CURS DE MODELS MATEMATICS
La teoria de la communicación de Shannon

Jean-Michel Morel.
morel@cmla.ens-cachan.fr

18 de diciembre de 2011

Índice general

Capítulo 1

Prefacio

Tres tipos de datos numéricos se han hecho omnipresentes en la comunicación humana: los sonidos digitales, las imágenes digitales y los datos alfanuméricos (los textos). Representan tres de cinco vectores principales de la comunicación humana. (Los otros dos son la conversación y la gestual, cuando los interlocutores están en presencia).

Las investigaciones sobre los tres modos digitales de representación de estos datos, que tienen cada uno su estructura propia, estan todavía empezando. No obstante, la teoría de la información y el análisis de Fourier permitieron delimitar un campo científico suficiente para que se pueda, entre otras cosas, tratar muy rigurosamente uno de los problemas más cruciales: la compresión de los datos transmitidos bajo una cualquiera de estas tres formas.

El curso irá de la teoría a la práctica más comn. Trataremos detalladamente de la teoría de la información y de la comunicación de Shannon. LLegaremos a describir los formatos más corrientes de compresión en informática y comunicación electronica, para textos y imágenes.

La teoría de la comunicación de Shannon, publicada en 1948, rebautizada luego teoría de la información, es el texto fundador, que analizaremos detalladamente. Esta teoría propone una cadena completa para codificar, comprimir y transmitir cualquier mensaje entre humanos, y particularmente las señales (voces e imágenes). La teoría matemática subyacente es la teoría de las probabilidades discretas a la cual se añade una noción nacida de la termodinámica de gases, la noción de entropía.

La teoría de Shannon, define la cantidad de información contenida en un mensaje y trata de su compresión óptima. La noción central es la de la entropía de un mensaje. Las codificaciones más usadas, Shannon, Huffman, Lempel-Ziv, serán explicadas, demostradas matemáticamente, y experimentadas.

Experimentos sobre textos serán llevados a cabo para verificar la validez de los algoritmos de compresión, y probaremos algoritmos de síntesis automática de texto y de comparación entropica de textos variados.

Capítulo 2

Modelos discretos de probabilidad

2.1. Curso 1: probabilidad discreta

Se parte de un conjunto de acontecimientos elementales o atómicos Ω, por ejemplo los resultados de un partido de fútbol, $\Omega = \{(0,0),(0,1),(1,0),\dots,\ (40,40)\}$. Mas generalmente $\Omega = \mathbb{N}\times\mathbb{N}$. Un acontecimiento en general es un subconjunto de Ω. Por ejemplo $A = \{(m,n) \mid m > n\}$ caracteriza el hecho de que el equipo 1 gana, y $B = \{(m,n) \mid m = n\}$ caracteriza el empate. Si $\omega = (m,n)$ está en A, se escribe que "ω es una realización de A."

Definición 2.1 *Algebra de conjuntos "interesantes": Es un conjunto $\mathcal{A}$ de conjuntos de Ω que satisface los axiomas siguientes:*

- $\emptyset, \Omega \in \mathcal{A}$
- $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
- $A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$ *(y luego también $A \cup B \in \mathcal{A}$).*

Definición 2.2 *Sea Ω un conjunto y $\mathcal{A}$ un algebra de conjuntos de Ω. Llamamos variable aleatoria discreta una aplicación $X : \Omega \to E$ donde E es un conjunto finito o enumerable y tal que los conjuntos $\{\omega \in \Omega \mid X(\omega) = e\}$ esten todos en $\mathcal{A}$. En el caso en que $E \subset \mathbb{R}$, se habla de variable aleatoria real.*

En el ejemplo del fútbol, $M(\omega) = M((m,n)) := m$ es una variable aleatoria real que puede llamarse "numero de goles del equipo 1". Miremos un ejemplo de Ω mas general y un gran clásico de teoria de probabilidades, el juego de cara o cruz. Se codifica cara por 0 y cruz por 1. Una serie infinita de jugadas da lugar al conjunto de acontecimientos $\Omega := \{0,1\}^{\mathbb{N}}$. Cada elemento de Ω se escribe $\omega = (\omega(1),\dots,\omega(n),\dots)$ y la aplicación $X_n : \omega \to \omega(n)$ es una variable aleatoria que se interpreta como el "resultado de la n-esima jugada". También podemos considerar el conjunto de las N primeras jugadas, $\Omega_N := \{0,1\}^N$. De una cierta manera Ω_N está contenido en Ω pero para formalizarlo hay que asociar a cada elemento $\omega_N = (\omega(1),\dots,\omega(N)) \in \Omega_N$ el conjunto de todas las sucesiones que empiezan por ω_N, que llamaremos $\Omega(\omega_N)$. Podemos considerar el álgebra generada por los $\Omega(\omega_N)$ cuando ω_N varia en Ω_N y N varia en $\mathbb{N}$. Se trata, por definición, del álgebra mas pequeña que contiene todos esos acontecimientos. Este álgebra se llama "álgebra de acontecimientos en tiempo finito".

Ejercicio 2.1 Probar que cualquier elemento del álgebra $\mathcal{A}$ de los acontecimientos en tiempo finito es una unión finita de acontecimientos de tipo $\Omega(\omega_N)$. Para probarlo, basta llamar $\tilde{\mathcal{A}}$ ese conjunto de uniones finitas y probar que tiene la estructura de una álgebra. Si asi es, es por fuerza el álgebra mas pequeña que contiene los $\Omega(\omega_N)$.

De hecho el álgebra de los acontecimientos en tiempo finito no nos dice nada sobre lo que pasa al infinito. Por ejemplo el conjunto : $A := \{\omega \in \Omega \mid \lim_n X_n(\omega) = 0\}$ no está en $\mathcal{A}$. Para verlo, basta notar que si estuviera en $\mathcal{A}$, tendriamos por el resultado del ejercio 2.1 que $A = \cup_{n\in I}\Omega(\omega_n)$ donde I es un subconjunto finito de $\mathbb{N}$ y $\omega_n \in \Omega_{N(n)}$. Llamemos k el indice mayor que hay en I, $k = \max_{n\in I} N(n)$. Entonces es inmediato verificar que si $\omega \in A$ y si consideramos cualquier otro elemento $\omega' \in \Omega$ tal que $\omega'(i) = \omega(i)$ por $i \leq k$, entonces ω' está en A. Luego podemos imponer que $\omega'(n) = 1$ por $n \geq k$ y eso implica que ω' no está en A, una contradiccion. Por eso Kolmogorov extendió la noción de álgebra a la de $\sigma-$álgebra, lo que permite considerar acontecimientos como A.

Definición 2.3 *Una σ-álgebra $\mathcal{F}$ de Ω es una álgebra tal que si A_n está en $\mathcal{F}$, entonces $\cup_n A_n$ está tambíen en $\mathcal{F}$. Dado un conjunto $\mathcal{A}$ de partes de Ω, se llama σ-álgebra engendrada por $\mathcal{A}$ y se escribe $\sigma(\mathcal{A})$ la intersección de todas las σ-álgebras que contengan $\mathcal{A}$.*

Tal intersección existe porque hay al menos una σ-álgebra que contiene $\mathcal{A}$: el conjunto $\mathcal{P}(\Omega)$ de todas las partes de Ω es en efecto una σ-álgebra.

Ejercicio 2.2 Demostrar que el conjunto $A := \{\omega \in \Omega \mid \lim_n X_n(\omega) = 0\}$ está en $\sigma(\mathcal{A})$, donde $\mathcal{A}$ es el álgebra de acontecimientos en tiempo finito. Indicación: probar que

$$A = \bigcup_{n\geq 1} \bigcap_{m\geq n} \{\omega \mid X_m(\omega) = 0\}.$$

En la práctica se empieza conociendo el valor de la probabilidad de algunos acontecimientos. Luego, se deduce la probabilidad de los acontecimientos del álgebra $\mathcal{A}$ generada por esos acontecimientos y finalmente la probabilidad de los acontecimientos de $\sigma(\mathcal{A})$. Las reglas para deducir probabilidades unas de otras son:

Definición 2.4 *Sea Ω un espacio de probabilidad con una σ-álgebra $\mathcal{F}$. Se dice que $\mathbb{P}$ es una probabilidad sobre $(\Omega, \mathcal{F})$ si para todo A en $\mathcal{F}$ y para toda sucesión disjunta A_n en $\mathcal{F}$,*

- $0 \leq \mathbb{P}(A) \leq 1$, $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$.

La ultima propriedad se llama "σ-aditividad" de la probabilidad.

Ejercicio 2.3 Deducir las consecuencias siguientes:

- si $A \subset B$, entonces $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- Si $A_n \in \mathcal{F}$, $\mathbb{P}(\cup_n A_n) \leq \sum_n \mathbb{P}(A_n)$.

2.2. Probabilidad condicional

En realidad muy a menudo las probabilidades a las que se tiene acceso son probabilidades condicionales.

Definición 2.5 *Dado un espacio de probabilidad* $(\Omega, \mathcal{F}, \mathbb{P})$ *y* $A, B \in \mathcal{F}$*, se llama probabilidad condicional de* A *sabiendo que* B

$$P(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ si } \mathbb{P}(B) \neq 0, = 0 \text{ si } \mathbb{P}(B) = 0.$$

Ejercicio 2.4 Probar que para todo B en $\mathcal{F}$, si $\mathbb{P}(B) \neq 0$, la aplicación $A \to \mathbb{P}(A \mid B)$ es una probabilidad sobre $(\Omega, \mathcal{F})$. Probar igualmente la "regla de las causas totales" : si los B_i, $i \in I\!N$ son acontecimientos formando una partición de Ω, entonces

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

2.2.1. Independencia

Definición 2.6 *Sea* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espacio de probabilidad.*

A y B son independientes si $\mathbb{P}(A \mid B) = \mathbb{P}(A)$*, lo que es equivalente a* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

- *Una familia* $(A_i)_{i \in I}$ *es una familia de acontecimientos independientes si para toda subfamilia finita* $A_{i_1}, \ldots, A_{i_n}$, $\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \ldots \mathbb{P}(A_{i_n})$.
- *Una sucesión* $(X_n)_{n \geq 1}$ *de variables aleatorias discretas* $X_n : \Omega \to E$ *es independiente si para todo* $(e_1, \ldots, e_n, \ldots) \in E^{I\!N}$*, los acontecimientos* $(X_n = e_n)$ *son independientes.*
- *Una sucesión* $(X_n)_{n \geq 1}$ *de variables aleatorias reales* $X_n : \Omega \to \mathbb{R}$ *es independiente si para toda sucesión de intervalos de* $\mathbb{R}$*,* $(I_1, \ldots, I_n, \ldots)$ *los acontecimientos* $(X_n \in I_n)$ *son independientes.*

Ejercicio 2.5 Dos ejemplos importantes:

- Sea $\Omega = [0,1]^N$, $\mathbb{P}(A) = \text{volumen}(A)$ por $A \subset \Omega$. Demostrar que las coordenadas $X_i : \omega = (\omega_1, \ldots, \omega_n) \in \Omega \to \omega_i$ son variables aleatorias independientes.
- Si en cambio $\Omega = B(0,1)$ es el disco de centro 0 y de radio 1 y $\mathbb{P}$ es la medida de Lebesgue multiplicada por un factor λ tal que $\mathbb{P}(B(0,1)) = 1$, probar que las variables X_i no son independientes.

2.2.2. Modelos, ejemplos

Problema de los tres presos

Tres presos A, B y C estan incomunicados en la carcel de un regimen totalitario. Saben que dos de ellos estan condenados a muerte por no saben cuáles. Por supuesto el carcelero no tiene derecho a decirles nada. Sin embargo uno de los presos hace el razonamiento siguiente

al carcelero: "Yo ya sé que al menos uno de B y C est condenado. De modo que no me darás ninguna información útil sobre mi suerte si me dices que B o C está condenado. Por favor, dame el nombre de uno de los condenados entre B y C." El carcelero se lo piensa un momento y contesta : "-tienes razón, te voy a decir que B está condenado. Pero no creas que con eso puedes ganar alguna tranquilidad." El preso contesta "- todo lo contrario ; antes tenia una probabilidad de $2/3$ de estar condenado ; ahora todo se juega entre mi y C. Luego mi probabilidad de morir es $1/2$ y no dos tercios. Y bien sabes, añade, que $1/2$ es inferior a $2/3$!". El carcelero no se inmuta, sonrie y se aleja.

Quien tiene razón del preso A o del carcelero? Habra ganado un poco de tranquilidad A con razón, o será víctima de una ilusión?

Solución

Para formalizar ese tipo de problema, hay que buscar los acontecimientos atómicos, o sea enumerar cada sucesión de posibilidades y luego intentar buscar su probabilidad. Una vez calculadas las probabilidades de esos acontecimientos atómicos, cualquier otra probabilidad se vuelve la probabilidad de un acontecimiento que es una unión de atómicos. Luego sera una suma de probabilidades, fàcil de calcular. Segun esa descripción el modo de proceder es:

- enumerar y nombrar todos los acontecimientos distintos (en general sucesiones de eventos);
- dar un nombre a las variables aleatorias de interés;
- expresar las probabilidades indicadas por el problema: muy a menudo veremos que son probabilidades condicionales;
- tomar en cuenta todas las independencias implícitas en el texto;
- formalizar el acontecimiento cuya probabilidad queremos calcular y expresarlo en términos de acontecimientos elementales;
- finalmente calcular la probabilidad .

Nuestro problema muestra dos sucesos aleatorios consecutivos: Primero la elección del par de presos condenados: AB, BC o AC (para mayor claridad los ordenamos A, B, C). Segundo la elección por el carcelero de B o de C en el caso de que el tenga que elegir. También hay que tomar en cuenta un gran principio de la probabilidad que a veces se llama *principio de probabilidad subjectiva*, segun el cuál cuando de varias posibilidades 1, 2, ... n no se sabe nada, entonces se atribuye la misma probabilidad a cada una, o sea $1/n$. En nuestro caso la probabilidad que AB, o BC, o AC esten condenados tiene que fijarse en un tercio. De la misma manera si B, C estan condenados el carcelero tiene que elegir a quién nombrar de B o de C. Como A no sabe que criterio adopta el carcelero, el tiene que contar con que las probabilidades de que el carcelero nombre a B o a C sean iguales a $1/2$. Los acontecimientos atómicos son ABC, ACC, BCB y BCC, donde las dos primeras letras indican a los condenados y la ultima es el condenado elegido por el carcelero. Las variables aleatorias naturales son X, Y, Z, donde XY es la lista ordenada de condenados y Z el condenado indicado por el carcelero. Por ejemplo $XY(ABC) = AB$ y $Z(ABC) = C$. Ahora bien, podemos expresar todas las probabilidades que conocemos:

- $\mathbb{P}(XY = AB) = \mathbb{P}(XY = BC) = \mathbb{P}(XY = AC) = \frac{1}{3}$
- $\mathbb{P}(Z = C \mid XY = BC) = \mathbb{P}(Z = B \mid XY = BC) = \frac{1}{2}$.

