UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

Tema 3. Codificación económica

José A. Montenegro

Dpto. Lenguajes y Ciencias de la Computación
ETSI Informática. Universidad de Málaga
monte@lcc.uma.es twitter

26 de septiembre de 2013

El concepto de fuente

- Básicamente podemos definir como una "**fuente**" como un productor de mensajes.
  - Los ejemplos de la vida real van desde un persona escribiendo correos, o un escáner digitalizando imágenes.
- Utilizando la terminología anteriormente descrita, una fuente emite una cadena de símbolos, denotado por:

$$\xi_1\xi_2\xi_3\ldots.$$

Cada ξ_k es una variable que puede tomar como valor cualquier símbolo que pertenece a un alfabeto dado.

- Una característica que define una fuente y la diferencia de otra es el alfabeto.
  - Idiomas orientales utilizan alfabetos distintos al alfabeto de 28 letras $\mathbb{A}$.
- Muchos comparten el alfabeto de 28-27 letras similar a $\mathbb{A}$, existen ciertas reglas de como el alfabeto es utilizado (gramática, sintaxis ...) y esos elementos están reflejados en la cadena emitida por la fuente.
  - El símbolo W aparece mucho más en el Inglés que en otro idioma.
  - Escáner, producirá más a menudo mensajes en el que W (representa un pixel blanco) que B (representa un pixel negro).

Ejemplos Lenguajes Naturales

- Como conclusión tenemos que la fuente debe ser descrita especificando el alfabeto y las probabilidades de los símbolos que van a ser emitidos.

Símbolo	Español	Portugues	Italiano	Inglés
a	12.53 %	14.63 %	11.74 %	8,167 %
b	1.42 %	1.04 %	0.92 %	1,492 %
c	4.68 %	3.88 %	4.5 %	2,782 %
d	5.86 %	4.99 %	3.73 %	4,253 %
e	13.68 %	12.57 %	11.79 %	12,702 %
f	0.69 %	1.02 %	0.95 %	2,228 %
g	1.01 %	1.30 %	1.64 %	2,015 %
h	0.70 %	1.28 %	1.54 %	6,094 %
i	6.25 %	6.18 %	11.28 %	6,966 %
j	0.44 %	0.40 %	0.00 %	0,153 %
k	0.01 %	0.02 %	0.00 %	0,772 %
. . .	. . .	. . .	. . .	. . .
q	0.88 %	1.20 %	0.51 %	0,095 %
r	6.87 %	6.53 %	6.37 %	5,987 %
s	7.98 %	7.81 %	4.98 %	6,327 %
t	4.63 %	4.74 %	5.62 %	9,056 %
u	3.93 %	4.63 %	3.01 %	2,758 %
v	0.90 %	1.67 %	2.10 %	0,978 %
w	0.02 %	0.01 %	0.00 %	2,36 %
x	0.22 %	0.21 %	0.00 %	0,15 %
y	0.90 %	0.01 %	0.00 %	1,974 %
z	0.52 %	0.47 %	0.49 %	0,074 %

Ejemplos Moneda

- Experimento de una moneda es lanzada repetidamente, resultado $H = Cara$, $T = Cruz$. Una fuente con alfabeto S = {H,T}
- Si la moneda no está trucada, la cadena emitida podría ser

$$HTTHTHTHTTHHTTTHTHHTTHTHTHTH\ldots,$$

  donde H y T ocurren con la misma frecuencia.
- La probabilidad que ξ_k sea H y la probabilidad que ξ_k sea T, es la misma $1/2$. Lo representaríamos como:

$$Pr(\xi_k = H) = Pr(\xi_k = T) = 0{,}5$$

- Por otro lado, una moneda trucada podría emitir un cadena como :

$$HHHHHHHTHHHHHTHHHHHHHHHHT\ldots$$

  donde las probabilidades son tales que $Pr(\xi_k = H)$ es significativamente mayores que $Pr(\xi_k = T)$.

Distribución de Probabilidad

La teoría necesaria para establecer un modelo matemático completo de la fuente, es bastante complejo pero podemos realizar algunos progresos basándonos en un modelo simple.

