad f(x_1 \dots x_n) \quad E(x) = \vec{\mu} \quad K = E_f[x\ x^t]$

$\rightarrow \phi(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |k|}}\, e^{-\frac{1}{2}(x^t k^{-1} x)}$ cov. K

$E_\phi[x_i x_j] = k_{ij} = E_f[x_i x_j]$ Matriz covarianza

Mayorante

Para aligerar, suponemos spq que restamos la media

media $\bar{\mu} = E[\bar{x}]$

$$\phi(\bar{x}) = \frac{1}{\sqrt{(2\pi)^n |k|}} e^{-\frac{1}{2} x^t k^{-1} x}$$

$$h(\phi) = \int_{-\infty}^{+\infty} \phi(x) \left[\frac{1}{2} x^t k^{-1} x - \frac{1}{2} \ln\left((2\pi)^n |k|\right) \right] dx$$

(A) $= \frac{1}{2} \ln (2\pi)^n |k|$

$$(A) = \frac{1}{2} E\left[\sum_{i,j} x_i \left(k^{-1}\right)_{ij} x_j \right] =$$

$$= \frac{1}{2} E\left[\sum_{i,j} x_j x_i \left(k^{-1}\right)_{ij} \right] = \frac{1}{2} \sum_{ij} E(x_j x_i) \left(k^{-1}\right)_{ij} =$$

$$= \frac{1}{2} \sum_{ij} K_{ji} \left(k^{-1}\right)_{ij} = \frac{1}{2} \sum_{j} \left(k k^{-1}\right)_{jj} = \frac{1}{2} n = \frac{1}{2} n \ln e$$

↑ traza de I

$$h(\phi) = \frac{1}{2} \ln\left\{ (2\pi e)^n |k| \right\} \text{ nats}$$

Producto de matrices

En general, $C = A \times B$ $\qquad C_{ij} = \sum_k A_{ik} B_{kj}$

este caso $\sum_{ij} K_{ji} \left[K^{-1}\right]_{ij} = \sum_j \sum_i K_{ji} (K^{-1})_{ij} =$

$= \sum_j \left(K K^{-1}\right)_{jj} = \sum_j I_{jj} = n$ (traza)

Las que siguen son edificios auxiliares o

agrícolas.

$X \sim f(x)$ $\quad \sigma^2 = E_f[x^2]$ $\quad H(f) \leq H(\phi)$

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

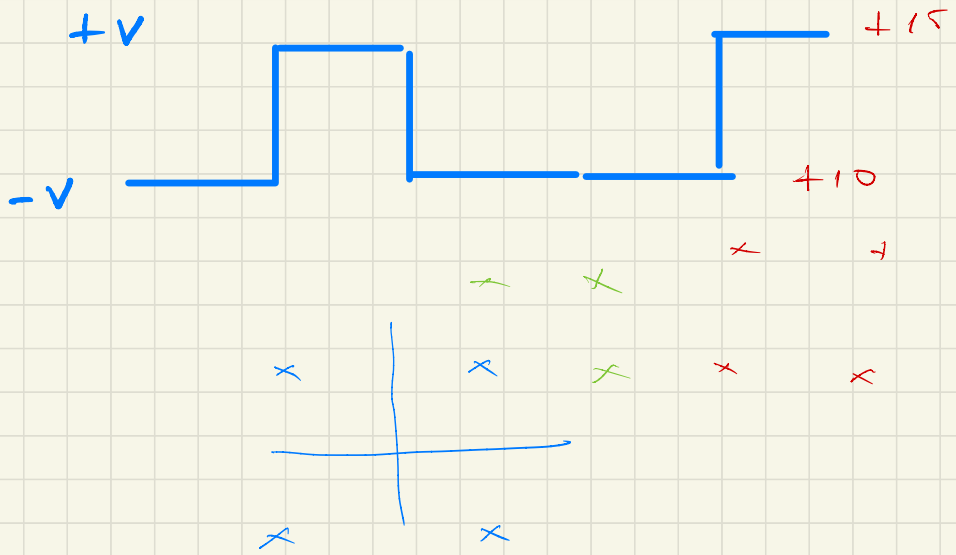
$E_\phi[x^2] = \sigma^2$

$D(f\|\phi) \geq 0$ $\quad D(f\|\phi) = \int f \log\left(\frac{f}{\phi}\right) =$

$$= \underbrace{\int f \log(f)}_{-H(f)} - \underbrace{\int f \log \phi}_{E_f(\log\phi)} = -H(f) + H(\phi) \geq 0$$