Es notable que dos de las probabilidades a las que tenemos acceso son probabilidades condicionales. Eso es del todo natural porque corresponde a preguntas del tipo : "cuál es la probabilidad de tal acontecimiento si tal otro sucede?" Ahora podemos traducir la cuestión que se plantea el preso : "cuál es mi probabilidad de no morir sabiendo que el carcelero me ha indicado que B está condenado?" Otra vez es una probabilidad conditional ; tenemos que calcular:

$$p = \mathbb{P}(XY = BC \mid Z = B)$$

Por definición de la probabilidad condicional,

$$p = \frac{\mathbb{P}(XY = BC \ \& \ Z = B)}{\mathbb{P}(Z = B)}.$$

Pero utilizando lo que sabemos y de nuevo la probabilidad condicional.

$$\mathbb{P}((XY = BC) \ \& \ (Z = B)) = \mathbb{P}(Z = B \mid XY = BC)\mathbb{P}(XY = BC) = \frac{1}{2}\frac{1}{3} = \frac{1}{6}.$$

Para calcular $\mathbb{P}(Z = B)$ podemos utilizar la *regla de las causas totales*:

$$\begin{array}{rcl} \mathbb{P}(Z{=}B) & = & \mathbb{P}(Z{=}B|XY{=}AB)\mathbb{P}(XY{=}AB) + \mathbb{P}(Z{=}B|XY{=}AC)\mathbb{P}(XY{=}AC) \\ & & \quad +\mathbb{P}(Z{=}B|XY{=}BC)\mathbb{P}(XY{=}BC) \\ & = & \frac{1}{3} + 0 + \frac{1}{2}.\frac{1}{3}{=}\frac{1}{2}. \\ & = & \end{array}$$

Finalmente obtenemos

$$p = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

La estimación de su propia probabilidad de supervivencia por A no cambia a consecuencia de la información dada por el carcelero.

Ejercicio 2.6 La conclusión anterior depende estrictamente de la hipótesis de que el preso no conoce nada sobre el modo de selección por el carcelero entre B y C. Supongamos que el preso intuye que el carcelero preferira nombrar a B si B y C son condenados. Entonces $q = \mathbb{P}(Z = B \mid XY = BC) > \frac{1}{2}$. Volver a hacer las cuentas precedentes y demostrar que el preso gana información grácias a las respuesta del carcelero.

2.3. Ejercicios

Ejercicio 2.7 Un modelo discreto para ganar a un juego muy sencillo. Ahora el juego es menos siniestro. Se trata de saber ganar a un juego muy sencillo. Tenemos tres cartas. Una tiene dos caras rojas y la llamamos RR. Otra tiene dos caras verdes (VV) y la tercera tiene un cara roja y otra verde (RV). Se tira una carta al azar y se tira tambíen al azar el lado

expuesto de la carta. Los jugadores observan esa cara expuesta y tienen que hacer apuestas sobre el color de su otra cara.

Supongamos por ejemplo que la cara expuesta sea roja. Cual es la probabilidad de que la otra cara sea roja? Y verde? Supongamos que cada jugador apueste por un color. Si el otro jugador pone 100 euros en la mesa apostando por rojo cuanto tengo que poner en la mesa, apostando por verde, para que mi esperanza de ganancia sea positiva?

Ejercicio 2.8 Sea $\mathcal{A}_n$ una sucesión creciente ($\mathcal{A}_n \subset \mathcal{A}_{n+1}$) de álgebras de Ω. Demostrar que $\bigcup_n \mathcal{A}_n$ es un álgebra.

Ejercicio 2.9 Sea A_i, $i \in I\!N$ una partición de Ω. Describir el álgebra generada por los A_i y la σ-álgebra generada por los A_i.

Ejercicio 2.10 Probar que $\mathbb{P}(A \mid B\&C)\mathbb{P}(B \mid C) = \mathbb{P}(A\&B \mid C)$.

Ejercicio 2.11 Si $A_1, \dots A_n$ son independientes entonces $\tilde{A}_i$ son independientes, donde $\tilde{A}_i$ puede ser arbitrariamente A_i o A_i^c.

Ejercicio 2.12 Sea un espacio Ω con cuatro elementos $\{\omega_1, \omega_2, \omega_3, \omega_4)$. Sea $\mathbb{P}$ la probabilidad definida por $\mathbb{P}(\omega_i) = \frac{1}{4}$, $i = 1, \dots, 4$. Consideremos los acontecimientos $A = \{\omega_1, \omega_2\}$, $B = \{\omega_2, \omega_3\}$, $C = \{\omega_1, \omega_3\}$. Comprobar que A y B son independientes, B y C también, C y A también, pero que A, B, C no son independientes.

Ejercicio 2.13 Tres turistas disparan al mismo tiempo hacia un elefante. El animal muere herido por dos balas. El valor de cada cazador se estima en la probabilidad de que alcanze su meta. Esas probabilidades son 1/4, 1/2, 3/4. Calcular para cada uno de los cazadores su probabilidad de haber fallado con el elefante. (Esa probabilidad depende del acontecimiento observado: si el elefante hubiera recibido tres balas sabriamos por ejemplo que la probabilidad de haber acertado es uno para cada cazador).

Ejercicio 2.14 El profesor Peplluis Serra Balaguer viaja de Nueva York a Palma pasando por Madrid y Barcelona. La probabilidad de que la maleta se quede atrás es identica en cada uno de esos aeropuertos y igual a p. Cuando llega a Palma, el profesor comprueba que su maleta ha desaparecido. Cuales son las probabilidades de que la maleta se haya quedado en Nueva York, en Madrid, y en Barcelona?

Capítulo 3

Distribuciones discretas de probabilidad

Definición 3.1 *Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, $\mathcal{X}$ un conjunto finito o enumerable y $X : \Omega \to \mathcal{X}$ tal que los conjuntos $\{\omega \in \Omega \mid X(\omega) = x\}$ esten todos en la tribu $\mathcal{F}$. Entonces se dice que X es una variable aleatoria sobre $(\Omega, \mathcal{F}, \mathbb{P})$.*

Por ejemplo si tenemos una sucesion de lanzamientos a cara o cruz, que se formaliza como una sucesión de variables de Bernoulli, X_1, ..., X_n con $X_i = 0$ (cara) o $X_i = 1$ (cruz) y $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$. Si los lanzamientos son independientes, tenemos

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n) \text{ por todo } (x_1, \ldots, x_n) \in \{0, 1\}^n.$$

Entonces si escribimos $X = (X_1, \ldots, X_n)$ y $x = (x_1, \ldots, x_n)$,

$$\mathbb{P}(X = x) = p^{h(x)}(1 - p)^{n - h(x)}, \text{ donde } h(x) := \sum_{i=1}^{n} x_i.$$

La function $h(x)$ se llama "peso de Hamming" de x.

Ejercicio 3.1 Comprobar que $\sum_{x \in \{0,1\}^n} p(x) = \sum_{k=0}^{n} C_n^k p^k (1 - p)^{n-k} = 1$.

Definición 3.2 *Si $\mathcal{X}$ es un conjunto enumerable y $(p(x))$, $x \in \mathcal{X}$ una funcion sobre $\mathcal{X}$ que satisface:*

1. $0 \leq p(x) \leq 1$
2. $\sum_{x \in \mathcal{X}} p(x) = 1$,

se dice que $(p(x))$, $x \in \mathcal{X}$, es una distribución de probabilidad sobre $\mathcal{X}$.
Ejemplo fundamental: Si X es una variable aleatoria definida sobre $(\Omega, \mathcal{F}, \mathbb{P})$ y con valores en $\mathcal{X}$, se considera por todo $A \subset \mathcal{X}$,

$$\mathbb{P}_X(A) := \mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} \mathbb{P}_X(x).$$

Llamando $p(x) = \mathbb{P}_X(x)$, se dice que $(p(x))$, $x \in \mathcal{X}$ se llama la distribución (o ley) de la variable X.

Demos algunos ejemplos de distribuciones clásicas. (Habra más en los ejercicios):

Si $\mathcal{X} := \{0,1\}^n$, entonces $p(x) := p^{h(x)}(1-p)^{n-h(x)}$ se llama distribución de Bernoulli de orden n y parametro p. Es la ley de la variable aleatoria $X := (X_1, \ldots, X_n)$ donde X_i son variables aleatorias de Bernoulli de parametro p y orden 1. Luego se puede mirar la distribución sobre $\{0, 1, \ldots, n)$ de la variable aleatoria $S_n := \sum_{i=1}^n X_i$. Es fácil comprobar que $\mathbb{P}(S_n = k) = C_n^k p^k (1-p)^{n-k} = p_k$. Esa distribución sobre $\{0, 1, \ldots, n\}$ se llama distribución binomial.

3.1. Esperanza y varianza

Sea X una variable aleatoria discreta con valores en $\mathcal{X}$, de distribución $p(x)$, $x \in \mathcal{X}$. Sea $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty, -\infty\}$ una aplicación.

Definición 3.3

- *Si $f \geq 0$, se define $\mathbb{E}[f(X)] := \sum_{x\in\mathcal{X}} \mathbb{P}(X = x) f(x)$. (Esa cantidad puede ser infinita).*
- *Si $\mathbb{E}[|f(X)|] < +\infty$, se define $\mathbb{E}[f(X)] := \sum_{x\in\mathcal{X}} \mathbb{P}(X = x) f(x)$ y se dice que $f(X)$ es integrable.*
- *$\mathbb{E}$ es lineal y monótona ($\forall x\ f(x) \leq g(x) \Rightarrow \mathbb{E}f(X) \leq \mathbb{E}g(X)$.)*

En el caso en que X es ella misma una variable con valores reales, se definen:

- media de X : $m(X) = m_X := \mathbb{E}X$;
- Si X^2 es integrable, varianza de X: $\sigma^2(X) = \sigma_X^2 := \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

$\sigma(X) = \sigma_X$ se llama desviación estandar de X.

Ejercicio 3.2 Comprobar que $\sigma_X^2 = \mathbb{E}[(X - m_X)^2]$. Comprobar que si X es Bernoulli de orden 1 y parametro p, $m_X = p$ y $\sigma_X^2 = p(1-p)$. En el caso de la binomial de parametro p y orden n, comprobar que $m_X = np$ y $\sigma_X^2 = np(1-p)$.

3.2. La convergencia en probabilidad

Empezemos con algunas desigualdades. La desigualdad de Markov dice que si X es una variable aleatoria y f una función medible tal que $f(X)$ sea integrable,

$$\mathbb{P}(|f(X)| \geq a) \leq \frac{\mathbb{E}|f(X)|}{a}.$$

De ahi sale la famosa desigualdad de Tchebychev, que para una variable aleatoria real y para todo $\varepsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{\sigma^2(X)}{\varepsilon^2}.$$

La prueba consiste en aplicar la desigualdad de Markov a $f(X) = |X - \mathbb{E}X|^2$.

Ejercicio 3.3 Probar esas dos desigualdades!

Vamos a aplicar la desigualdad de Tchebychev para probar una ley fundamental, la llamada ley débil de los números grandes.

Teorema 3.1 *Sea una sucesion de variables de Bernoulli independientes X_i de parametro p, $i = 1, \ldots, n, \ldots$ y $S_n := \sum_{i=1}^{n} X_i$. Entonces $\mathbb{P}(|\frac{S_n}{n} - p| \geq \varepsilon) \to 0$ cuando $n \to \infty$.*

En efecto la esperanza (o media) de S_n es np y la varianza de S_n es $np(1-p)$. Luego la media de $\frac{S_n}{n}$ es p y su varianza es $\sigma^2 = \frac{p(1-p)}{n}$. Applicando la desigualdad de Tchebytchev se obtiene

$$\mathbb{P}(|\frac{S_n}{n} - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon} \to 0 \text{ cuando } n \to \infty.$$

Definición 3.4 *Sea Y_n variables aleatorias reales definidas sobre el mismo espacio de probabilidad. Se dice que $Y_n \to Y$ en probabilidad si $\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \varepsilon) = 0$ for all $\varepsilon > 0$.*

La ley débil de los grandes números se extiende fácilmente a una situación lígeramente mas géneral. Consideremos una sucesión de variables aleatorias X_n independientes, equidistribuidas y de varianza acotada $\sigma^2(X)$ y esperanza $\mathbb{E}X$. Entonces con el mismo razonamiento de antes (aditividad de las varianzas y de las esperanzas y desigualdad de Tchebychev), se obtiene para $S_n = \sum_{k=1}^{n} X_k$, Así se has demostrado la ley débil de los números grandes con toda generalidad.

3.3. Ejercicios

Ejercicio 3.4 Un lote de relojes idénticos para turistas llega a un mercante de Palmanova. Ese lote puede provenir tan solo de dos fábricas: una en Singapur u otra en Hong Kong. El mercante sabe que en promedio la fábrica de Singapur produce un porcentaje de relojes defectuosos de 1/200 mientras que la fábrica de Hong Kong tiene un porcentaje de 1/1000. El mercante saca un reloj y comprueba que funciona. Cuál es la probabilidad de que el segundo relój funcione?

Ejercicio 3.5 Se sortean dos puntos en $[0, 1]$, independientemente. El más pequeño de los números obtenidos es mayor que 1/3. Cuál es la probabilidad de que el otro sea superior a 3/4?

Ejercicio 3.6 Por $\lambda > 0$ se define la ley de Poisson sobre $\mathbb{N}$ por

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, \ldots, .$$

Comprobar que $\sum_k p(k) = 1$ y calcular la media y la varianza de esa distribución.

Ejercicio 3.7 Calcular la media y la varianza de una ley binomial de orden n y parametro p.

Ejercicio 3.8 Sean $X_1, \ldots, X_n$ n variables aleatorias discretas independientes, identicamente distribuidas de varianza común σ^2. Calcular la varianza y la desviación tipica de $\frac{X_1+\cdots+X_n}{n}$.

Ejercicio 3.9 Se dice que una variable aleatoria X con valores en $\mathbb{N}$ sigue una ley geométrica de parametro $p \in [0, 1]$ si $\mathbb{P}(T = n) = p(1 - p)^{n-1}$, $n \geq 1$. Calcular la media y la varianza de T. Probar que T "no tiene memoria", o sea que $\mathbb{P}(T \geq n_0 + n \mid T > n_0) = \mathbb{P}(T \geq n)$, $n \geq 1$.

Ejercicio 3.10 Sea X una variable aleatoria con valores en $\mathbb{N}$. Probar que $\mathbb{E}X = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n)$.

Capítulo 4

Códigos prefijos

Consideremos una distribución discreta de probabilidad, $(p_1, \ldots, p_k)$. Se define la entropia de esta distribución por

$$H(p_1, \ldots, p_k) := -\sum_{i=1}^{k} p_i \log p_i.$$

En general el logaritmo tiene base 2, o sea $\log 2 = \log_2 = 1$. La interpretación de esa fórmula por Shannon es la siguiente: en la teoria de la comunicación el receptor ignora lo que el emisor le esta escribiendo, pero ya tiene una idea de la probabilidad de cada mensaje posible. Por ejemplo $p = (p_1, \ldots, p_k)$ puede ser la distribución de probabilidad de las palabras de un diccionario $\mathcal{X} := \{x_1, \ldots, x_k\}$. El grado de comunicación es superior cuando la incertidumbre sobre el mensaje es mayor. Para Shannon, $H(p)$ mide esa incertidumbre. Por ejemplo si $p_1 = 1$ y los demas son nulos, se comprueba que $H(p) = 0$, lo que significa que no hay nada de incertidumbre. Si todos los p_i son iguales a $\frac{1}{k}$, la entropia es $\log_2 k$ y tiene que ser maximal. Veremos que esa intuición es justa. La incertidumbre o entropia (Shannon utiliza esas dos palabras) se mide en *bits*, una abreviación de BInary digiT, término creado por John W. Tukey. Esa unidad fundamental de la informática se define como la cantidad de información recibida por un receptor que está esperando un mensaje 0 o 1 y que atribuye la probabilidad $\frac{1}{2}$ a ambas posibilidades. Efectivamente si $p = (\frac{1}{2}, \frac{1}{2})$, se comprueba inmediatamente que $H(p) = 1$.

