Introducción a la Teoría de la Información

Conceptos Básicos.

Facultad de Ingeniería, UdelaR

Año 2023

Agenda

1. Definiciones y Propiedades Básicas
   - Entropía
   - Divergencia, Entropía Relativa, o Distancia KL

Agenda

1. Definiciones y Propiedades Básicas
   - Entropía
   - Divergencia, Entropía Relativa, o Distancia KL
2. Propiedades
   - Desigualdad de Jensen
   - Desigualdad Log Sum

Agenda

1. Definiciones y Propiedades Básicas
   - Entropía
   - Divergencia, Entropía Relativa, o Distancia KL
2. Propiedades
   - Desigualdad de Jensen
   - Desigualdad Log Sum
3. Cadenas de Markov
   - Desigualdad de Fano

Definición de Entropía

Definición

$X \sim p$ variable aleatoria con valores en un alfabeto finito $\mathcal{X}$.

$$H(X) = E_p\Big[-\log p(X)\Big]$$

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

Logaritmos son en base 2 y convenimos $0\log 0 = 0$.
La entropía se expresa en *bits* y es una medida de la incertidumbre, o la cantidad de información de X en promedio.

Definición de Entropía

Definición

$X \sim p$ variable aleatoria con valores en un alfabeto finito $\mathcal{X}$.

$$H(X) = E_p\Big[-\log p(X)\Big]$$

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

Logaritmos son en base 2 y convenimos $0\log 0 = 0$.
La entropía se expresa en *bits* y es una medida de la incertidumbre, o la cantidad de información de X en promedio.

- $H(X) \geq 0$ ya que $-\log p(x) \geq 0$
- Si $p(x) = 1/|\mathcal{X}|$ para todo x, $H(X) = \log|\mathcal{X}|$.

Entropía como función de una distribución

Definición

Si $\mathbf{p}$ es un vector de probabilidad, $\mathbf{p} = (p_1 \dots p_m)$, la entropía de $\mathbf{p}$ está dada por

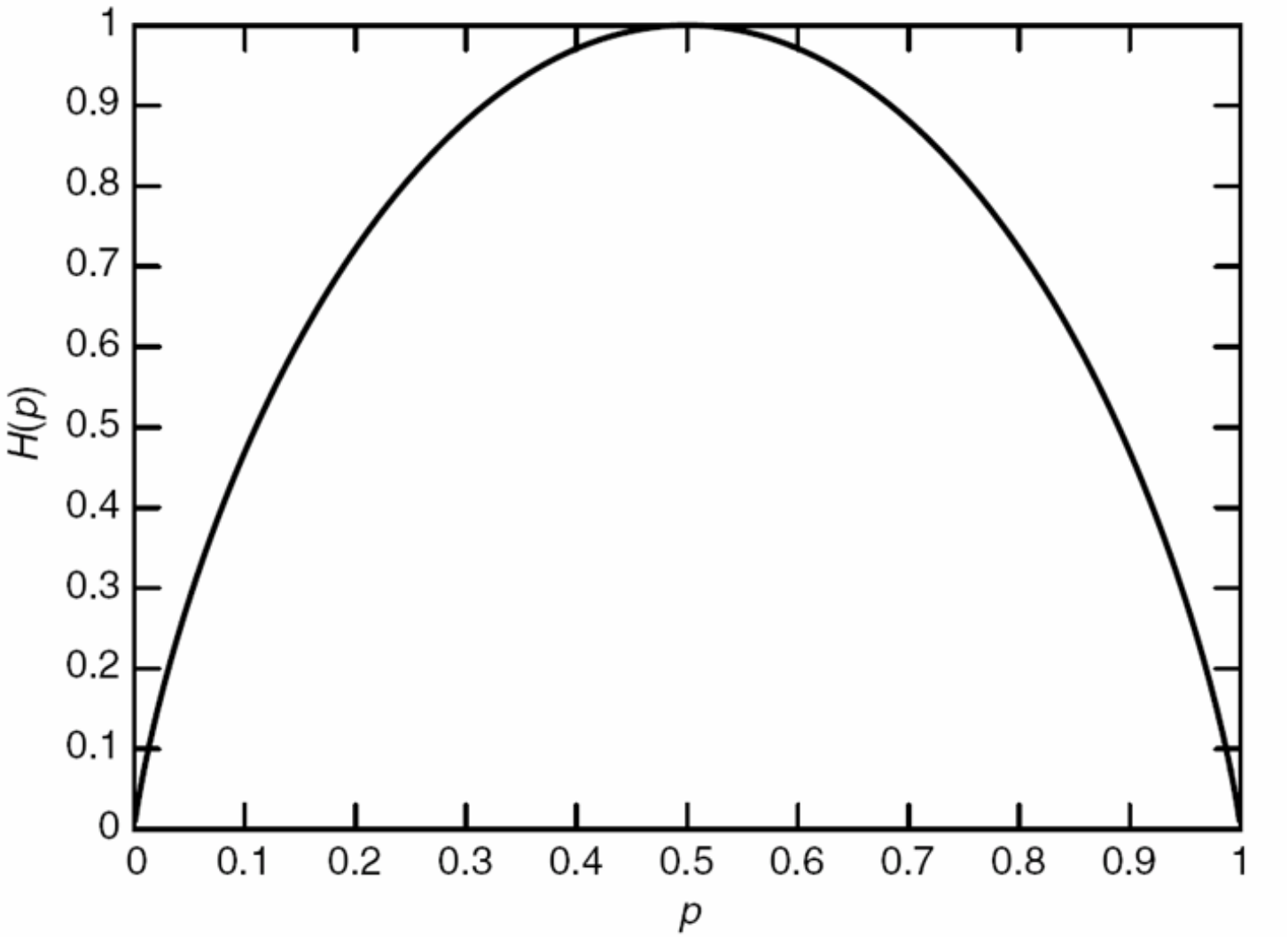
$$H(\mathbf{p}) = -\sum_{i=1}^{m} p_i \log p_i$$

