Teoría de la Información

Ejercicios Tema 1

Antonio Coín Castro

30 de octubre de 2020

Ejercicio 1. Sea $X \sim \mathcal{N}(\mu, \sigma^2)$. Probar que la entropía de X viene dada por

$$S(X) = \frac{1}{2}\log_2(2\pi e\sigma^2).$$

Solución. Sabemos que en el caso continuo la entropía se calcula como

$$S = -\int_{\mathcal{X}} f(x)\log_2 f(x)\,dx.$$

En nuestro caso $\mathcal{X} = \mathbb{R}$ y $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, y usando que $\log_b x = \log_c x/\log_c b$ se tiene:

$$\begin{aligned}
S &= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}\frac{1}{\log 2}\log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right)dx\\
&= -\frac{1}{\log 2}\left(\int_{\mathbb{R}} f(x)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)dx - \int_{\mathbb{R}} f(x)\frac{(x-\mu)^2}{2\sigma^2}\,dx\right)\\
&= -\frac{1}{\log 2}\left(\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\underbrace{\int_{\mathbb{R}} f(x)\,dx}_{=1} - \frac{1}{2\sigma^2}\underbrace{\int_{\mathbb{R}} f(x)(x-\mu)^2\,dx}_{=\mathbb{E}[(X-\mu)^2]}\right)\\
&= \frac{-\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)}{\log 2} + \frac{Var[X]}{2\sigma^2\log 2} = \frac{\log(\sqrt{2\pi}\sigma)}{\log 2} + \frac{\sigma^2}{2\sigma^2\log 2}\\
&= \log_2(\sqrt{2\pi}\sigma) + \frac{\log e}{2\log 2} = \log_2(\sqrt{2\pi}\sigma) + \frac{1}{2}\log_2 e\\
&= \frac{1}{2}(\log_2((\sqrt{2\pi}\sigma)^2) + \log_2 e) = \frac{1}{2}\log_2(2\pi\sigma^2 e).
\end{aligned}$$

Ejercicio 2. Probar que la información mutua promedio $MI(X, Y)$ entre dos variables aleatorias X e Y es simétrica.

Solución. Es consecuencia directa de la definición, entendiendo que los símbolos $p(x, y)$ y $p(y, x)$ representan la misma distribución conjunta (X, Y):

$$MI(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(y,x)\log_2\left(\frac{p(y,x)}{p(y)p(x)}\right) = MI(Y,X),$$

donde hemos podido intercambiar las sumas ya que son numerables y todos los sumandos son no negativos. En efecto, como

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

sabemos que $p(x,y) \geq p(x)$ y también que $p(x,y) \geq p(y)$, luego $p(x,y) \geq p(x)p(y)$, y el argumento del logaritmo en la expresión de $MI(X,Y)$ es mayor o igual que 1.

En el caso continuo se llega a la misma conclusión aplicando el teorema de Fubini para intercambiar el orden de las integrales.

Ejercicio 3. Probar que $MI(X,Y) = S(X) + S(Y) - S(X,Y)$ para dos variables aleatorias X e Y.

Solución. Tenemos que:

$$\begin{aligned}
MI(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)(\log_2 p(x,y) - \log_2 p(x) - \log_2 p(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y) \\
&= -S(X,Y) - \sum_{x\in\mathcal{X}} \log_2 p(x)\sum_{y\in\mathcal{Y}} p(x,y) - \sum_{y\in\mathcal{Y}}\log_2 p(y)\sum_{x\in\mathcal{X}} p(x,y) \\
&= -S(X,Y) - \sum_{x\in\mathcal{X}} \log_2 p(x)p(x) - \sum_{y\in\mathcal{Y}}\log_2 p(y)p(y) \\
&= -S(X,Y) + S(X) + S(Y).
\end{aligned}$$

El caso continuo es totalmente análogo.

Ejercicio 4. Probar las siguientes propiedades para dos variables aleatorias X e Y:

(i) $S(X,Y) = S(X) + S(Y|X) = S(Y) + S(X|Y)$.

(ii) $MI(X,Y) = S(X) - S(X|Y) = S(Y) - S(Y|X)$.

(iii) $MI(X,X) = S(X)$.