Definición 1 (Distribución de Probabilidad)

Sea $S = \{s_1, s_2, ..., s_m\}$ *un alfabeto. Una distribución de probabilidad sobre S es un conjunto de números reales* $p_1, p_2, ..., p_m$ *tal que*

$$p_1 + p_2 + ... + p_m = 1, 0 \leq p_i \leq 1 (i = 1, 2, ..., m).$$

Las probabilidades pueden ser descritas como un vector $\mathbf{p} = [p_1, p_2, ..., p_m]$*, y establece la relación que el símbolo* s_i *aparece con probabilidad* p_i*.*

Fuente con probabilidad de distribución

- Dada una distribución de probabilidad **p** en *S*, consideramos una fuente con la siguiente propiedad:
  - La fuente emite una cadena $\xi_1\xi_2\xi_3\ldots$, donde el valor de cada ξ_k es un miembro del alfabeto *S*, y para toda k, la probabilidad que ξ_k tome un valor determinado s_i es dado por la ecuación:

$$Pr(\xi_k = s_i) = p_i$$

  - Donde ξ_k es una *variable aleatoria*.
- Nos referimos a este modelo como una "fuente con probabilidad de distribución **p**" o simplemente una "fuente (S,**p**)" .
- Esta definición no implica que todas las características de la fuente son descritas mediante **p**. Representa que la probabilidad de distribución para cada ξ_k es la misma.

- En muchos casos esta es una suposición razonable, aunque la fuente puede tener otras características.
  - Por ejemplo, suponemos que el texto de un libro es una cadena de símbolos del alfabeto $\mathbb{A}$.
  - Si abrimos el libro en la página x y buscamos el y ésimo símbolo en esa página, entonces la probabilidad que el símbolo sea L puede ser asumido que es la misma, para toda x e y.
  - Pero si sabemos que la letra 17 en la página 83 es una L, entonces afectará a la probabilidad que la letra 18 de la página 83 sea una A.
- Imponiendo que las variables aleatorias ξ_k tienen la misma distribución $\mathbf{p}$, pero en general, no asumimos que esas variables aleatorias son $independientes$.

Fuente sin memoria

Definición 2 (Fuente sin memoria)

*Una fuente (S,**p**) que emite una cadena $\xi_1\xi_2\xi_3\ldots$ es sin memoria si las variables aleatorias ξ_k son independientes. O lo que es lo mismo, para todos k y l*

$$Pr(\xi_k = s_i \text{ y } \xi_l = s_j) = Pr(\xi_k = s_i)Pr(\xi_l = s_j)$$

- En una fuente sin memoria, el conocimiento de algunas partes del mensaje no afecta a la distribuciones de probabilidades asignadas a los otros términos.
- Esta situación ocurre en el experimento realizado al lanzar una moneda al aire y otras situaciones en la vida real.

Ejemplos

Ejemplo 1

Una de las apuestas más conocidas en España es la quiniela, que puede ser representada como una fuente que emite una cadena de símbolos del alfabeto S = {1,x,2}, donde 1 = gana casa, x = empate, 2 = gana foráneo. ¿Podría ser considerado este ejemplo como una fuente sin memoria?

Ejemplos

Ejemplo 1

Una de las apuestas más conocidas en España es la quiniela, que puede ser representada como una fuente que emite una cadena de símbolos del alfabeto S = {1,x,2}, donde 1 = gana casa, x = empate, 2 = gana foráneo. ¿Podría ser considerado este ejemplo como una fuente sin memoria?

Solución:
Es razonable asumir que la fuente es sin memoria, ya que el resultado de un partido de fútbol no afecta al resultado de los otros partidos. Un estudio en los resultados obtenidos es que la probabilidad de distribución es aproximadamente:

$$p_1 = 0{,}42, p_x = 0{,}26, p_2 = 0{,}31.$$

Ejercicio 1

Considere una fuente que emite símbolos del alfabeto S con probabilidad de distribución p, donde $S = \{a, b, c\}$ $p = [p_a, p_b, p_c] =$ *[0.6,0.3,0.1].*
Si la fuente emite una cadena de 100 símbolos, aproximadamente ¿Cuántas veces aparecerá el carácter a? Si la fuente es sin memoria, aproximadamente ¿Cuántas veces aparecerá el par ab ?