Definición

En particular cuando $m = 2$, $\mathbf{p}$ es de la forma $\mathbf{p} = (p, 1-p)$, $p \in [0, 1]$. La entropía de $\mathbf{p}$ como función del escalar p se denomina *función de entropía binaria*, y la denotamos $H(p)$,

$$H(p) = -p \log p - (1-p) \log(1-p)$$

Entropía binaria



Algunas propiedades y desigualdades que usaremos en varias demostraciones

1. Regla de la cadena: $p(x,y) = p(x)p(y|x) = p(y)p(x|y)$
2. Logaritmo: $log(xy) = log(x) + log(y)$
3. Sumatoria con variable muda: $\sum_j P(X = x_i, Y = y_j) = P(X = x_i)$

   $\sum_{x,y} p(x,y)f(x) = \sum_x f(x) \sum_y p(x,y) = \sum_x p(x)f(x)$
4. Jensen: si f convexa, $E[f(x)] \geq f(Ex)$.
5. LogSum: $a_1 log\frac{a_1}{b_1} + a_2 log\frac{a_2}{b_2} \geq (a_1 + a_2) log(\frac{a_1+a_2}{b_1+b_2})$

Entropía conjunta y condicional

Definición

$$
\begin{aligned}
H(X,Y) &= E_{p(x,y)}\Big[-\log p(X,Y)\Big] \\
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)
\end{aligned}
$$

Entropía conjunta y condicional

Definición

$$
\begin{aligned}
H(X,Y) &= E_{p(x,y)}\Big[-\log p(X,Y)\Big] \\
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)
\end{aligned}
$$

Definición

$$
\begin{aligned}
H(Y|X) &= E_{p(x,y)}\Big[-\log p(Y|X)\Big] \\
H(Y|X) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\
&= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x)
\end{aligned}
$$

Regla de la cadena

Teorema

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

Regla de la cadena

Teorema

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

Demostración.

$$\begin{aligned}
p(X,Y) &= p(X)p(Y|X) \\
-\log p(X,Y) &= -\log p(X) - \log p(Y|X) \\
E_{p(x,y)}\left[-\log p(X,Y)\right] &= E_{p(x,y)}\left[-\log p(X)\right] + E_{p(x,y)}\left[-\log p(Y|X)\right] \\
H(X,Y) &= H(X) + H(Y|X)
\end{aligned}$$

□

Regla de la cadena

Teorema

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

Demostración.

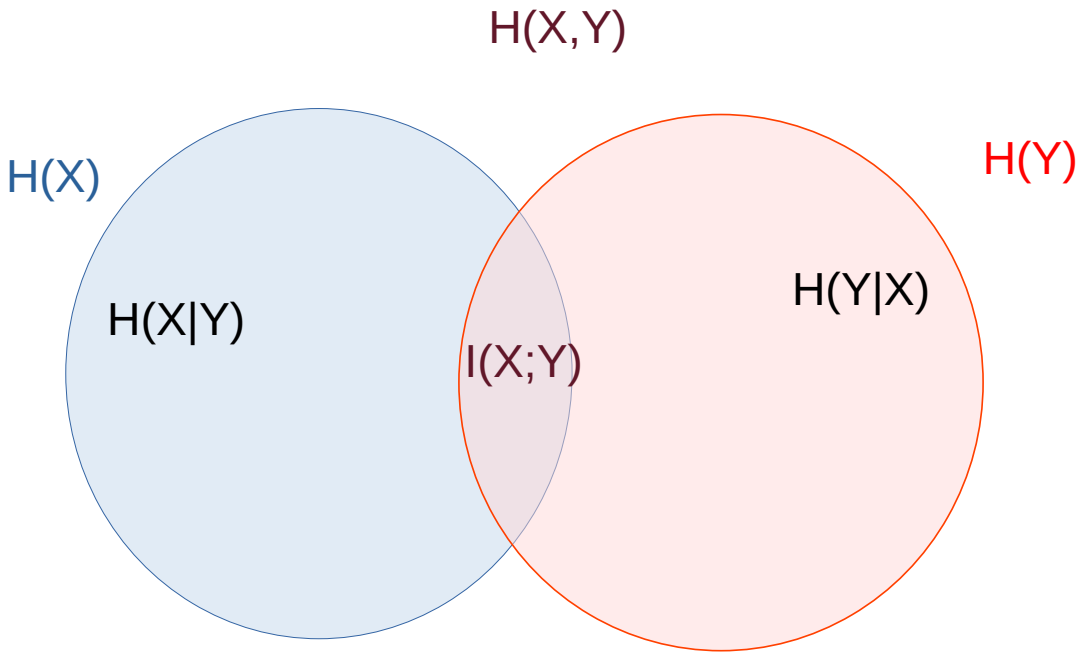
$$\begin{aligned}
p(X,Y) &= p(X)p(Y|X) \\
-\log p(X,Y) &= -\log p(X) - \log p(Y|X) \\
E_{p(x,y)}\left[-\log p(X,Y)\right] &= E_{p(x,y)}\left[-\log p(X)\right] + E_{p(x,y)}\left[-\log p(Y|X)\right] \\
H(X,Y) &= H(X) + H(Y|X)
\end{aligned}$$