Definición 4.1 *Dado un conjunto de mensajes $\mathcal{X} = \{x_1, \ldots, x_k\}$ llamaremos codificación de esos mensajes una aplicación $h : \mathcal{X} \to \{0,1\}^{(\mathbb{N})}$, conjunto de las sucesiones finitas de zeros y unos. Escribiremos $l_i := l(h(x_i))$, la lóngitud del código de x_i.*

El código mas elemental que podamos hacer es enumerar los mensajes de $i = 0$ a k, convirtiendo i en un número binario. Entonces $h(x_1) = 0$, $h(x_2) = 1$, $h(x_3) = 10$, $h(x_4) = 11$, etc. Vemos que con esa codificación la longitud máximal del código es

$$[\log k] \leq l_k := l(h(x_k)) \leq [\log k] + 1.$$

Shannon demuestra de varias maneras que, dada una fuente emisora de entropia p, entonces la longitud minimal media de los mensajes codificados en bits es exactamente $H(p)$. En otros terminos es posible transmitir lo que emite la fuente codificandolos con cifras hechas de zeros y unos, y de longitud media $H(p)$. Para desarollar mejor, necesitamos definir lo que llamaremos *codificación*. Vamos a empezar con una teoria lígeramente mas restringida, la teoria de las codificaciónes *prefijas*.

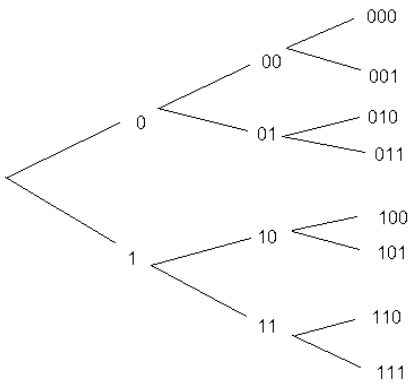

Figura 4.1: Arbol binario completo de las cifras de menos de tres símbolos. Cada nudo del arbol representa un código. Si una codificación es prefija, corresponde a una selección de nudos tales que ningun nudo es un subarbol de otro. Ejemplo: 00, 011, 10, 110 es una codificación prefija.

4.1. Teoria de las codificaciones prefijas

Un problema técnico es que si emitimos mensajes repetidos, no sabemos donde acaba uno y donde empieza el siguiente: tan solo observamos una sucesion de zeros y unos. Si de cualquier sucesión observada $h(x_{i_1}) \dots h(x_{i_n})$ deducimos de manera única $x_{i_1}, \dots, x_{i_n}$, diremos que la codificación es *únicamente descifrable*. Una manera muy sencilla de obtener tales codificaciones es de imponerles la propiedad de prefijo.

Definición 4.2 *Una codificación es llamada prefija si por todo* $i, 1 \leq i \leq k$, *cada cødigo* $h(x_i)$ *no es el prefijo de algun otro prefijo* $h(x_j)$.

Ejercicio 4.1 Demostrar que una codificación prefija es únicamente descifrable.

La teoria de las codificaciones prefijas es maravillosamement sencilla. Vamos a caracterizar todos las codificaciones prefijas y vamos a dar unos ejemplos optimales.

Teorema 4.1 (Desigualdad de Kraft). *Si* h *es una codificación prefija, llamemos* $l_1, \dots, l_i, \dots, l_k$ *las longitudes de los códigos para cada símbolo* x_i. *Si* h *es prefija, entonces*

$$\sum_{i=1}^{k} 2^{-l_i} \leq 1. \tag{4.1}$$

Reciprocamente, sean (l_i), $1 \leq i \leq k$ *enteros positivos tales que* (4.1) *sea cierta. Entonces existe una codificación* h *con la propiedad de prefijo y cuyos códigos tienen las longitudes* (l_i), $1 \leq i \leq k$.

Prueba La prueba se hace en modo muy intuitivo dibujando un arbol binario completo de profundidad n cuyas hojas son los números binarios de n cifras, de $00\ldots00$ a $11\ldots11$ (Confiera la figura 4.1.) La raiz del arbol es la palabra vacia. Sus hijos son 0 y 1. El primer, 0, tiene como hijos a 00 y 01 y el segundo, 1, a 10 y 11, etc. Es fácil enterarse de que en los nudos del arbol hay todos los códigos posibles de longitud inferior o igual a n. Cada nudo del arbol es la raiz de un subarbol. En ese subarbol estan todos códigos que tienen a la raiz como prefijo. De esa observacion inmediata se deduce que *una codificación prefija h es tal que ningun código pertenece al subarbol de otro.* En otros términos, los subarboles de los $h(x_i)$ son disjuntos. Es tambien fácil ver que como $h(x_i)$ tiene longitud l_i, su subarbol tiene profundidad $n - l_i$ y el número de hojas del subarbol de $h(x_i)$ es 2^{n-l_i}. Como esos conjuntos de hojas son todos disjuntos y como el numero total de hojas es 2^n se llega a la desigualdad

$$\sum_i 2^{n-l_i} \leq 2^n$$

que implica la desigualdad de Kraft.

Ahora veamos la recíproca. Si los números $(l_i)_{1\leq i\leq k}$ verifican la desigualdad de Kraft, podremos definir una codificación prefija h tal que $l_i = h(x_i)$? La construccion es afín al razonamiento precedente (vea la figura 4.2). Se considera de nuevo el arbol completo binario de profundidad $n = \text{máx}_i\, l_i$. Se ordenan los l_i por orden creciente, $l_1 \leq \cdots \leq l_i \leq \ldots l_k$ y se consideran las 2^{n-l_1} primeras hojas del arbol : son las hojas de un subarbol cuya raiz tiene longitud l_1: esa raiz se decide que sea el código de x_1. Luego se consideran las 2^{m-l_2} hojas siguientes. De nuevo son las hojas de un subarbol de raiz de longitud l_2 y esa raiz se vuelve el código de x_2. Podemos iterar con tal de que la cantidad de hojas utilizadas no exceda 2^n, pero eso es exactamente lo que nos garantiza la desigualdad de Kraft. La figura 4.2 trata el ejemplo siguiente :

$$l_1 = 2,\ l_2 = l_3 = 3,\ l_4 = 5.$$

En ese ejemplo $m = 5$ y tenemos $2^{5-2} = 8$, $2^{5-3} = 4$, $2^{5-5} = 1$, lo que nos da el tamaño de cada subarbol. Los códigos obtenidos para x_1, x_2, x_3, x_4 son $00, 010, 011 y 10000$. ∘

Ejercicio 4.2 Probar de manera mas formal que la codificación obtenida en la segunda parte de la prueba del teorema 4.1 es prefija. Para entender mejor, intentar la misma construcción con el mismo ejemplo, pero esa vez ordenando los subarboles con tamaño creciente de arriba abajo: que pasa? Porque no funciona?

Ejercicio 4.3 Importante! comprobar que la construcción de código de la segunda parte del teorema 4.1 se resume en el siguiente algoritmo:

Algoritmo calculando una codificación prefija $h(x_i)$
para una sucesión $l_1 \leq \cdots \leq l_i \leq \cdots \leq l_k$ de longitudes fijadas y verificando la desigualdad de Kraft.

1. $m = \text{máx}_i\, l_i$;
2. $h(x_1) =$ los l_1 primeros bits de 0 (o sea $0\ldots0$ l_1 veces);
3. $h(x_i) =$ los l_i primeros bits de $\sum_{k=1}^{i-1} 2^{m-l_k} = 2^{m-l_1} + \cdots + 2^{m-l_{i-1}}$, $i \geq 2$, escrito como un número binario de m bits, conservando los zeros a izquierda.

El teorema precedente nos da una condición necesaria sobre las longitudes de códigos para que una codificación prefija permita codificar k mensajes. Ahora bién, conviene hacer que las longitudes l_i de los códigos sean minimas. Por eso consideremos de nuevo un conjunto $\mathcal{X} = (x_1, \ldots, x_k)$ de k mensajes con una distribución $p = (p_1, \ldots, p_k)$. Si h es una codificación de los k mensajes con $l_i = h(x_i)$, queremos minimizar la longitud media, o sea la esperanza de la longitud de los códigos. En términos de esperanza esa longitud media es

$$L(h) := \sum_{i=1}^{k} p_i l(h(x_i)) = \sum_i p_i l_i = \mathbb{E}(l(h(X))).$$

Teorema 4.2 *Llamemos $L_{\text{ínf}} := \text{ínf}_h L(h)$, la longitud media mínima que se puede obtener con una codificación prefija. Entonces*

$$H(p) \leq L_{\text{ínf}} \leq H(p) + 1.$$

Lemma 4.1 *Sea $p = (p_1, \ldots, p_k)$ una distribución de probabilidad. La solución única del problema donde las incognitas son los q_i:*

$$\begin{cases} -\sum_{i=1}^{k} p_i \log q_i = \text{min!} \\ \sum_{i=1}^{k} q_i \leq 1,\ q_i \geq 0. \end{cases} \tag{4.2}$$

es $q_i = p_i$.

Prueba Si $p = (p_i)$ es una distribucion de probabilidad y $q_i \geq 0$ satisface $\sum_i q_i \leq 1$, utilizando la concavidad del logaritmo,

$$\sum_{i=1}^{k} p_i \log \frac{q_i}{p_i} \leq \log \sum_{i=1}^{k} q_i \leq \log 1 \leq 0,$$

que nos da

$$-\sum_{i=1}^{k} p_i \log p_i \leq -\sum_{i=1}^{k} p_i \log q_i. \tag{4.3}$$

Luego $q := p$ realiza el mínimo en el problema (4.1). La desigualdad es estricta a menos que $\sum_i q_i = 1$ y $p_i = q_i$ por todo i. ∘

Prueba del teorema 4.2. El problema que queremos resolver es minimizar la longitud media de código bajo la condición dada por la desigualdad de Kraft, o sea buscar $(l_i)_{i=1,\ldots,k}$ tales que

$$\begin{cases} -\sum_{i=1}^{k} p_i l_i = \text{min!} \\ \sum_{i=1}^{k} 2^{-l_i} \leq 1. \end{cases} \tag{4.4}$$

Resolvemos primero el problema sin preocuparnos del hecho de que los l_i tienen que ser enteros. Busquemos cualquier solución $l_i \geq 0$. Entonces poniendo $y_i := 2^{-l_i}$ el problema

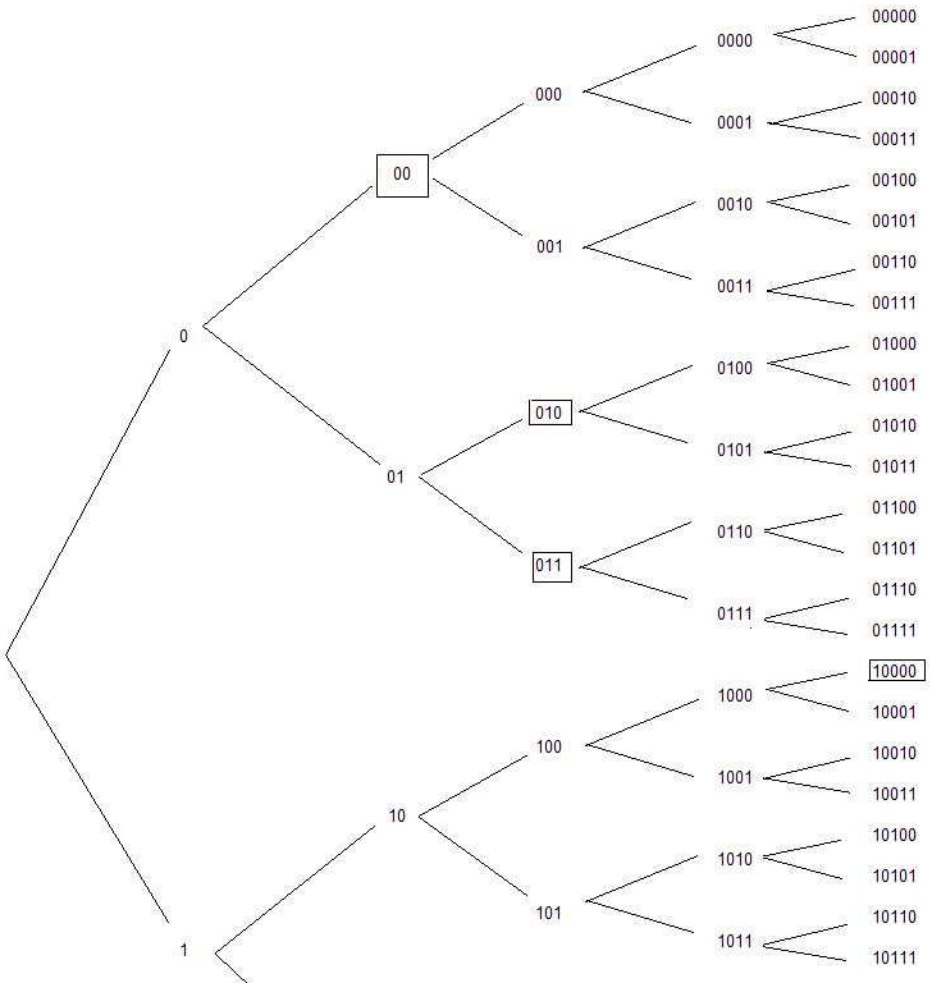

Figura 4.2: En un arbol binario de profundidad 5 se construye una codificación prefija para la serie de longitudes $l_1 = 2$, $l_2 = l_3 = 3$, $l_4 = 5$. Es facil ver que esa serie verifica la desigualdad de Kraft. En la figura se ven de arriba abajo los códigos obtenidos, creando subarboles disjuntos y de tamaño decreciente. Cada uno tiene un número de hojas 2^{5-l_j} y la raíz del subarbol se vuelve el código de x_j.

(4.4) es equivalente al problema (4.2). Luego su solución es $l_i^* = -\log p_i$ y obtenemos que el minimo $\sum_{i=1}^k p_i l_k^* = H(p)$ es la entropia. Sin embargo los l_i^* no son generalmente enteros y lo mejor que podemos hacer es fijar $l_i := \lceil l_i^* \rceil$, el minimo entero superior a l_i^*. Como $l_i \geq l_i^*$ la condición de Kraft es verificada:

$$\sum_{i=1}^k 2^{-l_i} \leq \sum_{i=1}^k 2^{-l_i^*} \leq 1.$$

Luego existe una codificación prefija cuyos códigos tienen longitud l_i y ademas:

$$H(p) = \sum_{i=1}^k p_i l_i^* \leq \sum_{i=1}^k p_i l_i = \mathbb{E}l(h(X)) \leq \sum_{i=1}^k p_i (l_i^* + 1) = H(p) + 1.$$

∘

4.1.1. Un primer ejemplo: el código de Shannon

Sea X una fuente con símbolos $\mathcal{X} = \{x_1, \ldots, x_n\}$ y distribución de probabilidad $p = (p_1, \ldots, p_n)$. Segun el teorema 4.2 podemos construir códigos casi optimales prefijos para una fuente $p = (p_1, \ldots, p_i, \ldots, p_k)$ cogiendo $l_i := \lceil \log p_i \rceil$ como longitud del código de x_i.

Algoritmo de codificación de Shannon

1. ordenar los p_i de manera decreciente: $p_1 \geq p_2 \geq \cdots \geq p_n$;
2. sea $P_i := \sum_{k=1}^{i-1} p_i$ por $i \geq 2$ y $P_1 = 0$;
3. $h(x_i)$ es compuesto de las $l_i = \lceil -\log p_i \rceil$ primeras cifras del desarrollo binario de P_i. En otros términos escribir $P_i = 0, a_1 a_2 \ldots a_{l_i} \ldots$ y guardar $a_1 \ldots a_{l_i}$.

Vamos a comprobar directamente que el código es prefijo y casí optimal, en pocas lineas!