Ejercicio 1

Considere una fuente que emite símbolos del alfabeto S con probabilidad de distribución p, donde $S = \{a, b, c\}$ $p = [p_a, p_b, p_c] =$ *[0.6,0.3,0.1].*
Si la fuente emite una cadena de 100 símbolos, aproximadamente ¿Cuántas veces aparecerá el carácter a? Si la fuente es sin memoria, aproximadamente ¿Cuántas veces aparecerá el par ab ?

Solución:
a) 60, b)18

El problema de optimización

- Consideremos que ocurre cuando la fuente (S,p) emite una cadena y es transformada a una cadena codificada utilizando un alfabeto T.
  - ¿En que grado depende la longitud de un mensaje del código utilizado $c : S \to T^*$?
- Partimos $S = \{s_1, s_2, \ldots, s_m\}$.
  - Sea y_i la longitud de la palabra codificada $c(s_i)(1 \leq i \leq m)$, y consideramos un mensaje en S de longitud N.
  - Si N es un número razonadamente grande, el mensaje contendrá el símbolo s_1 aproximadamente Np_1 veces, s_2 unas Np_2 veces ....
  - Después de codificar aplicando la función c, tendremos Np_1 cadenas $c(s_1)$, cada una de longitud y_1, Np_2 cadenas $c(s_2)$, cada una de longitud y_2 y así con todas.

Longitud media de palabras

- La longitud total del mensaje codificado es aproximadamente:

$$Np_1y_1 + Np_2y_2 + \ldots Np_my_m = N(p_1y_1 + p_2y_2 + \ldots + p_my_m).$$

- Por tanto, si el mensaje original tiene longitud N, el mensaje codificado tiene una longitud de LN, donde:

$$L = p_1y_1 + p_2y_2 + \ldots + p_my_m.$$

Definición 3 (Longitud media de palabras)

La longitud media de palabras de un código $c : S \to T^*$ para la fuente $(S, \mathbf{p})$ es

$$L = p_1y_1 + p_2y_2 + \ldots + p_my_m.$$

Ejemplo 2

Sea $S = \{s_1, s_2, s_3\}$ y p = [0.2, 0.6, 0.2]. Debemos encontrar el valor de L cuando utilizamos el código binario:

$$s_1 \mapsto 0, s_2 \mapsto 10, s_3 \mapsto 11$$

Además, ¿Es el código binario asignado para S el que alcanza su menor valor?

Ejemplo 2

Sea $S = \{s_1, s_2, s_3\}$ y p = [0.2, 0.6, 0.2]. Debemos encontrar el valor de L cuando utilizamos el código binario:

$$s_1 \mapsto 0, s_2 \mapsto 10, s_3 \mapsto 11$$

Además, ¿Es el código binario asignado para S el que alcanza su menor valor?

Solución:
El valor de L es: 1 x 0.2 + 2 x 0.6 + 2 x 0.2 = 1.8.
Ya que s_2 es el símbolo más utilizado, podríamos dar una codificación que utilice la palabra codificada más corta para este símbolo, tal que

$$s_1 \mapsto 10, s_2 \mapsto 0, s_3 \mapsto 11$$

Donde el valor de L es: 2 x 0.2 + 1 x 0.6 + 2 x 0.2 = 1.4.

Código Optimo

- En el ejemplo anterior ha sido fácil probar que existe un código mejor que el original, pero tenemos un problema mayor que resolver.
- Dado una fuente (S,**p**) y el alfabeto T, ¿Cómo podemos encontrar un código $c : S \to T^*$ para el cual L es el mínimo posible?