□

Corolario

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

Información conjunta e información mutua



$H(X,Y) <= H(X)+H(Y)$ $H(X,Y) = H(X)+H(Y|X) = \ldots$

$I(X;Y) = H(x)-H(X|Y) = H(Y)-H(Y|X)$

Regla de la cadena

Corolario

$$
\begin{aligned}
H(X_1 \dots X_n) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1) \\
H(X_1 \dots X_n | Z) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1, Z)
\end{aligned}
$$

Regla de la cadena

Corolario

$$\begin{aligned} H(X_1 \dots X_n) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1) \\ H(X_1 \dots X_n|Z) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1, Z) \end{aligned}$$

Demostración.

La prueba es por inducción

$$\begin{aligned} H(X_1 \dots X_n) &= H(X_1) + H(X_2 \dots X_n|X_1) \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1} \dots X_2, X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1) \end{aligned}$$

Regla de la cadena (2)

Corolario

$$\begin{aligned} H(X_1 \dots X_n) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1) \\ H(X_1 \dots X_n|Z) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1, Z) \end{aligned}$$

Demostración.

La prueba es por inducción

$$\begin{aligned} H(X_1 \dots X_n|Z) &= H(X_1|Z) + H(X_2 \dots X_n|X_1, Z) \\ &= H(X_1|Z) + \sum_{i=2}^{n} H(X_i|X_{i-1} \dots X_2, X_1, Z) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1, Z) \end{aligned}$$

Ejemplo

$P(X,Y)$	x_1	x_2	x_3	x_4	$P(Y)$
y_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
y_4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
$P(X)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

Ejemplo

$P(X,Y)$	x_1	x_2	x_3	x_4	$P(Y)$
y_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
y_4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
$P(X)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$H(X) = \frac{7}{4}$, $H(Y) = 2bit$

Ejemplo

$P(X,Y)$	x_1	x_2	x_3	x_4	$P(Y)$
y_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
y_4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
$P(X)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$H(X) = \frac{7}{4}$, $H(Y) = 2bit$
Regla de la cadena: $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)|X = x_i)$

Ejemplo

$P(X,Y)$	x_1	x_2	x_3	x_4	$P(Y)$
y_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
y_3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
y_4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
$P(X)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$H(X) = \frac{7}{4}$, $H(Y) = 2bit$

Regla de la cadena: $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)|X = x_i)$

$P(X\|Y)$	x_1	x_2	x_3	x_4
y_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
y_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$
y_3	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
y_4	1	0	0	0

Divergencia, Entropía relativa, o Distancia KL

Definición

La Divergencia, Entropía Relativa, o Distancia de Kullback Leibler entre dos distribuciones de probabilidad sobre un mismo alfabeto $\mathcal{X}$ está dada por

$$\begin{aligned} D(p||q) &= E_p\left[\log \frac{p(X)}{q(X)}\right] \\ &= \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \end{aligned}$$

Convenimos $0 \log 0/q = 0$ para todo q, y $p \log p/0 = \infty$ para $p \neq 0$.
La Divergencia se expresa en bits.

Divergencia, Entropía relativa, o Distancia KL

Propiedades

- $D(p||q) \geq 0$ con igualdad si y sólo si $p = q$. Sin embargo no es simétrica y no cumple la desigualdad triangular.

- $D(p||q) = E_p\left[-\log q(X)\right] - E_p\left[-\log p(X)\right]$.
  En este sentido la divergencia mide la ineficiencia por usar q cuando la verdadera distribución es p.

- Desigualdad de Pinsker: $D(p||q) \geq \frac{1}{2\ln 2}\|p - q\|_1^2$

Entropía relativa conjunta y condicional

Definición

$$D(p(x,y)||q(x,y)) = E_{p(x,y)}\left[\log \frac{p(X,Y)}{q(X,Y)}\right]$$

$$D(p(x,y)||q(x,y)) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{q(x,y)}$$

Entropía relativa conjunta y condicional

Definición

$$D(p(x,y)||q(x,y)) = E_{p(x,y)}\left[\log \frac{p(X,Y)}{q(X,Y)}\right]$$

$$D(p(x,y)||q(x,y)) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{q(x,y)}$$

Definición

$$D(p(y|x)||q(y|x)) = E_{p(x,y)}\left[\log \frac{p(Y|X)}{q(Y|X)}\right]$$

$$D(p(y|x)||q(y|x)) = \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{q(y|x)}$$

Regla de la cadena

Teorema

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Regla de la cadena

Teorema

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Demostración.