En efecto, como tenemos $-\log p_i \leq l_i \leq -\log p_i + 1$, se obtiene

$$H(p) = -\sum_i p_i \log p_i \leq \sum_i p_i l_i = El \leq -\sum_i p_i (\log p_i + 1) = H(p) + 1.$$

Eso implica que la longitud media (o sea esperanza de la longitud) de los códigos, $\mathbb{E}l$, verifica

$$H \leq \mathbb{E}l \leq H + 1.$$

Eso quiere decir que la codificación es casí optimal. En efecto verifica la misma desigualdad que una codificación optimal y se pierde todo lo más un bit por símbolo. Nos queda por demostrar que la códificacion así definida tiene la propiedad del prefijo. Observemos que si $j \geq 1$, $P_{i+j} - P_i \geq p_i \geq 2^{-l_i}$. Si los l_i primeros bits de P_i coincidieran con los de P_{i+j} eso implicaria $P_{i+j} - P_i < 2^{-l_i}$. Luego no coinciden y la codificación es prefija.

Ejercicios

Ejercicio 4.4 Comparar el código de Shannon con el codigo del algoritmo 4.3.

Ejercicio 4.5 Experimentos. Implementar en Scilab un algoritmo que:

1. dado un texto extrae las frecuencias (probabilidades empiricas) de todos los caracteres incluyendo espacios y cifras. Así se deduce una distribución empirica de esos carácteres $p = (p_1, \ldots, p_k)$;
2. Calcula la entropia binaria de p, que indica cuantos bits hay que gastar por carácter;
3. deduce cual es la longitud teórica prevista para el texto en bits (producto de la entropia por el número de carácteres);
4. calcula los códigos de Shannon asociados con p;
5. genera el código binario del texto, calcula su longitud y la compara con la longitud teórica prevista.

Ejercicio 4.6 De una codificación $x \in \mathcal{X} \to h(x) \in \{0,1\}^{(\mathbb{N})}$ se dice que es únicamente descifrable si se tiene la implicación:

$$h(x_{i_1}) \ldots h(x_{i_n}) = h(x_{j_1}) \ldots h(x_{j_k}) \Rightarrow n = k \text{ y } x_{i_1} = x_{j_1}, \ldots, x_{i_n} = y_{j_k}.$$

Demostrar que si h es una codificación con la propiedad de prefijo, entonces la codificación que consiste a invertir el orden de cada $h(x_i)$ es únicamente descifrable. En conclusión, hay códigos unicamente descifrables que no tienen la propriedad del prefijo.

Ejercicio 4.7 Encontrar una codificación optimal para la distribución

$$p := (0{,}01;\ 0{,}04;\ 0{,}05;\ 0{,}07;\ 0{,}09;\ 0{,}1;\ 0{,}14;\ 0{,}2;\ 0{,}3).$$

Ejercicio 4.8 Sean $p = (p_1, \ldots, p_n)$ y $q = (q_1, \ldots, q_l)$ dos distribuciones discretas y $p \otimes q$ su producto tensorial definido por

$$p \otimes q = (p_1 q_1, \ldots, p_1 q_l,\ p_2 q_1, \ldots, p_2 q_l,\ \ldots, p_m q_1, \ldots, p_m q_l).$$

Que interpretación puede darse de esa distribución en términos de variables aleatorias? Comprobar que

$$H(p \otimes q) = H(p) + H(q).$$

Deducir que $H(p^{(n)}) = nH(p)$ donde $p^{(n)}$ es el producto tensorial de p por si misma, n veces.

Ejercicio 4.9 Formulario de Shannon (fuente : Shannon)
Empecemos recordando un resultado que ya hemos utilizado:

Lemma 4.2 *Sean p y q dos distribuciones de probabilidad discretas. Entonces*

$$\sum p(x) \log \frac{p(x)}{q(x)} \geq 0.$$

La igualdad se produce si y solo si $p(x) = q(x)$ para todo x.

Demostración del Lemma 4.2. Se involucra la concavidad estricta del logaritmo.

$$-\sum p(x)\log\frac{p(x)}{q(x)} = \sum p(x)\log\frac{q(x)}{p(x)} \leq \log(\sum p(x)\frac{q(x)}{p(x)}) = \log 1 = 0,$$

esa desigualdad siendo una igualdad si y solo si todos los valores $\frac{p(x)}{q(x)}$ sont iguales. En tal caso su valor común es obviamente 1. ∘

Sean X y Y dos variables aleatorias discretas de distribución conjunta $p(x,y) = \mathbb{P}(X = x, Y = y)$. Luego se tiene $\mathbb{P}(X = x) = p(x) = \sum_{y\in\mathcal{X}} p(x,y)$ y $\mathbb{P}(Y = y) = p(y) = \sum_{x\in\mathcal{X}} p(x,y)$ La entropia (o incertidumbre) de una variable discreta X con valores en el alfabeto finito $\mathcal{X}$, y la entropia de una pareja de variables aleatorias (X,Y) con valores en $\mathcal{X}\times\mathcal{Y}$ se han definido por

$$H(X) = -\sum_x p(x)\log p(x), \quad H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y).$$

1) Comprobar que la entropia de X es la esperanza de $g(X)$, donde $g(X) = \log\frac{1}{p(X)} = -\log p(x)$.

2) Probar que $H(X) \leq \log \mathrm{Card}(\mathcal{X})$, y que esa desigualdad se convierte en una igualdad si y sólo si X tiene una distribución uniforme sobre $\mathcal{X}$. (Utilizar el lema 4.2 donde p es la distribución de X y q la distribución uniforme sobre $\mathcal{X}$).

3) Ejemplo importante : se elija para X la variable de Bernouilli, $X = 1$ con probabilidad p, $X = 0$ con probabilidad $1-p$. Entonces

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

función de p que llamaremos $H(p)$. Comprobar que $H(X) = 1$ bit cuando $p = \frac{1}{2}$. Dibujar el grafo de $H(p)$ y probar que H vale 0 en 0 y 1 y es maximal cuando $p = \frac{1}{2}$. Interpretación : la incertidumbre sobre X es maximal cuando $p = \frac{1}{2}$ y minimal cuando X es determinista. La entropia es una misura de la incertidumbre sobre el valor de X.

4) La entropia conjunta de dos variables aleatorias es mas pequeña que la suma de las entropias. La igualdad se produce si y solo si ambas variables aleatorias son independientes.

$$H(X,Y) \leq H(X) + H(Y). \tag{4.5}$$

Basta aplicar el lema 4.2 a las distributiones $p(x,y)$ y $p(x)p(y)$.

5) **Entropia conditional de Y sabiendo X :** es "la media de la entropia de Y para cada valor de X, ponderada por la probabilidad de observar ese valor particular de X". (Comprobar que la formula que sigue corresponde a la definición entre comillas, debida a Shannon).

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

"*This quantity measures how uncertain we are of Y on the average when we know X.*" Utilizando la definición de la probabilidad conditional :

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

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

Luego

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

"*The uncertainty (or entropy) of the joint event X, Y is the uncertainty of X plus the uncertainty of Y when X is known*". Pero sabemos que

$$H(X) + H(Y) \geq H(X,Y) = H(X) + H(Y|X).$$

Entonces

$$H(Y) \geq H(Y|X).$$

"*The uncertainty of Y is never increased by knowledge of X. It will be decreased unless X and Y are independent events, in which case it is not changed*".

Ejercicio 4.10 Entropia relativa y información mutua

1) Consideremos dos distributiones de probabilidad $p(x)$ y $q(x)$. Llamaremos distancia de Kullback Leibler, o entropia relativa de las dos distributiones $p(x)$ y $q(x)$ la cantidad

$$D(p||q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p(X)}{q(X)}.$$

Comprobar que $D(p||q) \geq 0$ y que $D(p||q) = 0 \Leftrightarrow p = q$. La entropia relativa es una medida de la distancia entre dos distribuciones. Se llama información mutua $I(X,Y)$ la entropia relativa entre la distribución conjunta $p(x,y)$ y la distribución producto $p(x)p(y)$,

$$I(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) = \mathbb{E}_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}.$$

2) Demostrar que

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

$$I(X,Y) = I(Y,X) = H(X) + H(Y) - H(X,Y), \quad I(X,X) = H(X).$$

Se observara que $I(X,Y) = 0$ si X et Y son independientes y que $I(X,Y) = 0$.

3) Sean X_1,, X_2, ..., X_n variables aleatorias discretas de distribución conjunta $p(x_1,\ x_2, \ldots x_n)$. Probar que

$$H(X_1,\ X_2, \ldots X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \tag{4.6}$$

4) Mostrar, utilizando las definiciones de D y de la probabilidad condicional que

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

5) Deducir finalemente que

$$H(X_1,\ X_2, \ldots X_n) \leq \sum_{i=1}^{n} H(X_i),$$

y que la igualdad se produce si y solo si X_i son independientes. (Utilizar la pregunta precedente y (4.6)).

Capítulo 5

La codificación de Huffman

En 1952 Huffman descubrió una codificación particularmente sencilla. Vamos a describirla y a probar su optimalidad. El algoritmo se explica de nuevo muy bien con una figura y un ejemplo. Sea

$$p = (p_1, \ldots, p_k) = (0{,}01; 0{,}02; 0{,}04; 0{,}13; 0{,}13; 0{,}14; 0{,}15; 0{,}15; 0{,}23),$$

ordenadas de forma creciente. A partir de esa sucesión se construye un árbol binario cuyas hojas seran las probabilidades. A cada paso de la construcción del arbol se agrupan las dos probabilidades que tienen la suma inferior y se sustituyen por su suma. O sea que se pasa a la sucesión obtenida reuniendo $(p_1 + p_2, \ldots, p_k)$. El nudo padre de p_1 y p_2 es entonces $p_1 + p_2$. Iterando el proceso $n - 1$ veces, se construye un arbol tal como indicado en la figura 5.1. Con la convención gráfica que la probabilidad mayor se pone a derecha y la probabilidad menor a izquierda, el arbol se reordena tal como indicado en la figura 5.2. Entonces podemos asociar con la convención ya habitual un código binario con cada nudo en el arbol, lo que fija en particular un código para cada hoja del arbol, o sea cada p_i. Como vamos a ver, esa procedimiento nos da una codificación optimal en el sentido ya adoptado en el capítulo precedente.

Ejercicio 5.1 Calcular la longitud media del código que se ha construido!

El lema que sigue nos va a explicar porqué un código de Huffman es optimal.

Lemma 5.1 *Sea $n \geq 3$ y consideremos una distribución decreciente de probabilidad $p = (p_1, p_2, \ldots, p_k)$, con $p_1 \geq p_2 \geq \cdots \geq p_k > 0$. Entonces todo código prefijo optimal h puede ser transformado en otro con las mismas longitudes (y por ello igualmente optimal), pero que encima verifica:*

$$h(k) = w0, \quad h(k-1) = w1,$$

para al menos una sucesion w hecha de zeros y unos, y de manera que el código h' definido por

$$h'(i) = h(i), \quad 1 \leq i \leq k-1, \quad h'(k-1) = w$$

sea prefijo optimal para la distribución $p' = (p_1, p_2, \ldots, p_{k-2}, p_{k-1} + p_k)$.

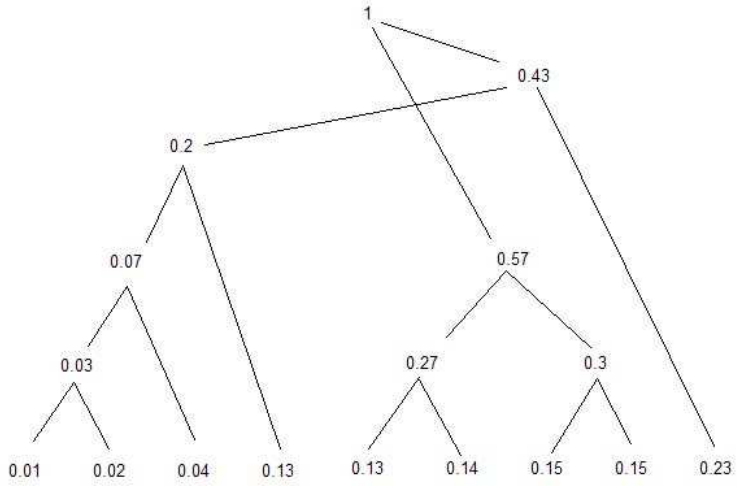

Figura 5.1: Sea $p = (p_1, \ldots, p_n) = (0{,}01; 0{,}02; 0{,}04; 0{,}13; 0{,}13; 0{,}14; 0{,}15; 0{,}15; 0{,}23)$ una distribución ordenada. A partir de esa sucesión se construye un árbol binario. A cada paso de la construcción del arbol se agrupan las dos probabilidades que tienen la suma más pequeña y su suma se vuelve el nudo padre de esas dos probabilidades. Iterando el proceso $n-1$ veces, se construye un arbol binario cuya raíz es 1, la suma de todas las probabilidades, y cuyas hojas son las probabilidades de partida.

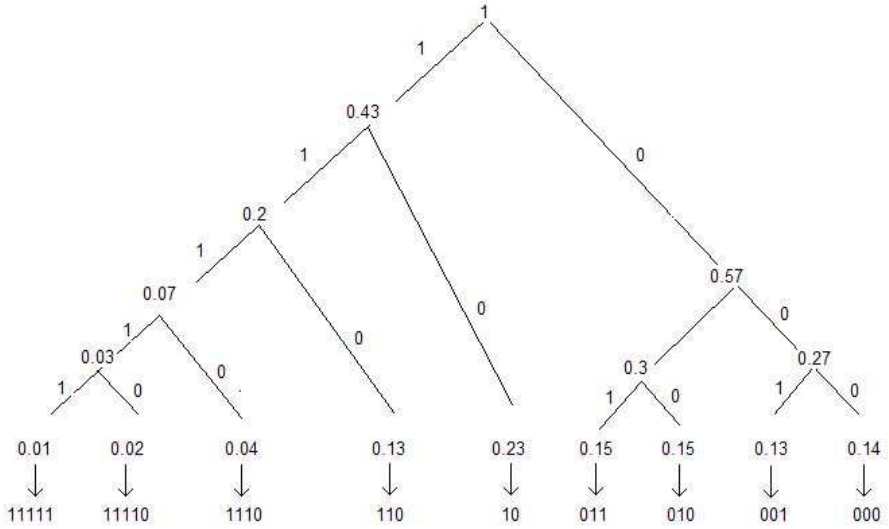

Figura 5.2: Con la convención gráfica que la probabilidad mayor se pone a derecha y la probabilidad menor a izquierda de arriba abajo, el arbol de la figura 5.1 se reordena. Con la convención ya habitual se asocia un código binario con cada nudo en el arbol, lo que fija en particular un código para cada hoja del arbol, o sea para cada p_i.

En todos los casos estamos hablando de codificación prefija. Veamos primero porqué el lema justifica la construcción del código de Huffman. El lema nos garantiza que para construir una codificación optimal de una distribución de n elementos, basta encontrar un código optimal para la distribución de $n-1$ elementos obtenida agrupando las dos probabilidades mas pequeñas. Luego el código de esas probabilidades p_{k-1} y p_k se obtiene añadiendo un zero y un 1 al código de $p_k + p_{k-1}$. Al ultimo paso, cuando se ha reducido n a 2, la codificación optimal es forzada, 0 y 1. Luego se deduce que todos los códigos sucesivos son optimales.

Prueba del lema 5.1. Sea h un código optimal para p. Podemos imponer, como una propiedad general de los códigos prefijos, que las longitudes verifiquen $l_1 \leq l_2 \leq \cdots \leq l_k$. En efecto si por ejemplo $l_1 > l_2$, hay dos casos : si $p_1 = p_2$ podemos intercambiar p_1 y p_2. En cambio $p_1 > p_2$ es imposible, ya que implicaria que podemos obtener un código de longitud media estrictamente inferior, intercambiando los códigos de p_1 y p_2. En efecto ese intercambio daría $p_2 l_1 + p_1 l_2 < p_1 l_1 + p_2 l_2$.