Definición 4 (Código optimo)

*Dado una fuente (S,**p**) y un alfabeto T, un código UD $c : S \to T^*$ es óptimo si no existe un código con longitud media de palabra menor.*

- El requisito que un código sea UD es una restricción significativa.
- En el tema anterior mostramos que esta restricción puede ser expresada por la condición $K \leq 1$, donde K es el número Kraft-McMillan asociado con los parámetros $n_1, n_2, ..., n_M$ y la base b del código:
  $K = n_1/b + n_2/b^2 + \ldots + n_M/b^M$

- En nuestra notación, n_i es el número de símbolos s_j tal que $y_j = i$, y M es el máximo valor de y_j.
- El término $\frac{n_i}{b^i}$ en K es la suma de n_i términos $\frac{1}{b^i}$, correspondientes a un término $\frac{1}{b^{y_j}}$ para cada j tal que $y_j = i$.
  - De esta forma K puede ser escrito como la suma de todos los términos $\frac{1}{b^{y_j}}$:

$$K = \frac{n_1}{b^{y_1}} + \frac{n_2}{b^{y_2}} + \ldots + \frac{n_M}{b^{y_m}}$$

- Reescribiendo la condición $K \leq 1$ en esta forma podemos formular el problema de encontrar los códigos óptimos como sigue:
  - Dado b y $p_1, p_2, \ldots, p_m$ encontrar $y_1, y_2, \ldots, y_m$ que

$$\text{minimizar } p_1 y_1 + p_2 y_2 + \ldots + p_m y_m$$

$$\text{sujeto a } \frac{1}{b^{y_1}} + \frac{1}{b^{y_2}} + \ldots + \frac{1}{b^{y_m}} \leq 1.$$

$$y_1, y_2, \ldots, y_m \geq 0$$

Ejercicio 2

Una fuente emite tres símbolos con probabilidades 0.5, 0.25, 0.25. Construya un código binario PF para esta fuente con un tamaño medio de palabra de 1.5.

Ejercicio 2

Una fuente emite tres símbolos con probabilidades 0.5, 0.25, 0.25. Construya un código binario PF para esta fuente con un tamaño medio de palabra de 1.5.

Solución:

Por ejemplo: $s_1 \mapsto 0, s_2 \mapsto 10, s_3 \mapsto 11$

$$\text{L} = 1\text{x}0{,}5 + 2 \text{ x } 0{,}25 + 2 \text{ x } 0{,}25 = 1{,}5$$

Entropía

Definición 5 (Entropía de una distribución)

La entropía base b de una distribución de probabilidad $p = [p_1, p_2, ..., p_m]$ *es*

$$H_b(p) = H_b(p_1, p_2, \ldots, p_m) = \sum_{i=1}^{m} p_i log_b(\frac{1}{p_i}).$$

Si $0 < p_i < 1$ *entonces* $\frac{1}{p_i} > 1$, *y cada término* $p_i log_b(\frac{1}{p_i})$ *es positivo.*

- Ya que la salida de una fuente sin memoria (S,**p**) esta determinada por **p**, podemos hablar a menudo de la entropía de **p** como la entropía de la fuente.
- Sin embargo, muchas fuentes son con memoria, y por tanto necesitamos una definición de entropía más sofisticada que veremos más adelante.

Ejemplo 3

¿Cual es la entropía base 2 de una fuente sin memoria con distribución $\mathbf{p} = [0{,}5, 0{,}25, 0{,}25]$*?*

Ejemplo 3

¿Cual es la entropía base 2 de una fuente sin memoria con distribución $\mathbf{p} = [0{,}5, 0{,}25, 0{,}25]$*?*

Solución:

$$H_2(p_1, p_2, p_3) = p_1 \times log_2(1/p_1) + p_2 \times log_2(1/p_2) + p_3 \times log_2(1/p_3)$$
$$= 0{,}5 \times log_2(1/0{,}5) + \times 0{,}25 log_2(1/0{,}25) + 0{,}25 \times log_2(1/0{,}25)$$
$$= 0{,}5 \times log_2 2 + 0{,}25 \times log_2 4 + 0{,}25 \times log_2 4$$
$$= 0{,}5 \times 1 + 0{,}25 \times 2 + 0{,}25 \times 2$$
$$= 1{,}5.$$

Ejercicio 3

¿Cuál es la entropía base 2 de una fuente sin memoria que emite 5 letras A,E,I,O,U con probabilidades 0.2, 0.3, 0.2, 0.2, 0.1? ¿Cuál sería la entropía si las 5 letras son igualmente probables?