Solución. A partir de ahora abreviaremos la notación de los subíndices en las sumatorias. En primer lugar comprobaremos (i), usando repetidamente que $p(x,y) = p(y|x)p(x)$. Se tiene que:

$$\begin{aligned}
S(X,Y) &= -\sum_{xy} p(x,y)\log_2 p(x,y) = -\sum_{xy} p(x,y)\log_2(p(y|x)p(x)) \\
&= -\sum_{xy} p(x,y)\log_2 p(x) - \sum_{xy} p(y|x)p(x)\log_2 p(y|x) \\
&= -\sum_{x}\log_2 p(x)p(x) - \sum_{x} p(x)\sum_{y} p(y|x)\log_2 p(y|x) \\
&= S(X) + S(Y|X).
\end{aligned}$$

Para probar la otra igualdad hacemos lo análogo pero esta vez descomponiendo $p(x,y) = p(x|y)p(y)$. Para ver (ii) observamos que:

$$
\begin{aligned}
MI(X,Y) &= \sum_{xy} p(x,y)\log_2\left(\frac{p(x|y)}{p(x)}\right)\\
&= \sum_{xy} p(x,y)(\log_2 p(x|y) - \log_2 p(x))\\
&= \sum_{xy} p(x|y)p(y)\log_2 p(x|y) - \sum_{x}\log_2 p(x)p(x)\\
&= -S(X|Y) + S(X).
\end{aligned}
$$

La otra igualdad se obtiene intercambiando los papeles de X e Y en la igualdad ya probada, recordando que la información mutua es simétrica. Por último, (iii) es una simple comprobación, teniendo en cuenta que $p(x|x) = 1$ para todo $x \in \mathcal{X}$. En efecto, utilizando (ii) tenemos que:

$$
\begin{aligned}
MI(X,X) = S(X) - S(X|Y) &= S(X) + \sum_x p(x)\sum_x p(x|x)\log_2 p(x|x)\\
&= S(X) + \sum_x p(x)\cdot 0 = S(X).
\end{aligned}
$$

Todas las propiedades para el caso continuo se demuestran de forma análoga.

Ejercicio 5. En el contexto de *temporal coding*, probar que cuando $r\Delta t \ll 1$, entonces la entropía viene dada por

$$S \approx Tr\log_2\left(\frac{e}{r\Delta t}\right).$$

Solución. En primer lugar establecemos el marco de trabajo. Supondremos que tenemos un tren de spikes de longitud fija T, dividido en *bins* o ventanas de tamaño Δt, lo que significa que tendremos en total $M = T/\Delta t$ ventanas. Codificaremos la presencia de *spikes* con un 1 y su ausencia con un 0, y supondremos que en cada ventana puede haber como mucho un *spike*. Así, las observaciones de las que disponemos son cadenas binarias de longitud M.

Definimos además el concepto de *firing rate* como la frecuencia de *spikes* observados en la muestra completa, es decir,

$$r = \frac{\#\text{spikes}}{T}.$$

De esta forma, la probabilidad de ocurrencia de un *spike* en una ventana puede expresarse como su frecuencia de ocurrencia, es decir,

$$p = \frac{\#\text{spikes}}{M} = \frac{\#\text{spikes}}{T/\Delta t} = r\Delta t.$$

En lo sucesivo, supondremos para que las aproximaciones que hacemos tengan sentido que la probabilidad de ocurrencia de un *spike* en una ventana es muy baja (cosa que se corresponde con el comportamiento real de las neuronas), es decir, que $r\Delta t \ll 1$. Esto podemos conseguirlo en la práctica tomando un tamaño suficientemente pequeño de ventana.

Para estimar la entropía, supondremos que la cadena observada ha sido extraída aleatoriamente y de manera uniforme de entre todas los posibles cadenas con el mismo número de 1s y 0s. Dicho de otra forma, todos los eventos considerados son equiprobables con probabilidad dada por el inverso del número de tales cadenas binarias distintas posibles, es decir,

$$p(x_i) = \frac{1}{\binom{M}{Tr}}, \quad i = 1,\ldots,\binom{M}{Tr}.$$

En el caso de sucesos equiprobables sabemos que la entropía viene dada por el logaritmo binario del número total de eventos. Si llamamos por comodidad $M_1 = Tr$ al número de 1s y $M_0 = M - M_1$ al número de 0s, se tiene que

$$S = \log_2 \binom{M}{M_1} = \log_2 \left(\frac{M!}{M_0! M_1!} \right) = \frac{1}{\log 2} (\log M! - \log M_0! - \log M_1!). \tag{1}$$