Observemos también como propiedad general de un código prefijo que $l_{k-1} = l_k$. Si así no fuera, podriamos mantener la propiedad del prefijo y disminuir la longitud media quitando los ultimos $l_k - l_{k-1}$ últimos bits de $h(k)$.

El código $h(k-1)$ puede escribirse $w0$ o bien $w1$. Pongamos que $w0$. Entonces podemos eligir $h(k) = w1$. En efecto si ese código ya esta utilizado para algun p_i, podemos intercambiar el código $h(i)$ de p_i y el de p_{k-1}. Si el código $w1$ no es utilizado, es obvio que lo podemos utilizar para $h(k)$ y a la vez mantener la propiedad del prefijo. En efecto si otro $h(j)$ fuera prefijo de $w1$, siendo de longitud estrictamente inferior ya que no es igual a $w1$, $h(j)$ seria prefijo de $w0 = h(k-1)$. Con todo, la longitud media no ha cambiado y luego se queda optimal.

Ahora bién. Consideremos el código $\tilde{h}$ inducido por h sobre $p' := (p_1, \ldots, p_{k-2}, p_{k-1} + p_k)$:

$$\tilde{h}(i) = h(i), \quad (1 \leq i \leq k-2) \quad \text{ y } \tilde{h}(k-1) = w.$$

Para concluir nos queda por demostrar que $\tilde{h}$ es de longitud media optimal. La longitud media de esa codificación, $\tilde{L}$, esta relacionada con la de h, L, por

$$L = \tilde{L} + p_{k-1} + p_k.$$

Sea L' la longitud media de un código optimal h' sobre p'. Partiendo de h' puede definirse un codigo $\hat{h}$ sobre p por

$$\hat{h}(i) = h'(i), \quad (1 \leq i \leq k-2), \quad \hat{h}(k-1) = h'(k-1)0, \quad \hat{h}(k) = h'(k-1)1.$$

Luego la longitud media de $\hat{h}$ es

$$\hat{L} = L' + p_{k-1} + p_k.$$

Pero sabemos que L' es la longitud optimal de p' y que L es la longitud optimal de p. Luego $\hat{L} \geq L$, $\tilde{L} \geq L'$ y obtenemos:

$$\tilde{L} + p_{k-1} + p_k = L \leq \hat{L} = L' + p_{k-1} + p_k \leq \tilde{L} + p_{k-1} + p_k,$$

que nos da $L' = \tilde{L}$. Luego $\tilde{h}$ es optimal. ∘

5.1. Ejercicios

Ejercicio 5.2 Experimentos. Implementar en Scilab un algoritmo que:

1. dado un texto de N carácteres extrae las frecuencias (probabilidades empiricas) de todos los caracteres incluyendo espacios y cifras. Así se deduce una distribución empirica de esos carácteres $p = (p_1, \ldots, p_k)$;

2. de la misma manera, fijado un entero k (en práctica $k = 2, 3, 4o5$), calcula las frecuencias de cada sequencia de k símbolos (por supuesto quedarse tan solo con las secuencias que aparecen al menos una vez.) Esa distribución de probabilidad se llama p^k;

3. calcula la entropia binaria de $H^k := H(p^k)$, que indica cuantos bits hay que gastar por carácter;

4. deduce cual es la longitud teórica prevista para el texto en bits (producto de la entropia $H(p^k)$ por el número de carácteres divido por k);

5. calcula los códigos de Shannon asociados con p^k;

6. genera el código binario del texto, calcula su longitud L^k y la compara con la longitud teórica prevista;

7. verifica que $\frac{N}{k}H^k \leq L^k \leq \frac{N}{k}(H^k + 1)$;

8. verifica que $k \to \frac{H^k}{k}$ es una función decreciente y da el factor optimo de compressión obtenido.

Ejercicio 5.3 La definición de la entropia, opus cit. p. 49 Consideremos un conjunto finito de acontecimientos cuyas probabilidades son $p_1, \ldots, p_n$. La entropia $H(p_1, \ldots, p_n)$ se va a definir como una medida de la incertidumbre sobre cual de los acontecimientos $i = 1, \ldots, n$ se va a producir. Para entender los axiomas que nos llevan a la definición de la entropia, hay que tomar en cuenta que una partición de acontecimientos disjuntos puede ser el objeto de agrupaciones parciales. Por ejemplo podemos agrupar los k primeros acontecimientos en un acontecimiento único de probabilidad p'_1. Así se obtiene una distribución de probabilidad, $(p'_1 = \sum_{i=1}^{k} p_i,\ p_{k+1}, \ldots,\ p_n)$ tal que a su vez el primer acontecimiento se descompone en $(\pi_1, \ldots,\ \pi_k)$ con $\pi_i = \frac{p_i}{p_1 + \cdots + p_k}$. En la distribución inicial teniamos un sorteo entre n acontecimientos. En el segundo caso tenemos primero un sorteo entre $n-k+1$ acontecimientos disjuntos, seguido de un segunto sorteo entre k acontecimientos en el caso de que el primer sorteo haya dado 1. En resumen, tenemos presentaciones del mismo sorteo final bajo dos formas:
$(p_1, \ldots\ p_n)$, o bién
$((p'_1, \pi_1), \ldots\ (p'_1, \pi_k), p_{k+1}, \ldots\ p_n)$
Formalicemos la entropia o incertidumbre con los axiomas intuitivos siguientes

1. H es continua

2. Supongamos que los p_i son iguales, $p_i = \frac{1}{n}$. Entonces H tiene que ser una función creciente de n

3. Si se recomponen los n acontecimientos por agrupación seguida de sorteo condicional, como indicado arriba, está claro que la incertidumbre final es la misma. Eso nos lleva a exigir
$$H(p_1, \ldots, p_n) = H(p'_1, p_{k+1}, \ldots p_n) + p'_1 H(\pi_1, \ldots \pi_k).$$

Nuestra meta es demostrar que con esos axiomas 1, 2 y 3 existe una constante positiva K tal que
$$H(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log p_i.$$

1) Escribamos $H(\frac{1}{n}, \ldots \frac{1}{n}) = A(n)$. Utilizando el axioma 3 probar que $A(s^m) = mA(s)$. Si t es otro entero, se puede encontrar, para n arbitrariamente grande, un m tal que $s^n \leq t^n < s^{m+1}$. Utilizando la monotonia (axioma 2) deducir que $\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n}$ y finalemente que $A(t) = K \log t$.

2) Supongamos que $p_i = \frac{n_i}{\sum_1^n n_i}$ son probabilidades racionales. Probar grácias al axioma 3 que
$$K \log \sum n_i = H(p_1, \ldots, p_n) + K \sum p_i \log n_i.$$

3) Tratar el caso general aproximando los p_i con racionales y utilizando el axioma 2 de continuidad.

Ejercicio 5.4 Mensajes típicos Consideremos una sucesión X_n de v.a.i.i.d. con valores en un conjunto finito de símbolos $\mathcal{X} = \{x_1, x_2, ..., x_k\}$ y tales que $P(X_n = x_i) = p_i$, $i = 1, ..., k$. Sea $\mathcal{X}^n$ el conjunto de las sucesiones de longitud n, que llamamos mensajes. Se define $\mathbb{P}$ sobre $\mathcal{X}^n$ por $\mathbb{P}(x_1, \ldots, x_n) := p(x_1) \ldots p(x_n)$. Hay k^n tales mensajes. Consideremos también la entropia de la repartición $(p_i)_{1 \leq i \leq k}$,
$$H(p_1, ..., p_k) = -\Sigma_{i=1}^{i=k} p_i \log p_i.$$

Esa cantidad la vamos a interpretar en relación con la probabilidad de un mensaje largo. La observación crucial es que los mensajes largos (n grande) tienen todos mas o menos la misma probabilidad de ser emitidos. Par darse cuenta de ello, basta aplicar la ley debil de los números grandes a la variable aleatoria definida como el logaritmo medio de la probabilidad de una sucesión larga,
$$\frac{1}{n} \log \mathbb{P}(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} \log p(X) \to E(\log p(X_i)) = \sum_i p_i \log p_i = H(p_1, \ldots, p_k).$$

Se deduce la fórmula fundamental
$$\mathbb{P}(X_1, \ldots, X_n) = 2^{-n(H(p_1, ..., p_k) + \epsilon(n))} \quad \text{q.s.},$$

con $\epsilon(n) \to 0$ cuando $n \to +\infty$.

Esa observación nos lleva a definir lo que llamaremos el conjunto de los "mensajes típicos",
$$C_n = \{(x_1, ..., x_n) \in \mathcal{X}^n, \ 2^{-n(H+\varepsilon)} \leq p(x_1)...p(x_n) \leq 2^{-n(H-\varepsilon)}\}.$$

1) Probar que $E(-\log p_X) = H$.

2) Deducir que

$$\mathbb{P}(C_n^c) \leq P(\{|\frac{1}{n}\Sigma_{l=1}^n(-\log_2 p_{X_l}) - H| \geq \varepsilon\}) \leq \frac{\mathrm{Var}(-\log_2 p_{X_1})}{n\varepsilon^2}.$$

3) Probar que

$$P(C_n)) \geq 2^{-n(H+\varepsilon)}\mathrm{Card}(C_n).$$

Deducir que $\mathrm{Card}(C_n) \leq 2^{(H+\varepsilon)n}$.

4) Recíproca. Supongamos que hayamos encontrado $\tilde{C}_n \subset \mathcal{X}^n$ tal que $\lim_{n\to+\infty} P((X_1,\ ...,\ X_n) \in \tilde{C}_n) = 1$ y $\mathrm{Card}(\tilde{C}_n) \leq 2^{Kn}$. Vamos a probar que $K \geq H$.

4a) Comprobar que $\lim_{n\to\infty} \mathbb{P}((X_1,\ ...,\ X_n) \in C_n \cap \tilde{C}_n) = 1$.

4b) Demostrar que $\mathbb{P}((X_1,\ ...,\ X_n) \in C_n \cap \tilde{C}_n) \leq 2^{-n(H-\varepsilon)}2^{Kn}$ et sacar la conclusión.

5) Sean $\mathcal{X} = \{0, 1\}$, $p_1 = p$, $p_2 = (1-p)$. Si $p_1 = p_2 = \frac{1}{2}$, comprobar que $H = 1$ y que el número de las sucesiones típicas es 2^n. (En otros términos no la compresión es imposible).
Caso general : estudiar la forma de $H(p) = -p\log p - (1-p)\log(1-p)$ y deducir de ella el comportamiento del número de sucesiones típicas.

6) Applicación a la codificación. Vamos a interpretar H como la longitud media de mensajes codificados en el alfabeto $\{0, 1\}$ de manera optimal. Empezemos con la descripción de una codificación que nos dé una tal longitud media. Para realizarla, fijemos $\varepsilon > 0$ y elijamos n bastante grande, de manera que $\mathbb{P}(C_n^c) \leq \varepsilon$. (Explicar porque eso es posible). Luego, se atribuye un código binario a cada uno de los elementos de C_n. Eso implica que el número de códigos binarios es inferior o igual a $2^{n(H+\varepsilon)}$. Ahora bién, consideremos los códigos non típicos, que son mucho mas numerosos! Su número se acerca a $k^n = 2^{n\log k}$. Luego los podemos enumerar con todo lo más k^n códigos binarios distintos de los precedentes o sea números inferiores a $2^{n\log k} + 2^{n(H+\varepsilon)} \leq 2^{n\log k+1}$. Así hemos logrado atribuir un código a todos los elementos de $\mathcal{X}^n$. Demostrar que la longitud media de un código binario en esa codificación es inferior o igual a $n(H + \varepsilon(1 + \log k))$. $H(p)$ puede en consecuencia interpretarse como la longitud media de código utilizada para cada simbolo cuando n es grande.

7) Finalmente nos tenemos que preguntar si no se podria encontrar una codificación todavia mas eficaz. Si fuera posible tendriamos un subconjunto de mensajes $\tilde{C}_n$ de cardinalidad mas pequeña que 2^{nK} con $K < H$ y tal que $\mathbb{P}(\tilde{C}_n) \to 1$ cuand $n \to \infty$. Demostrar que no eso no es posible. Concluir que acabamos de demostrar que:

La longitud mínima por símbolo de una codificación trasmitiendo mensajes de longitud n en el alfabeto $\mathcal{X}$ con la distribución $p_1, \ldots,\ p_k$ es $H(p_1, \ldots,\ p_k)$.

Capítulo 6

Lenguaje y comunicación segun Shannon

6.1. Introducción

Nuestra meta aqui es, siguiendo Shannon, explicar como podemos reducir el problema de medir la cantidad de información transmitida en un dispositivo de comunicación a un análisis tan sencillo como el de una distribución discreta de probabilidad. Shannon reduce la comunicación a la transmisión de una serie de símbolos emitidos por una fuente aleatoria. La fuente emite los símbolos. Cada uno de ellos representa, por ejemplo, una frase en castellano. La incertidumbre del receptor es grande, pero no tanto, ya que hay frases más probables que otras. Por ejemplo en una conversación la probabilidad de que una respuesta sea "si" o "no" no es nada negligible. Luego, la hipotesis fundamental es que tanto el receptor como el emisor conocen la probabilidad de cada frase. Pero tal hipotesis podría parecer fantástica si Shannon no nos diera el medio de calcular efectivamente una buena estimación de cada unidad significativa en el lenguaje, de una letra a una frase o incluso un texto. El modelo subyacente es un modelo Markoviano. Se supone por ejemplo que, dada una serie de símbolos $m_1 m_2 \dots m_k$, la probabilidad de que aparezca a continuación un símbolo m_{k+1} depende tan solo de m_k y no de las precedentes. Evidentemente, esa hipótesis markoviana es falsa, pero se vuelve cada vez más razonable cuando los símbolos crecen de tamaño.

Gracias a la hipotesis markoviana, la probabilidad de una sucesión de símbolos se puede calcular como

$$\mathbb{P}(m_1 m_2 \dots m_k) = \mathbb{P}(m_1)\mathbb{P}(m_2|m_1)\mathbb{P}(m_3 \mid m_2) \dots \mathbb{P}(m_k \mid m_{k-1}),$$

donde $\mathbb{P}(m_2 \mid m_1) = \frac{\mathbb{P}(m_1 m_2)}{\mathbb{P}(m_1)}$ es la probabilidad de que m_2 siga m_1 en un texto y $\mathbb{P}(m_1)$ es la probabilidad de que m_1 aparezca en un texto. Todas esas probabilidades pueden estimarse empiricamente con un texto o un conjunto de textos, donde se calcula la frecuencia de cada palabra y luego la frecuencia de cada par de palabras sucesivas. Calcular esas frecuencias y guardarlas en una especie de diccionario es informaticamente posible e incluso fácil.

Shannon, con su genio, y en un tiempo en que la manipulación informática no era posible, demuestra con un procedimiento muy simple que la "cadena de Markov del lenguaje" se puede simular con la máxima facilidad si se dispone de un libro. El interés de tal simulación es demostrar que *si se repectan las probabilidades y las probabilidades de transición aprendidas de un texto en inglés, entonces las frases sintetizadas se parecen al inglés!*

El procedimiento de Shannon consiste en :

- elegir al azar una palabra en un libro (abrir al azar, poner el dedo al azar)
- abrir de nuevo el libro al azar y buscar la misma palabra ; cuando se ha encontrado, eligir la palabra que la sigue
- iterar

Aqui tenemos los ejemplos históricos simulados por Shannon con ese procedimiento:

1. Aproximación de order zero (símbolos tirados independientes, equiprobables):
   XFOML RXKHRRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACI-UBZLHJQD.

