Introducción a la Teoría de la Información

Conceptos Básicos.

Facultad de Ingeniería, UdelaR

Agenda

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

2. Propiedades
   - Desigualdad de la Información
   - Cotas para la entropía
   - Convexidad de la divergencia
   - Concavidad de la entropía
   - Cadenas de Markov

Definición de Entropía

Definición (Entropía de un variable aleatoria)

La *entropía* de un variable aleatoria $X \sim p$ con valores en un alfabeto finito $\mathcal{X}$ se define como

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

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 media.

- $H(X) \geq 0$ 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 (Entropía de un vector de probabilidad)

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

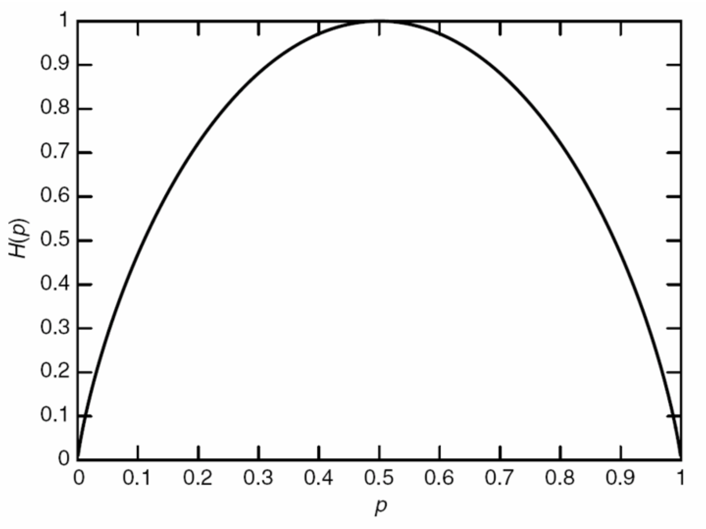
$$H(\mathbf{p}) = -\sum_{i=1}^{m} p_i \log p_i \,.$$

Definición (Entropía binaria)

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



Entropía conjunta y condicional

Definición (Entropía conjunta)

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

Definición (Entropía condicional)

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

Regla de la cadena

Teorema (Regla de la cadena)

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

Demostración.

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

$\square$

Teorema

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

Regla de la cadena

Corolario

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

Demostración.

La prueba es por inducción

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

$\square$

Regla de la cadena (2)

Corolario

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

Demostración.

La prueba es por inducción:

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

□

Divergencia, Entropía relativa o Distancia KL

Definición (Divergencia, Entropía Relativa o Distancia de Kullback Leibler)

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.
- Desigualdad de Pinsker: $D(p||q) \geq \frac{1}{2\ln 2}\|p-q\|_1^2$.
- $D(p||q) = E_p\left[-\log q(X)\right] - E_p\left[-\log p(X)\right]$.
  En este sentido la divergencia mide la ineficiencia por usar q cuando la verdadera distribución es p.

Entropía relativa conjunta y condicional

Definición (Entropía relativa conjunta)

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

Definición (Entropía relativa condicional)

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

Regla de la cadena

Teorema (Regla de la cadena para D)

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

Demostración.

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

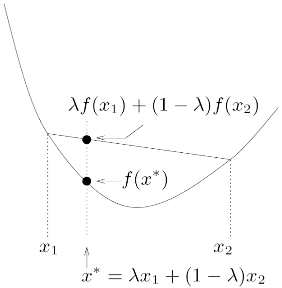
$\square$

Funciones convexas

Definición (Función convexa)

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

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



Funciones convexas

Definición (Función convexa)

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

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

Definición (Convexidad estricta)

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

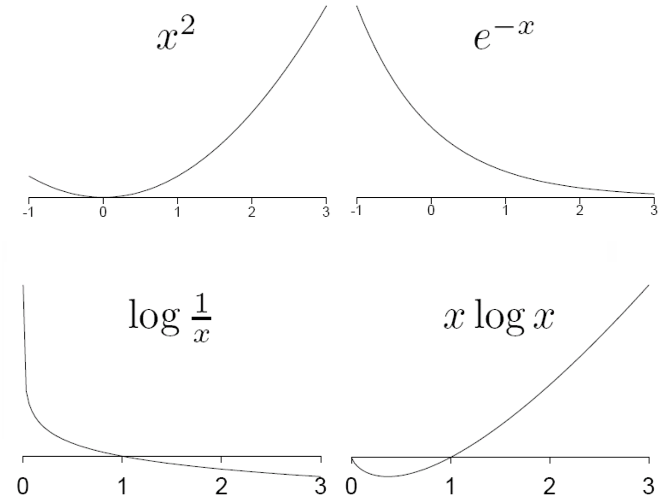
Definición (Función cóncava)

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

Teorema (Condición suficiente para la convexidad)

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, se cumple

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

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

Desigualdad de Jensen

Demostración.

PB (dos puntos):

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

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

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

□

Desigualdad de la Información

Teorema (Desigualdad de la Información)

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

Demostración.