Ejercicio 3

¿Cuál es la entropía base 2 de una fuente sin memoria que emite 5 letras A,E,I,O,U con probabilidades 0.2, 0.3, 0.2, 0.2, 0.1? ¿Cuál sería la entropía si las 5 letras son igualmente probables?

Solución:

$H_2(p_1, p_2, p_3, p_4, p_5)$
$= p_1 log_2(1/p_1) + p_2 log_2(1/p_2) + p_3 log_2(1/p_3) + p_4 log_2(1/p_4) + p_5 log_2(1/p_5)$
$= 0{,}2 log_2(1/0{,}2) + 0{,}3 log_2(1/0{,}3) + 0{,}2 log_2(1/0{,}2) + 0{,}2 log_2(1/0{,}2) + 0{,}1 log_2(1/0{,}1)$
$= 0{,}2 log_2 5 + 0{,}3 log_2 3{,}33 + 0{,}2 log_2 5 + 0{,}2 log_2 5 + 0{,}1 log_2 10$
$= 0{,}2 \times 2{,}32 + 0{,}3 \times 1{,}74 + 0{,}2 \times 2{,}32 + 0{,}2 \times 2{,}32 + 0{,}1 \times 3{,}32$
$= 0{,}46 + 0{,}52 + 0{,}46 + 0{,}46 + 0{,}33 = 2{,}23$

Solución:

$H_2(p_1, p_2, p_3, p_4, p_5)$
$= p_1 log_2(1/p_1) + p_2 log_2(1/p_2) + p_3 log_2(1/p_3) + p_4 log_2(1/p_4) + p_5 log_2(1/p_5)$
$= 0{,}2 log_2(1/0{,}2) + 0{,}2 log_2(1/0{,}2) + 0{,}2 log_2(1/0{,}2) + 0{,}2 log_2(1/0{,}2) + 0{,}2 log_2(1/0{,}2)$
$= 2{,}32$

Ejercicio 4

¿Cuál es la entropía de una fuente que emite un solo y específico símbolo?

Ejercicio 4

¿Cuál es la entropía de una fuente que emite un solo y específico símbolo?

Solución:

$H_2(p_1) = p_1 \times log_2(\frac{1}{p_1}) = 1 \times log_2(1/1) = 1 \times 0 = 0$

Ejercicio 4

¿Cuál es la entropía de una fuente que emite un solo y específico símbolo?

Solución:
$H_2(p_1) = p_1 \times log_2(\frac{1}{p_1}) = 1 \times log_2(1/1) = 1 \times 0 = 0$

Ejercicio 5

¿Cual es la entropía de una fuente sin memoria que emite símbolos de un alfabeto de tamaño m, siendo cada símbolo igualmente probable?

Ejercicio 4

¿Cuál es la entropía de una fuente que emite un solo y específico símbolo?

Solución:

$H_2(p_1) = p_1 \times log_2(\frac{1}{p_1}) = 1 \times log_2(1/1) = 1 \times 0 = 0$

Ejercicio 5

¿Cual es la entropía de una fuente sin memoria que emite símbolos de un alfabeto de tamaño m, siendo cada símbolo igualmente probable?

Solución:

$H_2(p) = m \times (p \times log_2(1/p)) = m \times (\frac{1}{m} \times log_2(\frac{1}{\frac{1}{m}})) = m(\frac{1}{m} log_2(m)) = log_2(m)$

Representación gráfica

- Hasta ahora la definición de la entropía no está contextualizada. ¿Qué utilidad tiene? ¿Cuál es el rol de la base b? ¿Como la entropía es relevante al problema de la codificación óptima?
- Para contestar estas preguntas, consideramos primero un ejemplo muy simple, una fuente sin memoria emite los símbolos 0 y 1.
  - Las probabilidades p_0 y p_1 pueden ser escritas como x y $1-x$ para cualquier x en el intervalo $0 \leq x \leq 1$.
  - La entropía de la fuente (en base 2) es

$$h(x) = x \times log_2(1/x) + (1-x) \times log_2(1/(1-x)).$$

- La figura 1 es la gráfica de h para los valores $0 \leq x \leq 1$.
  - Como podemos observar es simétrica sobre la línea $x = 1/2$ y el valor máximo es $h(1/2) = 1$.