2. Aproximación de orden 1 (símbolos independientes, pero con las frecuencias de un texto inglés).
   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTT-VA NAH BRL.

3. Aproximación de orden 2 (frecuencia de parejas de letras correcta para el inglés).
   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEATSONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Approximación de order 3 (frecuencias de los triples de letras correctas).
   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. Aproximación de orden 1 con palabras : la frecuencia de las palabras es correcta
   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFER-ENT NATURAL HERE HE THE A CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE;

6. Aproximación de orden 2 con palabras : las probabilidades de transición entre palabras son como en inglés.
   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED;

Shannon observa que : “The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that these samples have reasonably good structure out to about twice the range that is taken into account in their construction. Thus in 3. the statistical process insures reasonable text for two-letter sequences, but four-letter sequences from the sample can usually be fitted into good sentences. In 6. sequences of four or more words can easily be placed in sentences without unusual or strained constructions.”

6.2. Ejercicios

Ejercicio 6.1 Implementar en Scilab un algoritmo de **sintesis markoviana de texto probabilisticamente correcto**. Ese programa, por cada k:

- calcula la probabilidad empirica $\mathbb{P}(x_i \mid y^k) := \frac{\mathbb{P}(y^k x_i)}{\mathbb{P}(y^k)}$ que aparezca un carácter x_i sabiendo que esta precedido por un k-grama y^k. ($\mathbb{P}(x_i)$ es la frecuencia de x_i y $\mathbb{P}(y^k)$ es la frecuencia de y^k calculadas en el ejercicio precedente);
- sortea el k-grama inicial $x_{i_1} \dots x_{i_k}$ del texto sintetizado segun la distribución de probabilidad p^k;
- sortea el carácter siguiente segun la distribución condicional $x \to \mathbb{P}(x \mid x_{i_1} \dots x_{i_k})$;
- itera: dado un texto ya sintetizado de n caracteres, sintetiza el $n+1$-esimo carácter segun la distribución condicional $x \to \mathbb{P}(x \mid x_{i_{n-k+1}} \dots x_{i_n})$;
- saca el texto sintetizado.

Capítulo 7

Mensajes repetidos y entropia

Vamos a interpretar la entropia $H(p)$ como la media del logaritmo del número de mensajes "típicos" cuando una fuente manda una sucesión de n símbolos sucesivos independientes siguiendo todos la distribución p. Asimismo, se va a deducir que la entropia es la longitud media minimal, medida en bits/símbolo, requerida para codificar una fuente de entropia $H(p)$. Todas las sucesiones $x_1 x_2 \dots x_n$ son possibles, pero no son equiprobables. Sin embargo por la ley de los números grandes, la frecuencia de aparición de cada símbolo x en $x_1 \dots x_n$ tiende hacia $p(x)$ cuando n tiende al infinito. Por eso, habra mucho menos mensajes típicos que mensajes posibles. Vamos a ver que el número de mensajes típicos es de orden de magnitud $2^{nH(p)}$ mientras que el número de mensajes posibles es $(\mathrm{Card} X)^n = 2^{n \log \mathrm{Card}(X)}$.

7.1. Mensajes típicos

Consideremos una sucesión X_n de v.a.i.i.d. con valores en un conjunto finito de símbolos $\mathcal{X}$ y tales que $\mathbb{P}(X_n = x) = p(x)$. Sea $\mathcal{X}^n$ el conjunto de las sucesiones de longitud n, que llamamos mensajes. Hay k^n tales mensajes. $\mathcal{X}^n$ es un espacio de probabilidad, y escribiremos $\mathbb{P}$ su probabilidad producto, definida por $\mathbb{P}(x_{i_1} \dots x_{i_n}) = p(x_{i_1}) \dots p(x_{i_n})$. Consideremos también la entropia de la repartición $p = (p(x))_{x \in \mathcal{X}}$,

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

Vamos a interpretar esa cantidad en relación con la probabilidad de un mensaje largo. La observación crucial es que los mensajes largos (n grande) tienen todos mas o menos la misma probabilidad de ser emitidos. Para darse cuenta de ello, basta aplicar la ley fuerte de los números grandes a la variable aleatoria definida como el logaritmo medio de la probabilidad de una sucesión larga,

$$\frac{1}{n} \log \mathbb{P}(X_1, \dots, X_n) := \frac{1}{n} \sum_{i=1}^{n} \log p(X_i) \to E(\log p(X)) = \sum_{x \in \mathcal{X}} p(x) \log p(x) = H(p) \text{ (c.s.)},$$

donde (c.s.) significa casi siempre. Se deduce la fórmula fundamental

$$p(X_1, \dots, X_n) = 2^{n(H(p)+\epsilon(n))},$$

con $\epsilon(n) \to 0$ cuando $n \to +\infty$.

Esta applicación de la ley fuerte de los grandes números, que no vamos a utilizar (solo utilizaremos la ley débil de los números grandes), nos lleva sin embargo a la definición siguiente.

Definición 7.1 *Para cada $\varepsilon > 0$ llamaremos conjunto de los "mensajes típicos",*

$$C_n = \{(x_1, \ ..., \ x_n) \in \mathcal{X}^n, \ 2^{-n(H(p)+\varepsilon)} \leq \mathbb{P}(x_1 \dots x_n) = p(x_1)...p(x_n) \leq 2^{-n(H(p)-\varepsilon)}\}.$$

Lemma 7.1 *El conjunto C_n de los mensajes típicos asociado con p, n y ε verifica:*

1. *$\mathbb{P}(C_n) \geq 1 - \varepsilon$ para n suficientemente grande ;*
2. *Card$(C_n) \leq 2^{(H(p)+\varepsilon)n}$*

Prueba Pasando al logaritmo en las desigualdades definiendo C_n, vemos que

$$C_n = \{(x_1, \ ..., x_n) \mid -n(H(p) + \varepsilon) \leq \sum_{i=1}^{n} \log p(x_i) \leq -n(H(p) - \varepsilon)\}.$$

Luego

$$C_n = \{|\frac{1}{n}\Sigma_{i=1}^{n}(-\log p(X_i)) - H(p)| \leq \varepsilon\}.$$

Observemos que $E(-\log p(X)) = H(p)$. Eso sigue de la mera definición de la esperanza de una variable aleatoria $f(X)$ cuando X es otra variable aleatoria con valores en $\mathcal{X}$, $\mathbb{E}f(X) = \sum_{x\in\mathcal{X}} f(x)\mathbb{P}(X = x)$. Aqui se aplica con $f(x) = -\log p(x)$.

Considerando la variable aleatoria $S_n = \sum_{i=1}^{n} \log p(X_i)$, podemos aplicar la desigualdad de Tchebychev

$$\mathbb{P}(|\frac{S_n}{n} - \mathbb{E}(-\log p(X))| \geq \varepsilon) \leq \frac{\sigma^2(-\log p(X))}{n\varepsilon^2} \to 0 \text{ cuando } n \to \infty,$$

lo que nos da

$$\mathbb{P}((X_1, \ ..., \ X_n) \in C_n^c) = \mathbb{P}(\{|\frac{1}{n}\Sigma_{i=1}^{n}(-\log p(X_i)) - H(p)| > \varepsilon\}) \leq \frac{\sigma^2(-\log p(X))}{n\varepsilon^2}.$$

Fijado ε y eligiendo n suficientemente grande obtenemos $\mathbb{P}(C_n) \geq 1 - \varepsilon$.

En C_n las sucesiones tienen todas más o menos la probabilidad $2^{-nH(p)}$ y, más precisamente:

$$\mathbb{P}(C_n) \geq \sum_{(x_1 \dots x_n)\in C_n} \mathbb{P}(x_1 \dots x_n) \geq 2^{-n(H(p)+\varepsilon)}\text{Card}(C_n).$$

Como $\mathbb{P}(C_n) \leq 1$ deducimos que $\text{Card}(C_n) \leq 2^{(H(p)+\varepsilon)n}$.

∘

Vamos a interpretar $nH(p)$ como la longitud media de mensajes de n símbolos codificados en el alfabeto $\{0, 1\}$ de manera optimal.

Lemma 7.2 *Sea $\mathcal{X}$ un conjunto de mensajes elementales o símbolos y X una fuente emitiendo mensajes repitidos con la distribución $p = (p(x))_{x\in\mathcal{X}}$. Por todo $\varepsilon > 0$ podemos, si n es suficientemente grande, codificar los mensajes repitidos de tal manera que la longitud media por símbolo sea inferior o igual a $H(p) + \varepsilon$.*

Prueba Para realizar la codificación anunciada, fijemos $1 > \varepsilon > 0$ y elijamos n bastante grande, de manera que $\mathbb{P}(C_n^c) \leq \varepsilon$ y $\text{Card}(C_n) < 2^{n(H(p)+\varepsilon)}$. Luego podemos codificar todos los elementos de C_n utilizando números binarios mayores o iguales que $2^{[n(H(p)+\varepsilon)+1]}$ y estrictamente inferiores a $2^{[n(H(p)+\varepsilon)+2]}$. Son números binarios que tienen todos la misma longitud $[n(H(p)+\varepsilon)+2]$. Luego forman un código prefijo. Además al menos uno de esos códigos, m, queda inutilizado.

Ahora bién, consideremos los códigos no típicos, que son mucho mas numerosos! Su número es estrictamente inferior a $k^n = 2^{n \log k}$. De la misma manera que con los elementos de C_n, los enumeramos con numeros binarios que tengan todos la misma longitud $[n \log k + 2]$. Para obtener un codificación prefija, conviene añadir a todos esos numeros el prefijo m. Así obtenemos códigos de longitud $[n \log k + 2] + [n(H(p) + \varepsilon) + 2]$.

Hemos atribuido un código a todos los elementos de $\mathcal{X}^n$. La longitud de cada código binario de C_n es inferior o igual a $n(H(p) + \varepsilon) + 2$. La longitud de los demas códigos es inferior o igual a $n(\log k + H(p) + \varepsilon) + 4$. Como $\mathbb{P}_n(C_n) \leq 1$ y $\mathbb{P}_n(C_n^c) \leq \varepsilon)$, la longitud media $\mathbb{E}l_n$ de un mensaje de n símbolos con la precedente codificación verifica

$$\mathbb{E}l_n \leq \varepsilon(n(\log k + H(p) + \varepsilon) + 4) + (1 - \varepsilon)(n(H(p) + \varepsilon) + 2), \text{ que nos da:}$$

$$\lim \sup_{n \to \infty} \frac{\mathbb{E}l_n}{n} \leq (1 - \varepsilon)(H(p) + \varepsilon) + \varepsilon(\log k + H(p) + \varepsilon).$$

Como p y k son fijos y ε arbitrariamente pequeño obtenemos lo anunciado. ∘

Ejercicio 7.1 Observar que no necesitabamos imponer que la codificación precedente sea prefija. Porque? Dar una evaluación de cuantos bits se han perdido por hacerla prefija.

Ahora cabe demostrar la desigualdad inversa, o sea que no podemos codificar los mensajes de n símbolos con estricamente menos que $H(p)$ bits por símbolo.

Lemma 7.3 *Para todo $\epsilon > 0$, si n es suficientemente grande, la esperanza $L(h)$ de la longitud de cualquier codificacíon binaria h de C_n verifica*

$$L(h) \geq n(H(p) - \varepsilon).$$

Prueba Fijemos $\varepsilon > 0$ y consideremos el conjunto de los mensajes típicos C_n. La probabilidad π_i de cada uno de esos mensajes, enumerados de $i = 1$ a N, verifica

$$\pi(i) \geq 2^{-n(H(p)+\varepsilon)}. \tag{7.1}$$

La cardinalidad N del conjunto C_n verifica $2^{nH(p)-\varepsilon} \leq N \leq 2^{nH(p)+\varepsilon}$. Además, sabemos que

$$\mathbb{P}(C_n) \geq 1 - \varepsilon \tag{7.2}$$

para n suficientemente grande, o sea $\sum_{i=1}^{N} \pi_i \geq 1 - \varepsilon$. Sea $h(i)$ una codificación binaria de C_n. Para evitar que algunos de los códigos empiezen por zeros, les podemos, para que sean números binarios distintos en en el sentido habitual, juxtaponer a todos los $h(i)$ un 1 a izquierda, lo que aumenta su longitud de 1. Supongamos sin perdida de generalidad que los

nuevos códigos $1h(i)$ esten ordenados por orden creciente. Luego $1 \leq i \leq N$ es el rango de $1h(i)$. Finalmente observemos que $[\log i]$ es inferior o igual a la longitud de i escrito como un número binario. Como $1h(i) \geq i$, tenemos $l(1h(i)) \geq l(i) \geq [\log i]$, lo que nos da

$$\sum_{i=1}^{N} \pi_i l(h(i)) \geq \sum_{i=1}^{N} \pi_i [\log i] - 1 \geq -\sum_{i=1}^{N} \pi_i \log \frac{1}{i} - 2, \tag{7.3}$$

ya que $[r] \geq r - 1$. Para aplicar la desigualdad del lema 4.2, queremos transformar los $\frac{1}{i}$ en una distribución de probabilidad q_i, lo que se obtiene considerando

$$q_i := \frac{1/i}{\sum_{i=1}^{N} \frac{1}{i}}.$$

Así (7.3) se reescribe

$$\sum_{i=1}^{N} \pi_i l(h(i)) \geq -\sum_{i=1}^{N} \pi_i \log q_i - 2 - \log \left(\sum_{i=1}^{N} \frac{1}{i} \right). \tag{7.4}$$

Finalmente hacemos dos cosas : aplicamos el lema 4.2 que nos da, utilizando (7.1) y (7.2),

$$-\sum_{i=1}^{N} \pi_i \log q_i \geq -\sum_{i=1}^{N} \pi_i \log \pi_i \geq \sum_{i=1}^{N} \pi_i n(H(p) - \varepsilon) \geq n(H(p) - \varepsilon)(1 - \varepsilon).$$

Por otra parte hay que demostrar que el otro término en (7.4), $-\log \left(\sum_{i=1}^{N} \frac{1}{i} \right)$, es negligible ante n. Eso es fácil ya que

$$\sum_{i=1}^{N} \frac{1}{i} \leq \mathrm{Ln}N + 1 \leq (\mathrm{Ln}2) n(H(p) + \varepsilon) + 1.$$

Utilizando las dos ultimas estimaciones en (7.4) obtenemos

$$\sum_{i=1}^{N} \pi_i l(h(i)) \geq n(H(p) - \varepsilon)(1 - \varepsilon) + O(\log n).$$

Como podemos fijar ε arbitrariamente pequeño y n arbitrariamente grande, el resultado anunciado sigue.

∘

Combinando los lemas precedentes, ya podemos probar el resultado fundamental de Shannon :

Teorema 7.1 *La esperanza minimal $\mathbb{E}l$ de la longitud por símbolo de un mensaje emitido n veces por una fuente de entropia $H(p)$ es igual a $(H(p) + \varepsilon(n))$ con $\varepsilon(n) \to 0$ cuando $n \to \infty$.*

Prueba Consideremos por cada n una codificación h de longitud minimal de $\mathcal{X}^n$. Como $\mathcal{X}^n \supset C_n$, esa codificación codifica C_n y podemos aplicar el lema 7.3. Luego su longitud por símbolo es superior a $H(p) - \varepsilon$, para n grande. Como h es una codificación optimal, la esperanza de su longitud por símbolo es para n grande inferior o igual a $H(p) + \varepsilon$, gracias al lema 7.2. Combinando ambos resultados vemos que para todo $\varepsilon > 0$ y para n suficientemente grande, $H(p) - \varepsilon \leq \mathbb{E}l \leq H(p) + \varepsilon$.