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

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

□

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

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 $\mathcal{X}$.

Demostración.

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

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

$\square$

Cota de Independencia

Teorema

Sean $X_1 \ldots X_n \sim p$. Se cumple

$$\sum_{i=1}^{n} H(X_i) \geq H(X_1 \ldots X_n)\,,$$

con igualdad si y sólo si X_i son independientes.

Demostración.

$$\begin{aligned}
\left(\sum_{i=1}^{n} H(X_i)\right) - H(X_1 \ldots X_n) &= \left(\sum_{i=1}^{n} E_p\left[-\log p(X_i)\right]\right) - E_p\left[-\log p(X_1 \ldots X_n)\right] \\
&= E_p\left[\log \frac{p(X_1 \ldots X_n)}{\prod_{i=1}^{n} p(X_i)}\right] \\
&= D(p||q) \\
&\geq 0\,,
\end{aligned}$$

donde $q(x_1 \ldots x_n) = \prod_{i=1}^{n} p(x_i)$. □

Cota de Independencia condicional

Teorema

Sean $X_1 \dots X_n, Z \sim p$. Se cumple

$$\sum_{i=1}^{n} H(X_i|Z) \geq H(X_1 \dots X_n|Z)\,,$$

con igualdad si y sólo si X_i son condicionalmente independientes dado Z.

Demostración.

$$\begin{aligned}
&\left(\sum_{i=1}^{n} H(X_i|Z)\right) - H(X_1 \dots X_n|Z) \\
&= \left(\sum_{i=1}^{n} E_p\left[-\log p(X_i|Z)\right]\right) - E_p\left[-\log p(X_1 \dots X_n|Z)\right] \\
&= E_p\left[\log \frac{p(X_1 \dots X_n|Z)}{\prod_{i=1}^{n} p(X_i|Z)}\right] \\
&= D(p(x_1 \dots x_n|z)||q(x_1 \dots x_n|z)) \geq 0\,,
\end{aligned}$$

donde $q(x_1 \dots x_n|z) = \prod_{i=1}^{n} p(x_i|z)$. □

Condicionar reduce la entropía

Teorema

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

Demostración.

$$H(X|Y) = H(X,Y) - H(Y) = \Big(H(X,Y) - H(Y) - H(X)\Big) + H(X) \leq H(X)\,. \qquad \square$$

Teorema

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

Demostración.

$$\begin{aligned} H(X|Y,Z) &= H(X,Y,Z) - H(Y,Z) \\ &= H(X,Y|Z) + H(Z) - H(Y|Z) - H(Z) \\ &= \Big(H(X,Y|Z) - H(Y|Z) - H(X|Z)\Big) + H(X|Z) \leq H(X|Z)\,. \end{aligned}$$

$\square$

Desigualdad Log Sum

Teorema (Desigualdad Log Sum)

Sean $a_1 \dots a_n$ y $b_1 \dots b_n$ números no negativos. Entonces, se cumple

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

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 $\mathcal{X} = \{x_i = a_i/b_i : i = 1 \ldots n\}$. Como la función $f(x) = x \log x$ es estrictamente convexa en $x \geq 0$, entonces, por la desigualdad de Jensen, tenemos

$$E\left[X \log X\right] \geq E\left[X\right] \log E\left[X\right] .$$

En particular, para la distribución dada por $p_i = \frac{b_i}{\sum_j b_j}$, obtenemos

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

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

y la igualdad se da si y sólo si X es constante. $\square$

Desigualdad de la Información (II)

Teorema (Desigualdad de la Información)

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

Demostración.

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

□

Si se da la igualdad, $\frac{p}{q}$ es constante, necesariamente igual a 1 porque p y q suman 1.

Convexidad de $D(p||q)$

Teorema (Convexidad de la divergencia)

$D(p||q)$ es convexa en el par (p, q), es decir, para $0 \leq \lambda \leq 1$ se cumple

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

Observación

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

Convexidad de $D(p||q)$

Demostración.

Para cada $x \in \mathcal{X}$, por la desigualdad Log Sum tenemos

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

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

Convexidad de $D(p||q)$

Demostración.

Sumando en $x \in \mathcal{X}$ obtenemos

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

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

□

Concavidad de la entropía

Teorema (Concavidad de la entropía)

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

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

Demostración.

Sea u la distribución uniforme sobre $\mathcal{X}$ y sea $X \sim p$. Entonces podemos escribir

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

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

$X \to Y \to Z$

Definición (Variables que forman una cadena de Markov)

X, Y, Z *forman una cadena de Markov* y se denota $X \to Y \to Z$, si la distribución condicional de Z dadas X, Y depende sólo de Y. En este caso podemos escribir

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

$X \to Y \to Z$

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

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

- $X \to Y \to Z$ si y sólo si $Z \to Y \to X$.
- Si $Z = f(Y)$, $X \to Y \to Z$.
- Si $X \to Y \to Z$, entonces $H(Z|Y) = H(Z|X,Y) \leq H(Z|X)$.