$$\begin{aligned}
\frac{p(X,Y)}{q(X,Y)} &= \frac{p(X)p(Y|X)}{q(X)q(Y|X)} \\
\log \frac{p(X,Y)}{q(X,Y)} &= \log \frac{p(X)}{q(X)} + \log \frac{p(Y|X)}{q(Y|X)} \\
E_{p(x,y)}\left[\log \frac{p(X,Y)}{q(X,Y)}\right] &= E_{p(x,y)}\left[\log \frac{p(X)}{q(X)}\right] + E_{p(x,y)}\left[\log \frac{p(Y|X)}{q(Y|X)}\right] \\
D(p(x,y)||q(x,y)) &= D(p(x)||q(x)) + D(p(y|x)||q(y|x))
\end{aligned}$$

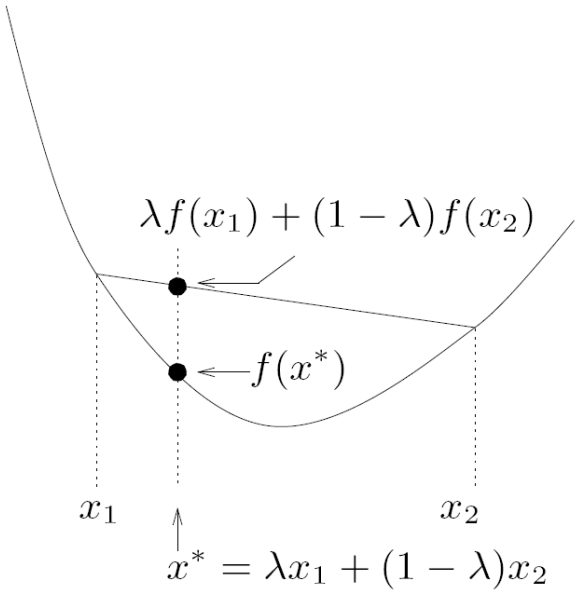
□

Funciones convexas

Definición

Una función f es *convexa* en un intervalo (a, b) si para todo $x_1, x_2 \in (a, b)$ y todo $\lambda \in [0, 1]$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$



Funciones convexas

Definición

Una función f es *convexa* en un intervalo (a, b) si para todo $x_1, x_2 \in (a, b)$ y todo $\lambda \in [0, 1]$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

Definición

f es *estrictamente convexa* si la desigualdad es estricta en $\lambda \in (0, 1)$.

Definición

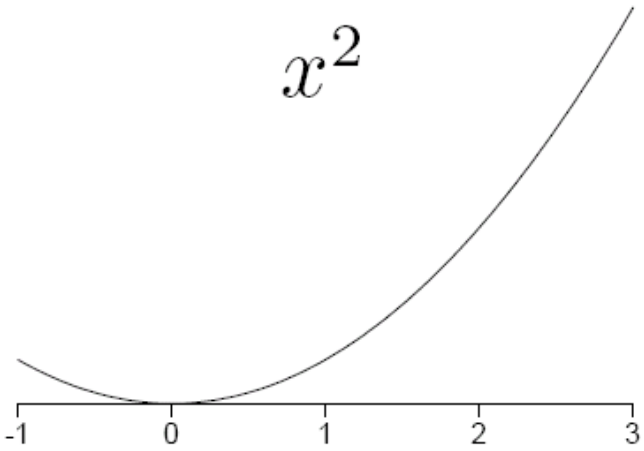
f es *(estrictamente) cóncava* si $-f$ es (estrictamente) convexa.

