EL4005 Principios de Comunicaciones

Clase No. 11: Modelamiento de Fuentes de Información

Patricio Parada | Néstor Becerra Yoma

Departamento de Ingeniería Eléctrica
Universidad de Chile

18 de Noviembre de 2011

Capítulo 3: Codificación de Fuentes de Información

Contenidos de la Clase (1)

Motivación

- Los sistemas de comunicación/almacenamiento se diseñan para transmitir/almacenar información.
- En cualquiera de estos sistemas existe al menos una fuente de información, la cual es necesario caracterizar para poder diseñar un sistema que la pueda transmitir o almacenar.
- Ejemplos:
  - Transmisión de TV: fuentes de audio y video
  - Transmisión de radio, descargas de podcasts o canciones: fuente de voz y/o música.
  - Email, descargas de pdf's: fuente de texto e imágenes.

La Naturaleza de las Fuentes de Información

- Las fuentes de información pueden ser de origen analógico o digital.
- Si son analógicas y vamos a utilizar un sistema comunicaciones digitales, debe ser digitalizada mediante procesos de **muestreo** y **cuantización**.
- Si son digitales, es habitual tratar de representarlas de manera eficiente (utilizando el menor número de bits posible).
- Este proceso se denomina **compresión** de fuente y es responsable por gran parte de las altas capacidades de almacenamiento que vemos hoy día y de la transmisión de grandes tasas de datos con enlaces de capacidad limitada.

Contenido de Información de una Fuente (1)

- Todos tenemos una idea intuitiva del concepto de información.
- Por ejemplo, si les muestro los últimos titulares de los diarios de hoy, la mayor parte de ustedes los considerarán informativos.
- Sin embargo, si presento este mismo objeto a alguien dentro de 5 años, probablemente no va ser muy informativo, excepto tal vez que la persona se encuentre haciendo investigación histórica.

Contenido de Información de una Fuente (2)

- Si queremos diseñar sistemas objetivos que permitan valorizar el contenido de información de una fuente debemos ser más precisos con nuestra noción de información.
- Hartley, Nyquist y Shannon fueron pioneros en definir medidas cuantitativas de información.
- Intuitivamente, la idea de información se refiere al conocimiento nuevo que se adquiere relativo a un objeto.

Modelamiento de Fuentes de Información (1)

- Una fuente de información produce salidas que el sistema no conoce en forma anticipada.
- Queremos construir un sistema que nos permita recuperar en el receptor la información transmitida, con la menor cantidad de errores posible.
- La naturaleza aleatoria de una fuente nos sugiere que la mejor forma de modelar una fuente de información es mediante un **proceso aleatorio**.

Modelamiento de Fuentes de Información (2)

- Ejemplos:
  - Señales de voz se modelan como un proceso aleatorio que ocupa la banda entre los 300 y $4{,}000$ Hz.
  - Las señales de video depende la resolución empleada. En el caso de transmisión estándar, la banda utilizada varía entre los 0 y 6 MHz.
- En general, las fuentes de información son procesos de banda limitada y por ello, pueden ser muestreados a la frecuencia de Nyquist o superior, y reconstruirla en forma exitosa.
- Por ello, podemos utilizar un proceso aleatorio de tiempo discreto como modelo de una fuente.

Modelo Matemático (1)

- *Definición*: Una fuente discreta de información es un proceso aleatorio de tiempo discreto

$$\{X_i\}_{i=-\infty}^{\infty}$$

que toma valores en un **alfabeto** $\mathcal{X}$, el que a su vez puede ser continuo o discreto.

- *Definición*: Una **fuente de información discreta sin memoria** es un proceso $\{X_i\}_{i=-\infty}^{\infty}$ que toma valores en un alfabeto discreto finito y donde cada símbolo es producido en forma independiente e igualmente distribuida.

Modelo Matemático (2)

- Esto es, el conjunto de valores queda caracterizado por

$$\text{Alfabeto:} \qquad \mathcal{A} = \{a_1, a_2, \ldots, a_N\} \tag{1a}$$

$$\text{Probabilidad:} \qquad p_i = \text{Pr}\{X = a_i\}. \tag{1b}$$

  con $i = 1, \ldots, N$.

- La medida de probabilidad de la fuente puede ser agrupada en un vector de largo N, que vive en un espacio $N-1$ dimensional llamado **Simplex de Probabilidad** $\mathscr{P}^N$ definido como sigue:

$$\mathscr{P}^N = \left\{(p_1, \ldots, p_N) \in \mathbb{R}^N : p_i \geq 0 \text{ y } \sum_{i=1}^{N} p_i = 1\right\} \tag{2}$$

Modelo Matemático (3)

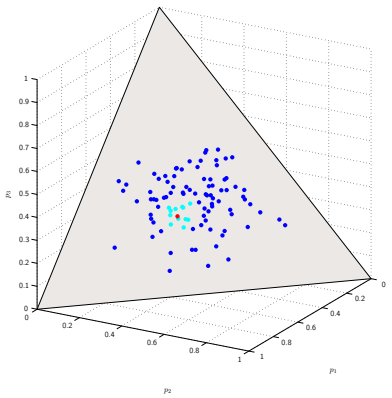

Figura 1. Simplex $\mathscr{P}^3$.