$E_\phi(\log\phi)$

$-H(\phi)$


+V
-V
+15
+10

Uniforme dimensión 2

X, Y v.a. $\quad X \sim U[0,a] \quad Y \sim [0,b]$

? Soporte (X, Y)

? Se desea cuantificar X con n bits precisión
$\qquad\qquad\qquad\qquad\quad Y \qquad m$

1) Dibujá soporte y su cuantificación

2) Expresión para las v.a. $X^\Delta \quad Y^\Delta$

3) Relacionar $H(X^\Delta, Y^\Delta)$ con $H(X, Y)$

4) Cuántos dígitos para (X, Y) con $n+m$ bits

Exponencial

Z v.a.c $\sim \frac{e^{-z/\lambda}}{\lambda}$ $z \in \mathbb{R}^+$ o mejor $A e^{-z/\lambda}$

$$1 = \int_0^{+\infty} A e^{-z/\lambda} dz = A\left(-\lambda e^{-z/\lambda}\right)\Big|_0^{+\infty} = A\lambda \; 1 \Rightarrow A = \frac{1}{\lambda}$$

1) $h(z)$

2) W v.a. ≥ 0 $\quad W \sim f$ de media λ

dem. $h(W) \leq h(z)$

$$\text{media } E(z) = \int_0^{+\infty} z \, \frac{e^{-z/\lambda}}{\lambda} dz = -z e^{-z/\lambda}\Big|_0^{+\infty} - \int_0^{+\infty} 1 \, e^{-z/\lambda} dz$$

$$= 0 - \lambda e^{-z/\lambda}\Big|_0^{+\infty} = \lambda$$

$$h(z) \text{ nats} = \int \frac{e^{-z/\lambda}}{\lambda} \ln\left(\frac{1}{\lambda} e^{-z/\lambda}\right) dz = \int \frac{e^{-z/\lambda}}{\lambda} \left(-\frac{z}{\lambda} + \ln\lambda\right)$$

$$= \ln\lambda + \frac{1}{\lambda}\lambda = \ln e\lambda \text{ nats} \qquad h(z) \text{ bits} = \log(\lambda e)$$

Mayor a V.A. de igual media

comparar con W de densidad $f(w)$

Exponencial

$D(f\|g) \geq 0$ $\quad D(f\|g) = E_f \log\left(\frac{f}{g}\right) = -h(f) - E_f(\log g)$

Construyo la exponencial $g(z) = \frac{1}{\lambda} e^{-z/\lambda}$ con $\lambda = E_w(W)$

Es decir igual media.

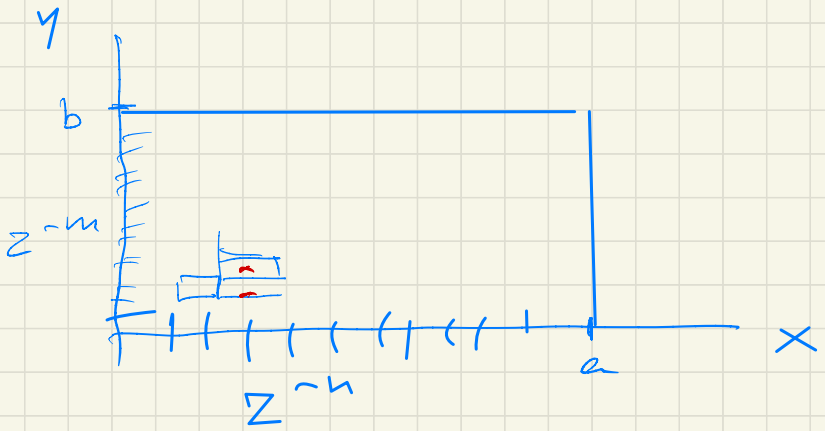
$$E_f(\log g) = E_f[\text{ctes} + \text{términos lineales}) =$$

$$= E_f[A + Bz] = E_g[\log g] = h(g)$$

cambio esencial

Entonces $h(g) \geq h(f)$

uniformes



$$h(x,y) = h(x) + h(y)$$

$$H(x^\Delta, y^\Delta) = H(x^\Delta) + H(y^\Delta)$$

$$H(x^\Delta, y^\Delta) = h(x,y) + n + m$$

$$h(x) = \log a$$

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

$$H(x^\Delta) = h(x) - \log 2^{-n} = h(x) + n$$

$$H(y^\Delta) = h(y) + m$$

Son independientes.