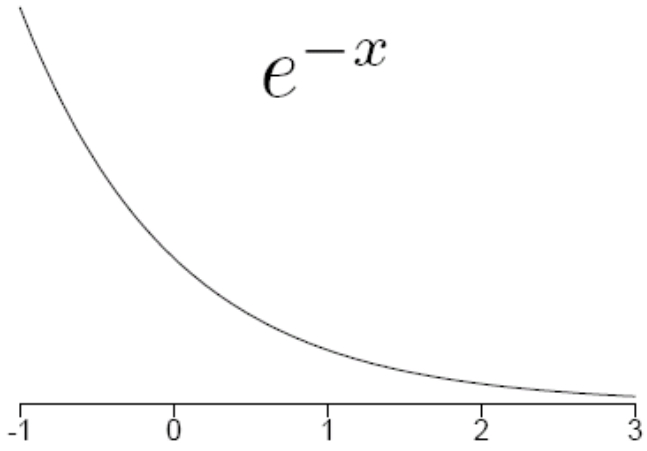
Condición suficiente para la convexidad

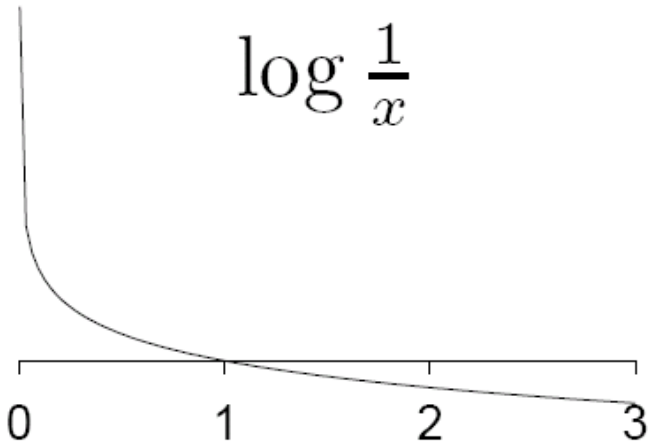
Teorema

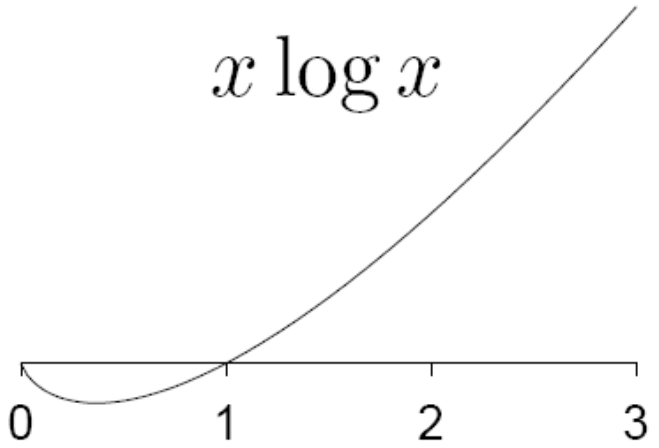
Si una función f tiene derivada segunda no negativa (positiva) en un intervalo (a, b), entonces f es convexa (estrictamente convexa) en (a, b)

Funciones convexas









Desigualdad de Jensen

Teorema

Desigualdad de Jensen: Si f es una función convexa y X una variable aleatoria,

$$E[f(X)] \geq f(E[X])$$

Desigualdad de Jensen

Teorema

Desigualdad de Jensen: Si f es una función convexa y X una variable aleatoria,

$$E\left[f(X)\right] \geq f(E\left[X\right])$$

Si se da la igualdad y f es estrictamente convexa, $X = E\left[X\right]$ con probabilidad 1 (X es una constante)

Desigualdad de Jensen

Demostración.

PB (dos puntos):

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) \quad \text{conv.}$$

PI (k puntos): definimos $p'_i = \frac{p_i}{1-p_k}$ para $i < k$

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \\
&\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) && \text{ind.} \\
&\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p'_i x_i\right) && \text{conv.} \\
&= f\left(\sum_{i=1}^{k} p_i x_i\right)
\end{aligned}$$

□

Desigualdad de la Información

Teorema

Desigualdad de la Información: $D(p||q) \geq 0$ con igualdad si y sólo si $p = q$

Demostración.

Definimos $Y = \begin{cases} q(X)/p(X) & \text{Si } p(X) > 0 \\ 0 & \text{Si } p(X) = 0 \end{cases}$

$$\begin{aligned} D(p||q) &= E_p\left[-\log Y\right] \\ &\geq -\log E_p\left[Y\right] \qquad\qquad Jensen \\ &= -\log \textstyle\sum_{p(x)\neq 0} p(x)\frac{q(x)}{p(x)} \\ &= -\log \textstyle\sum_{p(x)\neq 0} q(x) \geq 0 \end{aligned}$$

□

Desigualdad de la Información

Teorema

Desigualdad de la Información: $D(p||q) \geq 0$ con igualdad si y sólo si $p = q$

Demostración.

Definimos $Y = \begin{cases} q(X)/p(X) & \text{Si } p(X) > 0 \\ 0 & \text{Si } p(X) = 0 \end{cases}$

$$\begin{aligned} D(p||q) &= E_p\left[-\log Y\right] \\ &\geq -\log E_p\left[Y\right] \qquad\qquad Jensen \\ &= -\log \textstyle\sum_{p(x)\neq 0} p(x)\frac{q(x)}{p(x)} \\ &= -\log \textstyle\sum_{p(x)\neq 0} q(x) \geq 0 \end{aligned}$$

□

Si $D(p||q) = 0$, entonces $E_p\left[Y\right] = 1$.

Desigualdad de la Información

Teorema

Desigualdad de la Información: $D(p||q) \geq 0$ con igualdad si y sólo si $p = q$

Demostración.

Definimos $Y = \begin{cases} q(X)/p(X) & \text{Si } p(X) > 0 \\ 0 & \text{Si } p(X) = 0 \end{cases}$

$$\begin{aligned} D(p||q) &= E_p\left[-\log Y\right] \\ &\geq -\log E_p\left[Y\right] \qquad\qquad Jensen \\ &= -\log \textstyle\sum_{p(x)\neq 0} p(x)\frac{q(x)}{p(x)} \\ &= -\log \textstyle\sum_{p(x)\neq 0} q(x) \geq 0 \end{aligned}$$

□

Si $D(p||q) = 0$, entonces $E_p\left[Y\right] = 1$. $-\log$ estrictamente convexa $\Rightarrow Y = q/p = 1$

Vale para la divergencia condicional

Corolario

$D(p(y|x)||q(y|x)) \geq 0$ con igualdad si y sólo si $p(y|x) = q(y|x)$ para todo y y todo x con $p(x) > 0$

Demostración.

De la definición, $D(p(y|x)||q(y|x))$ es un promedio de valores no negativos.

$$D(p(y|x)||q(y|x)) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

□

Entropía máxima

Teorema

$H(X) \leq \log |\mathcal{X}|$ con igualdad si y sólo si X tiene distribución uniforme sobre $\mathcal{X}$.

Entropía máxima

Teorema

$H(X) \leq \log |\mathcal{X}|$ con igualdad si y sólo si X tiene distribución uniforme sobre $\mathcal{X}$.

Demostración.