Si suponemos que las cadenas observadas son suficientemente largas, podemos usar la fórmula de Stirling para aproximar $\log x!$, que toma la forma

$$\log x! = x(\log x - 1) + \mathcal{O}(\log x),$$

donde el resto es despreciable cuando $x \to \infty$. Sustituyendo esta aproximación en (1) tenemos:

$$\begin{aligned}
S &\approx \frac{1}{\log 2}(M(\log M - 1) - M_0(\log M_0 - 1) - M_1(\log M_1 - 1)) \\
&= \frac{1}{\log 2}(M \log M - M_0 \log M_0 - M_1 \log M_1 + (M_0 + M_1 - M)) \\
[M = M_0 + M_1] \quad &= \frac{1}{\log 2}(M_1(\log M - \log M_1) + M_0(\log M - \log M_0)) \\
&= -\frac{M}{\log 2}((M_1/M)\log(M_1/M) + (M_0/M)\log(M_0/M)) \\
[p = M_1/M] \quad &= -\frac{M}{\log 2}(p \log p + (1-p)\log(1-p)) \\
[M = T/\Delta t\,;\, p = r\Delta t] \quad &= -\frac{T}{\Delta t \log 2}(r\Delta t \log(r\Delta t) + (1 - r\Delta t)\log(1 - r\Delta t)). \qquad (2)
\end{aligned}$$

Ahora podemos considerar el desarrollo en serie de Taylor de la función $\log(1+x)$, es decir, $\log(1+x) = x + \mathcal{O}(x^2)$. Identificando $x = -r\Delta t$ y dado que estamos suponiendo que $r\Delta t$ es pequeño ($\ll 1$), podemos considerar que el resto de Taylor es despreciable. Sustituyendo la aproximación en (2) nos queda:

$$\begin{aligned}
S &\approx -\frac{T}{\Delta t \log 2}(r\Delta t \log(r\Delta t) + (1 - r\Delta t)(-r\Delta t)) \\
&\approx -\frac{T}{\Delta t \log 2}(r\Delta t \log(r\Delta t) - r\Delta t), \qquad (3)
\end{aligned}$$

donde de nuevo despreciamos los términos proporcionales a $(r\Delta t)^2$. Finalmente simplificamos la expresión (3) para obtener lo que queríamos:

$$\begin{aligned}
S &\approx -\frac{T}{\Delta t \log 2} r\Delta t(\log(r\Delta t) - 1) \\
&= -\frac{Tr}{\log 2}(\log(r\Delta t) - \log e) \\
&= \frac{Tr}{\log 2} \log\left(\frac{e}{r\Delta t}\right) \\
&= Tr \log_2 \left(\frac{e}{r\Delta t}\right).
\end{aligned}$$

Ejercicio 6. Demostrar la regla de la cadena para la entropía condicionada con tres variables aleatorias X, Y y Z, es decir,

$$S(X, Y|Z) = S(X|Z) + S(Y|X, Z).$$

Solución. Utilizando la definición de probabilidad condicionada y la regla del producto, podemos escribir:

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

Tomando logaritmos en la expresión anterior tenemos que

$$\log_2 p(x,y|z) = \log_2 p(x|z) + \log_2 p(y|x,z),$$

y tomando ahora esperanzas a ambos lados (sobre la distribución conjunta (X,Y,Z)) se tiene:

$$\mathbb{E}\left[\log_2 p(x,y|z)\right] = \mathbb{E}\left[\log_2 p(x|z)\right] + \mathbb{E}\left[\log_2 p(y|x,z)\right].$$

Atendiendo a la definición de entropía condicionada, la expresión anterior es equivalente a lo que queríamos probar, es decir,

$$S(X,Y|Z) = S(X|Z) + S(Y|X,Z).$$

Solución alternativa. También podemos entender el resultado como una consecuencia directa de la regla de la cadena para dos variables,

$$S(X,Y) = S(X) + S(Y|X).$$

En efecto, esta regla para tres variables nos dice que

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

y sucesivas aplicaciones de la misma nos permiten escribir:

$$\begin{aligned}
S(X,Y,Z) = S(Z) + S(X,Y|Z) \iff \\
S(X,Z) + S(Y|X,Z) = S(Z) + S(X,Y|Z) \iff \\
S(Z) + S(X|Z) + S(Y|X,Z) = S(Z) + S(X,Y|Z) \iff \\
S(X|Z) + S(Y|X,Z) = S(X,Y|Z).
\end{aligned}$$

Notamos que tanto esta demostración como la anterior son válidas también en el caso continuo.