- El conjunto $\mathscr{P}^N$ es convexo y contiene todas las posibles fuentes de información que se pueden definir sobre un alfabeto de N símbolos.
- Los vértices de la región corresponden a las fuentes deterministas con

$$p_1 = 1, p_2 = \ldots = p_N = 0$$

$$p_1 = 0, p_2 = 1, p_3 = 0, \ldots = p_N = 0$$

$$\vdots$$

$$p_1 = p_2 = \ldots, p_{N-1} = 0, p_N = 1.$$

Entropía de la Fuente

- Existen varias formas de medir el contenido de información de una fuente de información.
- Intuitivamente, este número debería estar asociado al número promedio de bits necesarios para representar los símbolos de una fuente.
- Por ejemplo, el alfabeto ASCII contiene 256 símbolos, por lo que se representa mediante 8 bits.
- Sin embargo, si consideramos el texto de una novela, veremos que no todos los caracteres aparecen con la misma frecuencia.

Ejemplo

- Si construimos una tabla de frecuencias de aparición de símbolos utilizando el primer capítulo de El Quijote de la Mancha:

  ```
  En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no
     ha mucho tiempo que vivía un hidalgo de los de lanza en ...
  ```

  veremos que sólo 69 de los 256 símbolos aparecen.

- El símbolo más frecuente es " ", con una probabilidad igual al $18{,}17\,\%$, seguido de la letra "e" con un $10{,}33\,\%$, la letra "a" con un $9{,}8\,\%$ y así sucesivamente hasta llegar a la "ü" con un $0{,}01\,\%$.

Tabla de Frecuencias de Aparición de Símbolos ASCII

Símbolo	%	Símbolo	%	Símbolo	%	Símbolo	%
“ ”	18.17	“i”	3.54	“í”	0.89	“;”	0.18
“e”	10.33	“c”	2.97	“g”	0.77	“ñ”	0.17
“a”	9.80	“t”	2.59	“v”	0.66	“:”	0.10
“o”	7.01	“m”	2.23	“ó”	0.65	“A”	0.10
“s”	5.50	“,”	1.96	“f”	0.44	“Q”	0.09
“n”	5.27	“b”	1.68	“j”	0.34	“M”	0.08
“r”	4.94	“p”	1.47	“z”	0.33	“D”	0.07
“l”	4.80	“q”	1.27	“á”	0.30	“ú”	0.07
“d”	4.09	“y”	1.21	“.”	0.28	“C”	0.06
”u”	3.78	“h”	0.90	“é”	0.28	“E”	0.06

Derivando una Medida de Información (1)

- Cabe preguntarse entonces, necesitamos realmente 8 bits para representar los símbolos de este texto?
- Notemos que aquellos símbolos más frecuentes como " " o la letra "e" deberían ser representados con menos bits, mientras que la letra "A" o "E" debería utilizar más.
- Vamos a considerar

$$-\log(p_i) \tag{3}$$

como una medida del grado de incertidumbre relacionada a un símbolo.

Derivando una Medida de Información (2)

- En el ejemplo

$$p_e = 0{,}1033 \Rightarrow -\log_2 p_e = 2{,}46 \text{ bits}$$
$$p_E = 0{,}0006 \Rightarrow -\log_2 p_E = 10{,}75 \text{ bits}$$

- Luego, el contenido promedio de información de la fuente es

$$-\sum_{i=1}^{N} p_i \log_2 p_i \text{ bits} \tag{4}$$

- En el caso del texto analizado este número es igual a $4{,}26$ bits, que es casi la mitad de los bits utilizados por la codificación ASCII.

Entropía de una Fuente

- Si es una fuente discreta y sin memoria, entonces el contenido de información de una fuente X es igual a

$$H(\mathbf{p}) = -\sum_{i=1}^{p} p_i \log(p_i) \tag{5}$$

  Esta cantidad recibe el nombre de **entropía** de la fuente.

- Si la base del logaritmo es 2 la información se mide en bits.
- Si la base del logaritmo es e la información se mide en nats.

Entropía Binaria (1)

- Consideremos una fuente discreta definida sobre $\mathcal{A} = \{0, 1\}$ y donde

$$\Pr(X = 0) = p \text{ y } \Pr(X = 1) = 1 - p.$$

Si $p = 0{,}5$ la fuente recibe el nombre de **fuente binaria simétrica**.

- La entropía de una fuente binaria es igual a

$$H(p) = -p \log p + (1 - p) \log(1 - p), \ p \in [0, 1] \tag{6}$$

Notar que si $p = 0$ o $p = 1$, por L'Hopital se puede mostrar que $H(0) = H(1) = 0$.

Entropía Binaria (2)

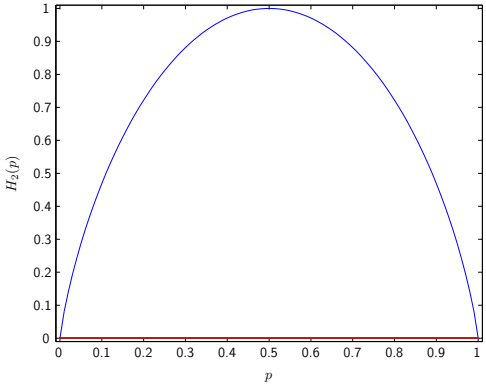

Figura 2. $H_2(p)$.