Sea $u(x) = \frac{1}{|\mathcal{X}|}$ la distribución uniforme sobre $\mathcal{X}$.

$$\begin{aligned} D(p||u) &= E_p\left[\log \tfrac{1}{u(X)}\right] - E_p\left[\log \tfrac{1}{p(X)}\right] \\ &= \log |\mathcal{X}| - H(X) \\ &\geq 0 \end{aligned}$$

□

Cota de Independencia

Sean $X_1 \dots X_n$ con distribución $p(X_1, \dots X_n)$
$H(X_1 \dots X_n) \leq \sum_{i=1}^{n} H(X_i)$ con igualdad si y sólo si las variables X_i son independientes.

Cota de Independencia

Sean $X_1 \dots X_n$ con distribución $p(X_1, \dots X_n)$
$H(X_1 \dots X_n) \leq \sum_{i=1}^{n} H(X_i)$ con igualdad si y sólo si las variables X_i son independientes.

Consideremos la divergencia entre dos conjuntos de probabilidad: $p(X_1, \dots, X_n)$ y $p(X_1) \dots p(X_n)$. Ambos son conjuntos de probabilidad.

$$D(p(X_1, \dots, X_n)||p(X_1) \dots p(X_n)) = E_{p(X_1,\dots,X_n)} \left[\log \frac{(p(X_1 \dots X_n))}{p(X_1) \dots p(X_n)}\right]$$

$$D(p(X_1, \dots, X_n)||p(X_1) \dots p(X_n)) \geq 0$$

$$E_{p(X_1,\dots,X_n)} [\log(p(X_1, \dots, X_n))] = -H(X)$$

$$E_{p(X_1,\dots,X_n)} [\log((p(X_1) \dots p(X_n))] = E_{p(X_i,\dots,X_n)} \left[\sum_i \log(p(X_i)\right] =$$
$$\sum_i E_{p(X_1,\dots,X_n)} [\log p(x_i)] = \sum_i E_{p(X_i)} [\log p(X_i)] = -\sum_i H(X_i)$$

Cota de Independencia

Sean $X_1 \dots X_n$ con distribución $p(X_1, \dots X_n)$
$H(X_1 \dots X_n) \leq \sum_{i=1}^{n} H(X_i)$ con igualdad si y sólo si las variables X_i son independientes.

Consideremos la divergencia entre dos conjuntos de probabilidad: $p(X_1, \dots, X_n)$ y $p(X_1) \dots p(X_n)$. Ambos son conjuntos de probabilidad.

$$D(p(X_1, \dots, X_n)||p(X_1) \dots p(X_n)) = E_{p(X_1,\dots,X_n)} \left[\log \frac{(p(X_1 \dots X_n))}{p(X_1) \dots p(X_n)}\right]$$

$$D(p(X_1, \dots, X_n)||p(X_1) \dots p(X_n)) \geq 0$$

$$E_{p(X_1,\dots,X_n)} [\log(p(X_1, \dots, X_n))] = -H(X)$$

$$E_{p(X_1,\dots,X_n)} [\log((p(X_1) \dots p(X_n))] = E_{p(X_i,\dots,X_n)} \left[\sum_i \log(p(X_i)\right] =$$
$$\sum_i E_{p(X_1,\dots,X_n)} [\log p(x_i)] = \sum_i E_{p(X_i)} [\log p(X_i)] = -\sum_i H(X_i)$$

Cota de independencia condicional
$H(X_1 \dots X_n|Z) \leq \sum_{i=1}^{n} H(X_i|Z)$ con igualdad si y sólo si X_i son condicionalmente independientes dado Z.

Condicionar reduce la entropía

Teorema

$H(X|Y) \leq H(X)$ con igualdad si y sólo si X, Y son independientes.

Condicionar reduce la entropía

Teorema

$H(X|Y) \leq H(X)$ con igualdad si y sólo si X, Y son independientes.

Demostración.

$$
\begin{aligned}
& H(X) - H(X|Y) = \\
&= H(X) + H(Y) - (H(Y) + H(X|Y)) \\
&= H(X) + H(Y) - H(X,Y) \\
&= E_{p(x,y)}\left[\log \frac{1}{p(X)} + \log \frac{1}{p(Y)}\right] - E_{p(x,y)}\left[\log \frac{1}{p(X,Y)}\right] \\
&= E_{p(x,y)}\left[\log \frac{1}{p(X)p(Y)}\right] - E_{p(x,y)}\left[\log \frac{1}{p(X,Y)}\right] \\
&= D(p(x,y)||p(x)p(y)) \geq 0
\end{aligned}
$$

□

Cota de Independencia

Teorema

$H(X_1 \dots X_n) \leq \sum_{i=1}^{n} H(X_i)$ con igualdad si y sólo si X_i son independientes.

Cota de Independencia

Teorema

$H(X_1 \dots X_n) \leq \sum_{i=1}^{n} H(X_i)$ con igualdad si y sólo si X_i son independientes.

Demostración.

$$\begin{aligned} H(X_1 \dots X_n) = & \sum_{i=1}^{n} H(X_i | X_{i-1} \dots X_1) \\ \leq & \sum_{i=1}^{n} H(X_i) \end{aligned}$$

La desigualdad se da término a término $\Rightarrow$ cuando hay igualdad, cada X_i debe ser independiente de $X_{i-1} \dots X_1$ $\square$

Cota de Independencia condicional

Teorema

$H(X_1 \dots X_n|Z) \leq \sum_{i=1}^{n} H(X_i|Z)$ con igualdad si y sólo si X_i son condicionalmente independientes dado Z.

