Compresión de Datos sin Pérdida

Álvaro Martín

[1]Instituto de Computación,
Facultad de Ingeniería
almartin@fing.edu.uy

[2]PEDECIBA Informática

Recordar

- $\log xy = \log x + \log y$.
- $\log \frac{x}{y} = \log x - \log y$.
- $\log x^n = n \log x$.
- $-\log x = \log \frac{1}{x}$.
- $\log_a x = \log_a b^{\log_b x} = (\log_b x)(\log_a b)$.
- En particular, $\log x = (\log e) \ln x$, y $\ln x = (\ln 2) \log x$

Notación

- Mayúsculas, como X, Y, etc., denotan variables aleatorias.
- El alfabeto de valores que pueden tomar esas variables lo denotamos $\mathcal{X}$, $\mathcal{Y}$, etc., o también $\mathcal{A}$ en general.
- Minúsculas, como x, y, etc., denotan valores específicos en esos alfabetos.
- X_i^j denota la secuencia $X_i, X_{i+1}, \dots, X_j$.
- X^n denota la secuencia $X_1, X_2, \dots, X_n$.
- Notación análoga para valores específicos, con letras minúsculas.

Notación

- $P\{X = x\}$ o $P(X = x)$ es la probabilidad de que X tome el valor x.
- Cuando no hay ambigüedad, escribimos abreviadamente $p(x)$ para denotar $P\{X = x\}$.
- Si es necesario, escribimos P_X para hacer explícito que nos referimos a la distribución de probabilidad que gobierna a X.
- $E_p[f(X)]$ denota la esperanza de $f(X)$ con respecto a la distribución de probabilidad p. Omitimos p cuando no hay ambigüedad.

$X \in \{1, 2, 3\}$

$P(1) = 0{,}2$

$P(2) = 0{,}5$

$P(3) = 0{,}3$

$P(X)$

Si sale $x = 2$

$\Rightarrow P(x) = 0{,}5$

$p(x) = 1/6$, para $x = 4$

$E[Y]$

$\boxed{P(X)?}$

Definición de entropía

- La *entropía* de una variable aleatoria discreta $X \sim p$ definida sobre un *alfabeto* $\mathcal{X}$ es

$$H(X) = E\Big[\underbrace{-\log p(X)}_{f(x)}\Big] = \underbrace{-\sum_{x\in\mathcal{X}} p(x)\log p(x)}\,,$$

con la convención $0 \log 0 = 0$.

- También la denotamos $H(p)$, donde p es una distribución de probabilidad.

$p = (p_1\ p_2\ \ldots\ p_m)$

$-\sum_{i=1}^{m} p_i \log p_i$

- $H(X)$ es una medida de la cantidad de información provista por X en media.

- $H(X)$ también se interpreta como una medida de la incertidumbre que tenemos con respecto a X.

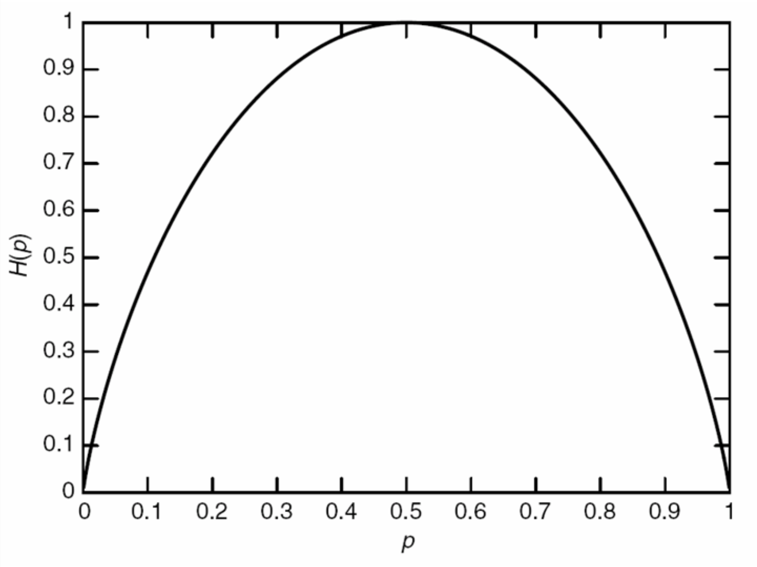
Entropía binaria

- Para alfabetos binarios, la distribución de probabilidad queda determinada por un valor escalar, $p \in [0, 1]$ (la probabilidad de uno de los dos símbolos).

- La entropía, como función del escalar p, se denomina *función de entropía binaria*

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

Gráfica de entropía binaria



Algunas propiedades de la entropía

$$H(X) = \sum_{x \in X} \frac{1}{|X|} \cdot \underbrace{\left(-\log \frac{1}{|X|}\right)}_{\log |X|}$$

- Simétrica.
- Cóncava.
- No negativa.
- Para X con distribución uniforme en $\mathcal{X}$, $H(X) = \log |\mathcal{X}|$.
- Si $p(x) = 1$ para algún $x \in \mathcal{X}$, entonces $H(X) = 0$.

Entropía conjunta y condicional

- Entropía conjunta

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

- Entropía condicional

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

$-\log p(y|x)$

Teorema (Regla de la cadena)

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

Demostración.

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

$\square$

Corolario

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

Regla de la cadena (generalización)

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

Para $i = 1$, $H(X_i | X_{i-1}, \ldots, X_1)$ denota $H(X_i)$.

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

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

$\ell(x)$ = # bits con q-e codifico x

$E[\ell(x)] = \sum_{x \in \mathcal{X}} p(x) \cdot \ell(x)$

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

donde convenimos $0 \log 0/q = 0$ para todo q, y $p \log p/0 = \infty$ para $p \neq 0$.

- Se expresa en bits, y mide la ineficiencia por usar q cuando la verdadera distribución es p

$$D(p\|q) = \underbrace{E_p\left[-\log q(X)\right]} - \overbrace{E_p\left[-\log p(X)\right]}^{H(x)}.$$

Desigualdad de la Información

Teorema (Desigualdad de la Información)
$D(p||q) \geq 0$ *con igualdad si y solo si* $p = q$.