Propiedades de la Entropía

La entropía satisface 4 propiedades que son relevantes para el análisis y la aproximación empírica de esta cantidad.

1. La entropía es una función continua de $\mathbf{p} = (p_1, \ldots, p_N)$.
2. $H(\mathbf{p}) \geq 0$ y $H(\mathbf{p}) = 0$ si y sólo si $p_n = 0$ para todo n excepto por un símbolo.
3. Dado N, $H(\mathbf{p}) \leq \log N$ y son iguales si y sólo si $p_n = 1/N$.
4. $H(\mathbf{p})$ es concava en $\mathbf{p}$.

Interpretando la Entropía (1)

- La entropía es una cantidad que tiene varias interpretaciones.
- La primera es como contenido de información de una fuente, en el sentido del número de símbolos requeridos para representarla.
- También se entiende como el grado de incertidumbre que uno tiene respecto de la fuente. A mayor entropía mayor incertidumbre existe sobre los símbolos que va a producir.

Interpretando la Entropía (2)

- El nombre entropía se inspira en el concepto de entropía de la teoría de termodinámica (denotada como $S(p)$ por los físicos).
- Finalmente un tema notacional.
  - La entropía es una función de la distribución de probabilidad de una variable aleatoria (o proceso en general).
  - Sin embargo, es habitual denotarla como $H(X)$ en lugar de $H(\mathbf{p}_X)$ o simplemente $H(\mathbf{p})$.
  - Nosotros vamos a utilizar ambas notaciones en lo que sigue del curso, por lo que ambas formas deben ser interpretadas como el mismo objeto.

Distancia entre Fuentes

- Consideremos ahora dos fuentes de información X e Y definidas sobre el mismo alfabeto $\mathcal{A} = \{a_1, \ldots, a_N\}$.
- Supongamos que existen dos distribuciones $\mathbf{p} = (p_1, \ldots, p_N)$ y $\mathbf{q} = (q_1, \ldots, q_N)$ que caracterizan a X e Y respectivamente.
- Podemos medir la "distancia" entre estas dos fuentes mediante el **discriminante de Kullback** que definimos como

$$D(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}. \tag{7}$$

Distancia entre Fuentes

- Notemos que en realidad $D(\mathbf{p}||\mathbf{q})$ no define una métrica propiamente tal, ya que no es simétrica ni satisface la desigualdad triangular, pero si satisface que

$$D(\mathbf{p}||\mathbf{q}) = 0 \text{ si y sólo si } \mathbf{p} = \mathbf{q}$$

- Sin embargo, este funcional es representa con buen nivel de fidelidad la "similitud" entre dos distribuciones dadas y tiene usos que van desde análisis estadístico, selección de modelos, etc.

Entropía Conjunta

- De la misma forma que una distribución conjunta contiene toda la información estadística respecto de las relaciones entre un conjunto de variables, la entropía conjunta mide el contenido de información en un conjunto de fuentes que se encuentran eventualmente ligadas.
- La entropía conjunta entre dos variables aleatorias X e Y es igual a

$$H(p(X,Y)) \equiv H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y). \tag{8}$$

- En general, si $\mathbf{X} = (X_1, \ldots, X_n)$ entonces

$$H(X_1, \ldots, X_n) = -\sum_{x_1,\ldots,x_N} p(x_1, \ldots, x_N) \log p(x_1, \ldots, x_N). \tag{9}$$

Entropía Condicional

- La entropía condicional de una variable aleatoria X dado otra variable aleatoria Y está definida como:

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y) \tag{10}$$

y en general

$$H(X_n|X_{n-1}, \ldots, X_1) = -\sum_{x_1,\ldots,x_n} p(x_1, \ldots, x_n) \log p(x_n|x_{n-1}, \ldots, x_1) \tag{11}$$

Tasa de Entropía

- Cuando una fuente de información tiene memoria (por ejemplo un sensor de temperatura) entonces no podemos hablar de entropía propiamente tal, ya que la distribución de $X_1, \ldots, X_n, \ldots$ va cambiando en el tiempo.
- Para estos casos, tenemos la definición más general dada por:

$$H = \lim_{n \to \infty} H(X_n | X_{n-1}, \ldots, X_1) \tag{12}$$

que llamamos **tasa de entropía** de la fuente.

- Si la fuente es estacionaria, entonces

$$H = \lim_{n \to \infty} \frac{1}{n} H(X_n, X_{n-1}, \ldots, X_1) \tag{13}$$

Resumen

Hemos revisado:

- Motivación para el modelamiento de fuentes.
- Modelo probabilístico de una fuente de información.
- Entropía, discriminante de Kullback, entropía conjunta y condicional.
- Tasa de entropía

Lecturas

- Salehi & Proakis, *Communication System Engineering*, Sección 6.1.