Cota de Independencia condicional

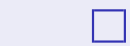
Teorema

$H(X_1 \dots X_n|Z) \leq \sum_{i=1}^{n} H(X_i|Z)$ con igualdad si y sólo si X_i son condicionalmente independientes dado Z.

Demostración.

$$\begin{aligned} H(X_1 \dots X_n|Z) &= \sum_{i=1}^{n} H(X_i|X_{i-1} \dots X_1, Z) \\ &\leq \sum_{i=1}^{n} H(X_i|Z) \end{aligned}$$

La desigualdad se da término a término $\Rightarrow$ cuando hay igualdad, cada X_i debe ser independiente de $X_{i-1} \dots X_1$ dado Z.



Cota de Independencia condicional

Teorema

$H(X_1 \dots X_n|Z) \leq \sum_{i=1}^{n} H(X_i|Z)$ con igualdad si y sólo si X_i son condicionalmente independientes dado Z.

Demostración.

$$\begin{aligned} H(X_1 \dots X_n|Z) = & \sum_{i=1}^{n} H(X_i|X_{i-1} \dots X_1, Z) \\ \leq & \sum_{i=1}^{n} H(X_i|Z) \end{aligned}$$

La desigualdad se da término a término $\Rightarrow$ cuando hay igualdad, cada X_i debe ser independiente de $X_{i-1} \dots X_1$ dado Z.

□

También se puede ver como la divergencia entre $p(X_1, \dots X_n|Z)$ y $p(X_1|Z) \dots p(X_n|Z)$.

Desigualdad Log Sum

Teorema

Sean $a_1 \dots a_n$ y $b_1 \dots b_n$ números no negativos.

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

con igualdad si y sólo si a_i/b_i es constante.
Nuevamente $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ si $a \neq 0$, y $0 \log \frac{0}{0} = 0$

Desigualdad Log Sum

Demostración.

Sea X v.a. en $\mathcal{X} = \{x_i = a_i/b_i : i = 1 \dots n\}$. $f(x) = x \log x$ es estrictamente convexa en $x \geq 0$:

$$E[X \log X] \geq E[X] \log E[X]$$

Para una distribución $p_i = \frac{b_i}{\sum b_j}$

$$\sum_i \left(\frac{b_i}{\sum_j b_j}\right) \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \left(\sum_i \left(\frac{b_i}{\sum_j b_j}\right) \frac{a_i}{b_i}\right) \log \sum_i \left(\frac{b_i}{\sum_j b_j}\right) \frac{a_i}{b_i}$$

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

Igualdad si y sólo si $X = cte$ □

Desigualdad de la Información (II)

Teorema

$D(p||q) \geq 0$ con igualdad si y sólo si $p = q$

Demostración.

$$\begin{aligned} D(p||q) = & \sum p(x) \log \frac{p(x)}{q(x)} \\ \geq & (\sum p(x)) \log \frac{\sum p(x)}{\sum q(x)} \\ = & 1 \log \frac{1}{1} = 0 \end{aligned}$$

□

Desigualdad de la Información (II)

Teorema

$D(p||q) \geq 0$ con igualdad si y sólo si $p = q$

Demostración.

$$\begin{aligned} D(p||q) = & \sum p(x) \log \frac{p(x)}{q(x)} \\ \geq & (\sum p(x)) \log \frac{\sum p(x)}{\sum q(x)} \\ = & 1 \log \frac{1}{1} = 0 \end{aligned}$$

□

Si se da la igualdad, $\frac{p}{q} = cte = 1$ porque ambas suman 1.

Convexidad de $D(p||q)$

Teorema

$D(p||q)$ es convexa en el par (p, q), es decir

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

para $0 \leq \lambda \leq 1$

Convexidad de $D(p||q)$

Teorema

$D(p||q)$ es convexa en el par (p, q), es decir

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

para $0 \leq \lambda \leq 1$

Observación

$D(p||q)$ es convexa en p para q fija y viceversa.

Convexidad de $D(p||q)$

Demostración.

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\overbrace{\lambda p_1(x)}^{a_1} + \overbrace{(1-\lambda)p_2(x)}^{a_2}}{\underbrace{\lambda q_1(x)}_{b_1} + \underbrace{(1-\lambda)q_2(x)}_{b_2}}$$

$$\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}$$

Convexidad de $D(p||q)$

Demostración.

$$\sum_x (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\overbrace{\lambda p_1(x)}^{a_1} + \overbrace{(1-\lambda)p_2(x)}^{a_2}}{\underbrace{\lambda q_1(x)}_{b_1} + \underbrace{(1-\lambda)q_2(x)}_{b_2}}$$

$$\leq \lambda \sum_x p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) \sum_x p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}$$

Concavidad de la entropía

Teorema

$H(p)$ es una función cóncava de p, es decir para $0 \leq \lambda \leq 1$

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$$

Concavidad de la entropía

Teorema

$H(p)$ es una función cóncava de p, es decir para $0 \leq \lambda \leq 1$

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$$

Demostración.