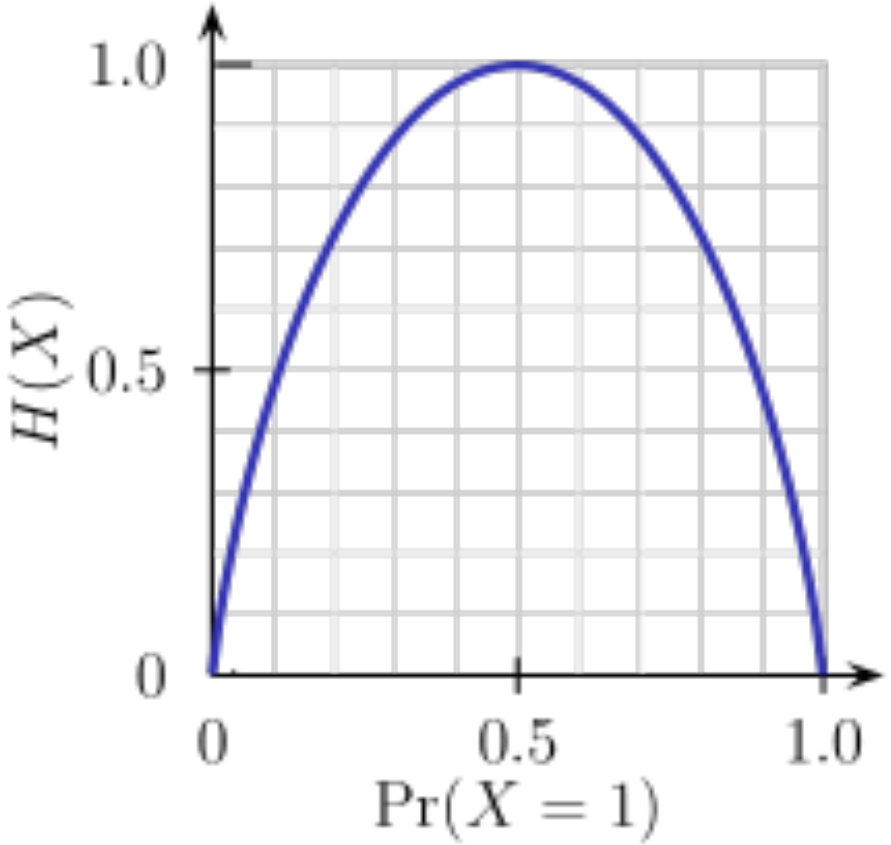

Figura 1 : Representación gráfica de h(x)

Incertidumbre

- Una posible interpretación que nos proporciona la representación gráfica es que la entropía es una medida de **incertidumbre**.
- La mayor incertidumbre sobre cual símbolo es emitido ocurre cuando dos símbolos son igualmente probables $(x = \frac{1}{2})$, por otro lado no existe incertidumbre si ninguno de los símbolos son emitidos (x=0 ó x=1).
- A modo de resumen, podemos establecer que la entropía es una medida de incertidumbre sobre la identidad del símbolo que es obtenido acorde con una distribución de probabilidad **p**.

El concepto de información está muy vinculado a la incertidumbre. Básicamente, proporcionar información sobre un evento reduce nuestra incertidumbre sobre él.

Teorema 1 (El teorema de comparación)

Si $p_1, p_2, \ldots, p_m$ *y* $q_1, q_2, \ldots, q_m$ *son dos distribuciones de probabilidad entonces:*

$$H_b(p) = \sum_{i=1}^{m} p_i log_b(\frac{1}{p_i}) \leq \sum_{i=1}^{m} p_i log_b(\frac{1}{q_i})$$

El caso de igualdad es dado sii $q_i = p_i$ *para todos los i* $(1 \leq i \leq m)$.

Teorema 2

La entropía (incertidumbre) de una distribución p en m símbolos es al menos $log_b m$. *El máximo valor ocurre sii todos los símbolos son equiprobables.*

Ejercicio 6

Suponemos una fuente sin memoria que emite tres símbolos a,b,c con probabilidades 0.6, 0.3, 0.1, y otra fuente sin memoria emite los mismos símbolos, con probabilidades 0.5, 0.3, 0.2. ¿Qué fuente presenta mayor incertidumbre? ¿Qué distribución de probabilidad produciría la mayor incertidumbre?