∘

Deducimos de ese resultado que la entropia $H(p)$ no es tan solo un limite inferior de longitud para una codificación prefija, tal y como se probó en el teorema 4.2; lo es también para cualquier codificación únicamente descifrable.

Teorema 7.2 *La esperanza minimal de la longitud por símbolo de una codificación únicamente descifrable de (X, p) es mayor o igual que $H(p)$. En otros términos, por toda codificación h únicamente descifrable de X,*

$$\mathbb{E}l(h(X)) \geq H(p).$$

Prueba Codificamos X^n con h por $h(X_1 \dots X_n) = h(X_1) \dots h(X_n)$. Luego por el teorema 7.1,

$$\mathbb{E}l(h(X_1 \dots X_n)) = n\mathbb{E}l(h(X)) \geq n(H(p) - \varepsilon) \Rightarrow \mathbb{E}l(h(X)) \geq H(X) - \varepsilon,$$

por todo $\varepsilon > 0$. ∘

7.2. Ejercicios

Ejercicio 7.2 Vamos a probar que la longitud media $\lambda(N)$ de los números binarios inferiores N verifica $\lambda(N) \sim \log N$, donde $\varepsilon(N) \to 0$ cuando $N \to \infty$.

1) Sea m tal que $2^m \leq N < 2^{m+1}$. Verificar que los números binarios mayores que 2^{j-1} y estrictamente menores que 2^j tienen longitud igual a j y su número es igual a 2^{j-1} y deducir que

$$m + 1 \geq \log N \geq \lambda(N) = \frac{1}{N}(\sum_{j=1}^{m} j2^{j-1} + (m+1)(N - 2^m + 1)). \tag{7.5}$$

2) Sea $p \in \mathbb{N}$ tal que $2^{-p} < \varepsilon$. Escribir

$$\lambda(N) \geq \frac{1}{N}(\sum_{j=m-p+1}^{m} j2^{j-1} + (m+1)(N - 2^m + 1))$$

y deducir que $\lambda(N) \geq (m - p + 1)(1 - 2^{-p})$.

3) Deducir que

$$1 \geq \lim \sup_{N \to \infty} \frac{\lambda(N)}{\log N} \geq (1 - 2^{-p})$$

y concluir.

Ejercicio 7.3 La codificación de Lempel Ziv
Ziv y Lempel descubrieron un algoritmo de codificación universal y que es optimal (o sea la compresión se acerca a la entropia de la fuente).

El algoritmo es fácil de describir. Dada una sucesión de longitud n, se descompone en cadenas de 0 y 1 de longitud mínima y que no hayan aparecido antes. Así la sucesión 10110010100011110101110001101010 1101 se descompone como

$$(1)(0)(11)(00)(10)(100)(01)(111)(010)(1110)(001)(101)(010)$$

1) probar que cada una de las cadenas se escribe w0 o w1 donde w es una cadena aparecida anteriormente. Se llama $c(n)$ el número de cadenas en la sucesión.

2) Entonces se codifica la sucesión poniendo por cada cadena w0 o w1 el número de orden de la cadena anterior w que está entre 1 y $c(n)$) y su ultimo bit (0 o 1). El codigo de la sucesión tomada por ejemplo es

$(0000, 1)(0000, 0)(0001, 1)(0010, 0)(0001, 0)(0101, 0)(0010, 1)(0011, 1)(0111, 0)(1000, 0)$
$(0010, 1)(0011, 1)(0101, .)$

3) Demostrar que la longitud total de la sucesión comprimida es $c(n)(\log c(n) + 1)$ bits.

4) Escribir el algoritmo, aplicarlo a un texto largo y comparar la longitud en bits resultante con la obtenida con una codificación de Huffman.

Ejercicio 7.4 Volver a utilizar Lempel-Ziv en en modo mas intuitivo siguiente. El programa de Lempel-Ziv programa puede, con modificaciones menores, aplicarse directamente a un texto sin necesidad de conversion del fichero a binario. Quizas entonces los codigos sean mas intuitivos. En lugar de ver sucesiones de zeros y unos veriamos paquetes de letras. El algoritmo es el mismo pero cada codigo esta hecho de una sucesion de letras seguida del codigo binario indicando su posicion en el texto. Por supuesto si el texto esta convertido en ascii, eso tan solo significa que los grupos de zeros y unos del texto se agrupan por paquetes de ocho bits.

Ejercicio 7.5 Consideramos el lenguaje concebido como una cadena de Markov de orden p, o sea que $\mathbb{P}(X_n \mid X_{n-1}, \ldots, X_1) = \mathbb{P}(X_n \mid X_{n-1} \ldots, X_{n-p})$. Se supone también la estacionaridad del lenguaje, o sea que $\mathbb{P}(X_n \mid X_{n-1} \ldots, X_{n-p}) = \mathbb{P}(X_{p+1} \mid X_p \ldots X_1)$. En otros terminos la fuente de lenguaje es completamente determinada por el conocimiento de $\mathbb{P}(x_p \ldots x_1)$ por cualquier grupo de p palabras. Se define la entropia de la cadena como $H(X_p \mid X_{p-1} \ldots X_1)$.

1) Demostrar que $H(X_n, X_{n-1}, \ldots X_2, X_1) = \sum_{i=1}^{n} H(X_i \mid X_{i-1} \ldots X_1)$.

2) Utilizando la ley de los grands números, deducir que

$$\frac{1}{n} H(X_1, \ldots, X_n) \to H(X_p \mid X_{p-1} \ldots X_1).$$

7.3. Ejercios de implementacion

1) Algoritmo que calcula el codigo de Shannon asociado con una distribucion de probabilidad

2) Algoritmo que calcula el codigo de Huffman asociado con una distribucion de probabilidad

3) Aplicar esos algoritmos a textos para comprobar que la longitud del binario obtenido corresponde bien a la longitud teorica optimal n H(p), donde n es el numero de simbolos (numero de caracteres, o de digramas, o trigramas, o palabras, o pares de palabras).

4) Para poner la codificacion en un marco aplicativo realista, nos dimos cuenta de que habia que definir el codigo con un primer texto de "aprendizaje" que nos daria una distribucion de probabilidad de referencia. Ese texto tiene que ser largo o muy largo, para garantizar que la gran mayoria de los simbolos codificados ocurra varias veces y permita estimar su probabilidad.

5) Una vez establecido un codigo de referencia (por ejemplo diccionario de palabras, y diccionario de pares de palabras par el castellano, obtenido de un libro), utilizarlo para codificar OTRO texto, y ver que grado de compresion se ha obtenido.

6) Hacer experimentos de SINTESIS DE TEXTO segun una distribucion de probabilidad. El algoritmo parte de una distribucion de probabilidad para palabras y de otra para pares de palabras y genera un texto tal que la primera esta sorteada segun la distribucion de probabilidades de palabras. La segunda esta sorteda segun la distribucion de probabilidades de pares, que permite sortear una palabra nueva sabiendo la precedente (probabilidad condicional).

Ejercicio 7.6 ALGUNOS DETALLES TECNICOS sobre sintesis de textos.

a) Hay que establecer para cada codigo un diccionario
simbolo–¿codigo
y el diccionario inverso
codigo–¿simbolo

b) (Los simbolos pueden ser monogramas (letras), digramas, trigramas, palabras, o pares de palabras). No conviene ir mas alla porque las probabilidades se hacen demasiado pequeñas y no son observables.)

c) Cuando se codifica un NUEVO texto, es posible que aparezcan simbolos que NO ESTAN en el diccionario. En tal caso, el simbolo se queda tal para cual en el codigo. Asi el codigo es una sucesion de zeros, unos, y simbolos no codificados. La longitud de lo no codificado se cuenta como (numero de letras)*8, ya que cada letra se codifica "tontamente" con ocho bits.

d) En el caso de codificar por palabras o pares de palabras: los separadores (, ; . ( ) : ! ? ") se cuentan como palabras. Se negligen las mayusculas ya que despues de un punto es facil reconocer que hay mayuscula. Cada palabra empieza por un espacio.

Capítulo 8

La communicación segura es posible a pesar del ruido

8.1. Transmisión en un canal ruidoso

El problema principal que Shannon abordó y resolvió es la transmisión en un canal con ruido, o sea errores aleatorios durante la transmisión. El problema principal es decidir si la transmisión es posible y, sobre todo, a que precio en términos de redundancia. Shannon parte de sus experimentos de juventud cuando, niño en una granja americana inmensa, lograba comunicar con sus compañeros por telegrafo gracias a los alambres eléctricos de las vallas. El se dió cuenta de que un mensaje en inglés muy corrupto, donde hasta la mitad de las letras faltan, puede ser reconstruido correctamente por el receptor. Eso es debido a la redundancia del lenguaje. El destino de Shannon era hacer esa observación empirica y demostrarla años mas tarde, cuando logró calcular la entropia del inglés corriente. Se verificara que el grado de compresión de un texto de inglés corriente es del orden de cinquenta por ciento. Otra intuición que Shannon traducira en su famoso teorema: cuanto más larga la sequencia transmitida, más fácil de reconstruir.

El teorema de Shannon, a pesar de que habia sido obtenido por medios de matematica pura no constructiva, provocó el entusiasmo de los ingenieros por su simplicidad y por el desafio tecnologico que lanzaba. Shannon considera una fuente emisora X, o fuente de entrada. Pero si hay ruido en el canal la recepción esta representada por una variable aleatoria Y distinta de X que se llama la salida del canal. Para medir la relación de incertidumbre sobre X observando Y, Shannon introduce la noción de *entropia relativa*, $H(X|Y)$. Puede tambien interpretarse $H(Y|X)$: es la incertidumbre de Y conociendo X. Luego $H(Y|X)$ mide la incertidumbre creada por el ruido.

En su teorema fundamental, Shannon demuestra que es posible codificar los mensajes emitidos por X de tal manera que la proporción de errores en la transmisión sea arbitrariamente baja. En otros términos la comunicación segura es posible a pesar del ruido. Shannon define la capacidad de un canal de transmisión como $\text{máx}_X\, H(X) - H(X|Y)$ por todas las fuentes de entrada posibles. Esa capacidad es la diferencia entre la cantidad de información emitida $H(X)$ y la incertidumbre sobre lo mandado una vez observado Y, $H(X|Y)$. El gran teorema de Shannon es cuantitativo: cualquier fuente de entropia menor que la capacidad puede ser transmitida en el canal.

Capacidad de un canal con ruido. Consideremos un canal discreto, o sea un sistema disponiendo de un alphabeto de entrada $\mathcal{X}$ y de un alphabeto de salida $\mathcal{Y}$ así como de una matriz de probabilidades de transición $p(y|x)$ dando la probabilidad de observar un símbolo y como salida cuando el símbolo x estuvo emitido. Diremos que un tal canal es "sin memoria", ya que la distribución de probabilidad de la salida depende unicamente de la entrada y es independiente de las entradas y salidas precedentes. Ahora vamos a definir el *transmission rate* de la fuente X por el canal, o tasa de transmisión dadas la fuente X y su salida correspondiente Y por

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y).$$

The first defining expression has already been defined as the amount of information sent less the uncertainty of what was sent. The second measures the amount received less the part of this which is due to noise. The third is the sum of the two amounts less the joint entropy and therefore in a sense is the number of bits per second common to the two. Thus, all three expressions have a certain intuitive significance.

Definición 8.1 *Llamaremos capacidad de un canal discreto sin memoria la cantidad*

$$C = \max_{p(x)} I(X;Y),$$

donde el maximum se calcula por todas las distributiones de probabilidad posible $p(x)$, $x \in \mathcal{X}$ de entrada.

La primera cosa que hay que notar es que ese problema es un problema de optimisación en $[0,1]^{\mathrm{Card}(\mathcal{X})}$), ya que conocemos los valores de $p(y|x)$.

Proposición 8.1 *La capacidad máxima de un canal con ruido es alcanzada por una cierta fuente X de entrada*

Prueba Los datos del canal son las probabilidades condicionales $p(y \mid x)$. Como $p(x,y) = p(x)p(y \mid x)$, $p(y) = \sum_x p(y \mid x)p(x)$, vemos que $I(X;Y) = H(X) + H(Y) - H(X,Y)$ es definida como una función continua del vector $(p(x))$ definida en $[0,1]^{\mathrm{Card}(\mathcal{X})} \cap \{(p(x)) \mid \sum_x p(x) = 1\}$, que es un compacto. Luego existe al menos una distribución $p(x)$ que maximiza $I(X,Y)$. ∘

Ejemplo 1 Si el canal transmite integramente una entrada binaria sin error, la matriz de transición es la identidad. Entonces $Y = X$ et luego $I(X;X) = H(X) - H(X|X) = H(X)$. Entonces la capacidad es maximal cuando la entropia de la fuente emisora es maximal, lo que implica lo esperado $p(0) = p(1) = \frac{1}{2}$ y $C = H(p) = 1$ bit.

Ejemplo 2 : canal binario symétrico Ahora $\mathcal{X} = \mathcal{Y} = \{0,\ 1\}$ y

$$p(y=1|x=1) = p(y=0|x=0) = 1-p, \quad p(y=1|x=0) = p(y=0|x=1) = p.$$

Como la entropia de una variable de Bernouilli $\mathcal{B}(p, 1-p))$ es $H(p) = -p\log p-(1-p)\log(1-p)$, obtenemos

$$H(Y)-H(Y|X) = H(Y)-\sum p(x)H(Y|X=x) = H(Y)-\sum p(x)H(p) = H(Y)-H(p) \leq 1-H(p).$$

Habra igualdad en esa desigualdad si y solo si X es uniforme, ya que entonces Y es tambien uniforme y $H(Y) = 1$. Luego $C = 1 - H(p)$. Cuando $p = \frac{1}{2}$, la capacidad es nula y el canal no transmite nada.

Mensajes y salidas típicos para un canal con ruido (Cover-Thomas 8.6 página 195) Consideremos una fuente X con ley $p(x)$, $x \in \mathcal{X}$. Un canal de transmisión con ruido transmite X con errores. El resultado observado es Y, con ley $p(y)$, $y \in \mathcal{Y}$. La ley conjunta de X y Y se escribirá $p(x, y) = \mathbb{P}(X = x, Y = y)$, definida en $\mathcal{X} \times \mathcal{Y}$. En el caso de un canal sin ruido tendriamos $p(x, x) = p(x)$ por todo x y $p(x, y) = 0$ si $x \neq y$. Las entropias de esas variables se escribiran $H(X)$, $H(Y)$ y $H(X, Y)$. Si la fuente emite n mensajes independientes e identicamente distribuidos, el mensaje resultante visto como variable aleatoria se escribe X^n y la salida Y^n, con valores en $\mathcal{X}^n$ y $\mathcal{Y}^n$ respectivamente. Los valores posibles de X^n se escribiran $x^n \in \mathcal{X}^n$ y los de Y^n, $y^n \in \mathcal{Y}^n$. Para un sistema de comunicación con ruido, tendremos las siguientes probabilidades para los mensajes emitidos y recibidos:

$$p(x^n) = \Pi_{i=1}^n p(x_i);\ \ p(y^n) = \Pi_{i=1}^n p(y_i),\ \ p(x^n, y^n) = \Pi_{i=1}^n p(x_i, y_i)$$

utilizando la independencia de los mensajes sucesivos, y el hecho de que y_i depende tan solo de x_i y luego es independiente de los demás y_j. La primera cosa que hay que notar es que si el canal emite una sola vez, no hay nada que hacer para corregir errores. Por ello, vamos a intentar utilizar mensajes repitidos, que nos permitiran de nuevo recurrir a la noción de mensajes típicos.