Sea $X \sim p$, u distribución uniforme en $\mathcal{X}$

$$\begin{aligned} D(p||u) = & \quad E_p\left[\log \tfrac{1}{u(X)}\right] - E_p\left[\log \tfrac{1}{p(X)}\right] \\ = & \quad \log|\mathcal{X}| - H(p) \end{aligned}$$

La concavidad de H surge de la convexidad de D. □

$X \to Y \to Z$

Definición

X, Y, Z *forman una cadena de Markov* y se denota $X \to Y \to Z$ si la distribución condicional de Z depende sólo de Y.

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

$X \to Y \to Z$

Definición

X, Y, Z *forman una cadena de Markov* y se denota $X \to Y \to Z$ si la distribución condicional de Z depende sólo de Y.

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

En general $p(x, y, z) = p(x)p(y, z|x) = p(x)p(y|x)p(z|y, x)$
Es decir, si es Markov, $p(z|y, x) = p(z|y)$

$X \to Y \to Z$

- $X \to Y \to Z$ si y sólo si X, Z son condicionalmente independientes dado Y.

$$
\begin{aligned}
(\Rightarrow) \quad & p(x,z|y) = \frac{p(x,z,y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y) \\
(\Leftarrow) \quad & p(z|y,x) = \frac{p(x,z|y)}{p(x|y)} = \frac{p(x|y)p(z|y)}{p(x|y)} = p(z|y)
\end{aligned}
$$

$X \to Y \to Z$

- $X \to Y \to Z$ si y sólo si X, Z son condicionalmente independientes dado Y.

$$
\begin{aligned}
(\Rightarrow) \quad & p(x,z|y) = \frac{p(x,z,y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y) \\
(\Leftarrow) \quad & p(z|y,x) = \frac{p(x,z|y)}{p(x|y)} = \frac{p(x|y)p(z|y)}{p(x|y)} = p(z|y)
\end{aligned}
$$

- $X \to Y \to Z \Leftrightarrow Z \to Y \to X$

$X \to Y \to Z$

- $X \to Y \to Z$ si y sólo si X, Z son condicionalmente independientes dado Y.

$$
\begin{aligned}
(\Rightarrow) \quad & p(x,z|y) = \frac{p(x,z,y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y) \\
(\Leftarrow) \quad & p(z|y,x) = \frac{p(x,z|y)}{p(x|y)} = \frac{p(x|y)p(z|y)}{p(x|y)} = p(z|y)
\end{aligned}
$$

- $X \to Y \to Z \Leftrightarrow Z \to Y \to X$
- Si $Z = f(Y)$, $X \to Y \to Z$

$X \to Y \to Z$

- $X \to Y \to Z$ si y sólo si X, Z son condicionalmente independientes dado Y.

$$
\begin{aligned}
(\Rightarrow) \quad & p(x,z|y) = \frac{p(x,z,y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y) \\
(\Leftarrow) \quad & p(z|y,x) = \frac{p(x,z|y)}{p(x|y)} = \frac{p(x|y)p(z|y)}{p(x|y)} = p(z|y)
\end{aligned}
$$

- $X \to Y \to Z \Leftrightarrow Z \to Y \to X$
- Si $Z = f(Y)$, $X \to Y \to Z$
- Si $X \to Y \to Z$, entonces $H(Z|Y) = H(Z|X,Y) \leq H(Z|X)$

Desigualdad de Procesamiento de Datos

Teorema

Si $X \to Y \to Z$, entonces $I(X;Y) \geq I(X;Z)$

Desigualdad de Procesamiento de Datos

Teorema

Si $X \to Y \to Z$, entonces $I(X;Y) \geq I(X;Z)$

Demostración.

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \end{aligned}$$

□

Desigualdad de Procesamiento de Datos

Teorema

Si $X \to Y \to Z$, entonces $I(X;Y) \geq I(X;Z)$

Demostración.

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \end{aligned}$$

□

Corolario

$I(X;Y|Z) \leq I(X;Y)$

Desigualdad de Procesamiento (Determinista) de Datos

Corolario

En particular $I(X;Y) \geq I(X;f(Y))$

Demostración.

$X, Y, Z = f(Y)$ forman una cadena de Markov. □

Desigualdad de Fano

Teorema

Sean X, Y variables aleatorias y $\hat{X} = f(Y)$ una estimación de X. Entonces $X, Y, \hat{X}$ son una cadena de Markov. Sea $P_e = P\{\hat{X} \neq X\}$

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

Desigualdad de Fano

Teorema

Sean X, Y variables aleatorias y $\hat{X} = f(Y)$ una estimación de X. Entonces $X, Y, \hat{X}$ son una cadena de Markov. Sea $P_e = P\{\hat{X} \neq X\}$

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

Corolario

$$1 + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)}$$

Desigualdad de Fano

Demostración.

Definimos $E = \begin{array}{ll} 1 & \text{si } \hat{X} \neq X \\ 0 & \text{si } \hat{X} = X \end{array}$

$$\begin{array}{rcccc} H(E,X|Y) = & H(X|Y) & + & \overbrace{H(E|X,Y)}^{=0} \\ = & \underbrace{H(E|Y)}_{\leq H(P_e)} & + & \underbrace{H(X|E,Y)}_{\leq P_e \log(|\mathcal{X}|-1)} \end{array}$$

$$\begin{array}{rcccc} H(X|E,Y) = & (1-P_e)H(X|Y,E=0) & + & P_e H(X|Y,E=1) \\ \leq & 0 & + & P_e \log(|\mathcal{X}|-1) \end{array}$$

□