Definición 8.2 *Llamaremos conjunto de* pares entrada-salida típicos relativemente a la distribución $p(x, y)$ *el conjunto* A_ε^n *de las sucesiones* $\{(x^n, y^n)\}$ *cuyas probabilidades son típicas en el sentido siguiente:* $A_\varepsilon^n = B_\varepsilon^n \cap C_\varepsilon^n \cap D_\varepsilon^n$ *con*

$$B_\varepsilon^n = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |-\frac{1}{n}\log p(x^n, y^n) - H(X, Y)| < \varepsilon\}; \tag{8.1}$$

$$C_\varepsilon^n = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |-\frac{1}{n}\log p(x^n) - H(X)| < \varepsilon\}; \tag{8.2}$$

$$D_\varepsilon^n = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |-\frac{1}{n}\log p(y^n) - H(Y)| < \varepsilon\}. \tag{8.3}$$

Lemma 8.1 *Sea* (X^n, Y^n) *una sucesión de longitud* n *de v.a.i.i.d. según la ley* $p(x^n, y^n) = \Pi_{i=1}^n p(x_i, y_i)$. *Entonces*

1. $\mathbb{P}((X^n, Y^n) \in A_\varepsilon^n) \to 1$ *quand* $n \to \infty$.
2. *Card*$(\{y^n,\ (x^n, y^n) \in A_\varepsilon^n\}) \leq 2^{n(H(Y|X)+2\varepsilon)}$
3. *Card*$(\{x^n,\ (x^n, y^n) \in A_\varepsilon^n\}) \leq 2^{n(H(X|Y)+2\varepsilon)}$

La segunda relación estima el número de mensajes y^n que pueden ser producidos por x^n. La tercera estima el número de mensajes x^n que han podido producir una cierta salida y^n. En la prueba que sigue y en el resto de ese capítulo escribiremos $2^{a\pm\varepsilon}$ para indicar un número cualquiera b tal que $2^{a-\varepsilon} \leq b \leq 2^{a+\varepsilon}$, de manera que por ejemplo $2^{a\pm\varepsilon} \times 2^{b\pm\varepsilon} = 2^{a+b\pm 2\varepsilon}$.

Prueba Vamos a probar la primera relación. Consideremos $\pi_1 : (x^n,\ y^n) \to x^n$ et $\pi_2 : (x^n,\ y^n) \to y^n$ las aplicaciones de proyección de $\mathcal{X}^n \times \mathcal{Y}^n$ sobre $\mathcal{X}^n$ y $\mathcal{Y}^n$ respectivamente. El resultado del lema 7.1 permite afirmar que cuando $n \to \infty$,

$$\mathbb{P}(B_\varepsilon^n) \to 1 \tag{8.4}$$
$$\mathbb{P}(C_\varepsilon^n) = \mathbb{P}(\pi_1(C_\varepsilon^n) \times Y^n) = \mathbb{P}(\pi_1(C_\varepsilon^n)) \to 1 \tag{8.5}$$
$$\mathbb{P}(D_\varepsilon^n) = \mathbb{P}(X^n \times \pi_2(D_\varepsilon^n)) = \mathbb{P}(\pi_2(D_\varepsilon^n)) \to 1. \tag{8.6}$$

La relación (8.4) sigue entonces de la observación general que si sucesiones de conjuntos B_i^n, $i = 1, \ldots,\ k$ verifican $\mathbb{P}(B_i^n) \to 1$ cuando $n \to \infty$, entonces $\mathbb{P}(\cap_{i=1}^k B_i^n) \to 1$ tambien. Se aplica esa observación a la intersección de los tres conjuntos precedentes,

$$A_n^\varepsilon = B_\varepsilon^n \cap C_\varepsilon^n \cap D_\varepsilon^n.$$

Ahora pasemos a la segunda relación del lema. Por hipotesis $\mathbb{P}(X^n = x^n) = p(x^n) = 2^{-n(H(X)\pm\varepsilon)}$ y $\mathbb{P}((X^n, Y^n) = (x^n, y^n)) = p(x^n, y^n) = 2^{-n(H(X,Y)\pm\varepsilon)}$. Luego

$$p(x^n) = \sum_{y^n} p(x^n, y^n) \geq \sum_{y^n,\ (x^n,y^n)\in A_\varepsilon^n} p(x^n, y^n) \geq \mathrm{Card}(\{y^n,\ (x^n, y^n) \in A_\varepsilon^n\}) \inf_{(x^n,y^n)\in A_\varepsilon^n} p(x^n, y^n).$$

Así que

$$\mathrm{Card}(\{y^n,\ (x^n, y^n) \in A_\varepsilon^n\}) \leq 2^{-n(-H(X,Y)+H(X)+2\varepsilon)} = 2^{n(H(Y|X)+2\varepsilon)}.$$

∘

8.2. El teorema fundamental para un canal discreto con ruido

Teorema 8.1 *(Shannon páginas 71, 72, 73) Consideremos un canal discreto de capacidad C y una fuente discreta X de entropia $H(X)$.*
(a) Si $H(X) \leq C$ existe un sistema de codificación tal que la salida de la fuente se transmita por el canal con una frecuencia arbitrariamente pequeña de error.
(b) Si $H(X) > C$ es todavia posible codificar la fuente de manera que la incertitumbre sobre los mensajes $H(X|Y)$ sea inferior a $H(X) - C + \varepsilon$, donde ε es arbitrariamente pequeño.
(c) No hay método de codificación que permita alcanzar una incertidumbre sobre los mensajes, $H(X|Y)$, inferior a $H(X) - C$.

Prueba (a) Consideremos una fuente X_0 de tasa de transmisión realizando la capacidad maximal C, y sea Y_0 su salida. Vamos a utilizar X_0 como entrada en el canal. Consideremos todas las sucesiones posibles transmitidas y recibidas, de duración n. En todo lo que sigue, los ε, η, θ seran reales positivos que podemos fijar arbitrariamente pequeños con tal de que la duración de transmisión n sea suficientemente grande. Cada vez que tengamos $C\varepsilon$ con C independiente de n, seguiremos escribiendo ε en lugar de $C\varepsilon$, para evitar notaciones pesadas.

Sea A_ε^n el conjunto de los pares típicos (x^n, y^n). (Definición 8.2). Tenemos la situación siguiente si lanzamos X_0 a emitir por el canal (vea el esquema de la figura 8.1):

1. Las sucesiones transmitidas van a estar en dos grupos : los mensajes típicos de X_0, cuyo número es $2^{n(H(X_0)\pm\varepsilon)}$ y los demas cuya probabilidad total es inferior a η.

2. De la misma manera, las sucesiones recibidas de X_0 son como emitidas por la fuente Y_0 y forman un conjunto de sucesiones típicas de numero $2^{n(H(Y_0)\pm\varepsilon)}$ y de probabilidad total superior a $1-\eta$. Vamos a llamar $\mathcal{M}_0$ ese conjunto de mensajes típicos.

3. Por el lema 8.1, propiedad 3, cada salida típica ha podido ser producida por no mas de $2^{n(H(X_0|Y_0)+\varepsilon)}$ entradas.

4. Finalmente, deducemos que cada entrada típica de X_0 puede producir todo lo mas $2^{n(H(Y_0|X_0)\pm\varepsilon)}$ salidas (pero no vamos a utilizar esa ultima propiedad.)

Ahora bien. Consideremos una fuente X de entropia $H(X) < C$. Escribamos $H(X) = C - \theta$. Como X_0 satisface $H(X_0) - H(X_0|Y_0) = C$, tenemos

$$H(X) - (H(X_0) - H(X_0|Y_0)) < -\theta. \tag{8.7}$$

En un tiempo de transmisión n, la fuente X puede producir $2^{n(H(X)\pm\varepsilon)}$ mensajes típicos. Vamos a llamar $\mathcal{M}$ ese conjunto de mensajes típicos y los vamos a codificar asociandolos con mensajes típicos de longitud n de la fuente X_0 que seran utilizados como códigos. Una codificación se definira como una application $\mathcal{C} : \mathcal{M} \to \mathcal{M}_0$ obtenida sorteando (con distribución uniforme) para cada mensaje en $\mathcal{M}$ un elemento de $\mathcal{M}_0$ que es eligido como su código. Los demas mensajes, no típicos, no tienen código, lo que nos da una probabilidad de error mas pequeña que η.

El algoritmo de decodificación asociado es el siguiente: si llega y_1^n, mirar cual de los x^n entre los mensajes típicos de $\mathcal{M}$ podria haber producido y_1^n. La cuestión es evaluar con que probabilidad hay un error, o sea con que probabilidad hay más de un código x^n que pueda producir y_1^n.

En otros términos: Supongamos que una salida y_1 es observada. Vamos a evaluar la probabilidad de error, o sea la probabilidad de que y_1 haya sido asociada con mas de un mensaje de $\mathcal{M}$. Por la observación (3), sabemos que y_1 pudo ser producida por no mas de $2^{n(H(X_0|Y_0)+\varepsilon)}$ mensajes x_0 en $\mathcal{M}_0$. Ademas, la probabilidad de que cada $x_0 \in \mathcal{M}_0$ sea un código es $2^{n(H(X)-H(X_0)\pm 2\varepsilon)}$. En efecto hemos distribuido $2^{n(H(X)\pm\varepsilon)}$ mensajes tipicos de $\mathcal{M}$ uniformemente en $2^{n(H(X_0)\pm\varepsilon)}$ códigos. Eso implica que la probabilidad de que y_1 sea el código de otro mensaje de X (ademas del mensaje que lo provocó) verifica

$$\mathbb{P}(\text{error sobre } y_1) \leq 2^{n(H(X)-H(X_0)\pm 2\varepsilon)} 2^{n(H(X_0|Y_0)+\varepsilon)}.$$

Luego por (8.7),

$$\mathbb{P}(\text{error sobre } y_1) \leq 2^{n(H(X)-H(X_0)+\varepsilon+H(X_0|Y_0)+2\varepsilon)} = 2^{-n(\theta-3\varepsilon)}$$

Como η ha sido fijado (arbitrariamente pequeño) y, η una vez fijado, podemos eligir ε tan pequeño como querramos para n grande, deducimos que la probabilidad de error para cada mensaje es arbitrariamente pequeña, lo que prueba (a) para $H(X) < C$.

(b) Si $H(X) \geq C$, se puede lo mismo aplicar la construcción precedente pero no podemos codificar mas de $2^{n(C-\varepsilon)}$ mensajes de los $2^{n(H(X)\pm\varepsilon)}$ típicos. Los demas no se transmiten, lo que no deja de perder interés, ya que la mayoria de los mensajes típicos no seran transmitidos!.

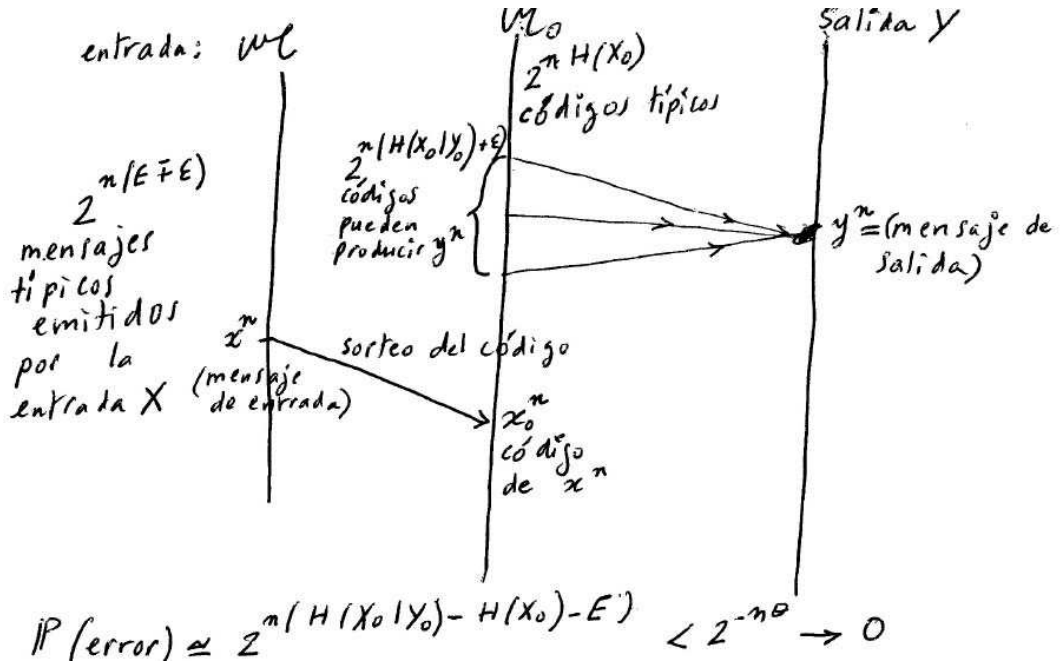

Figura 8.1: Prueba del teorema fundamental de Shannon (E denotes $H(X)$)

(c) Supongamos que se pueda transmitir por un canal de capacidad C los mensajes de una fuente X de entropia $E = H(X) + a$, con $a > 0$ y que la incertidumbre sobre el mensaje verifique $H(Y|X) = a - \varepsilon$ con $\varepsilon > 0$. Entonces $H(X_0) - H(X_0|Y_0) = C + \varepsilon$, lo que contradice la definición de C como el máximo de $H(X_0) - H(X_0|Y_0)$ por todas las fuentes de entrada.

Una pregunta final.
Explicar el razonamiento con que Shannon concluye su prueba:

Actually more has been proved than was stated in the theorem. If the average of a set of positive numbers is within ε of zero, a fraction of at most $\sqrt{\varepsilon}$ can have values greater than $\sqrt{\varepsilon}$. Since ε is arbitrarily small we can say that almost all the systems are arbitrarily close to the ideal.

En otros términos casí todos los códigos considerados, eligidos al azar, son códigos que corrigen espontaneamente los errores!

Entonces, porque sigue siendo dificil en la práctica concebir un código corrector de errores?

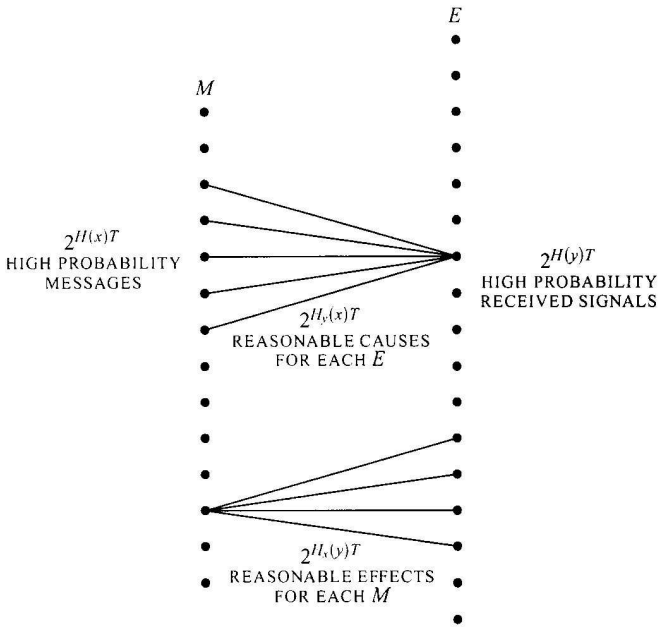

Figura 8.2: El esquema original de Shannon

Consejos de lectura :
Claude E. Shannon et Warren Weaver The mathematical theory of communication University of Illinois Press 1998.
Thomas M. Cover et Joy A. Thomas Elements of Information Theory, Wiley Series Telecommunications, (capítulos 2, 5 et 8), 1991.
Pierre Brémaud Introduction aux probabilités, Springer.