Ejercicio 6

Suponemos una fuente sin memoria que emite tres símbolos a,b,c con probabilidades 0.6, 0.3, 0.1, y otra fuente sin memoria emite los mismos símbolos, con probabilidades 0.5, 0.3, 0.2. ¿Qué fuente presenta mayor incertidumbre? ¿Qué distribución de probabilidad produciría la mayor incertidumbre?

Solución:

a) $H_2(p) = 0{,}6log_2(1/0{,}6) + 0{,}3log_2(1/0{,}3) + 0{,}1log_2(1/0{,}1)$
$= 0{,}6 \times 0{,}74 + 0{,}3 \times 1{,}73 + 0{,}1 \times 3{,}32$
$= 0,44 + 0,52 + 0,33 = 1{,}29$

b) $H_2(p) = 0{,}5log_2(1/0{,}5) + 0{,}3log_2(1/0{,}3) + 0{,}2log_2(1/0{,}2)$
$= 0{,}5 + 0{,}3x1{,}73 + 0{,}2x2{,}32 = 0,5 + 0,52 + 0,46 = 1{,}48$

c) La incertidumbre es mayor cuando todos los símbolos son igualmente probable, siendo la distribución de probabilidad $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. $H_2(p) = log_2(m = 3) = 1,58$.

Códigos Óptimos

- Después de las definiciones planteadas, volvemos a una pregunta que teníamos sin resolver: ¿Cómo encontrar un código b-ario UD para una fuente (S,p) tal que la longitud media de palabra L es mínima? .
- Para realizarlo hacemos uso del concepto de entropía $H_b(p)$ y los siguientes teoremas:

Teorema 3

La longitud media de cualquier código UD b-ario para la fuente (S,p) verifica $L \geq H_b(p)$.

Definición 6 (Regla de Shannon-Fano (SF))

Podemos intentar construir un código con una media de longitud de palabra cercana a L, escogiendo b^{y_i} sea tan cercano como sea posible a $\frac{1}{p_i}$. O lo que es lo mismo:

$$y_i \text{ es el menor entero positivo tal que } b^{y_i} \geq \frac{1}{p_i}.$$

El valor resultante de K es:

$$K = 1/b^{y_1} + 1/b^{y_2} + \ldots + 1/b^{y_m} \leq p_1 + p_2 + \ldots + p_m = 1$$

y por tanto un código PF (por teorema 2.2) con estos parámetros debe existir.

El siguiente teorema muestra que la longitud media L de un código esta muy próxima a $H_b(p)$.

Teorema 4

Existe un código PF b-ario para una fuente con una probabilidad de distribución **p** *que verifica la inegualdad* $L < H_b(p) + 1$.

Ejemplo 4

Una fuente emite 5 símbolos con una probabilidad de distribución $p = [0{,}4, 0{,}2, 0{,}2, 0{,}1, 0{,}1]$. *¿Cual es la entropía H(p)? Construya un código binario para esta fuente con longitud media de palabra menor que* $H(p) + 1$.

Ejemplo 4

Una fuente emite 5 símbolos con una probabilidad de distribución $p = [0{,}4, 0{,}2, 0{,}2, 0{,}1, 0{,}1]$. *¿Cual es la entropía H(p)? Construya un código binario para esta fuente con longitud media de palabra menor que* $H(p) + 1$.

Solución:
Utilizamos el hecho que H(p) es la entropía con respecto a la base b = 2. Un simple cálculo muestra que $H(p) \approx 2{,}12$.

$$H_2(p) = 0{,}4 \times log_2(\tfrac{1}{0{,}4}) + 0{,}2 \times log_2(\tfrac{1}{0{,}2}) + 0{,}2 \times log_2(\tfrac{1}{0{,}2}) + 0{,}1 \times log_2(\tfrac{1}{0{,}1}) + 0{,}1 \times log_2(\tfrac{1}{0{,}1}) = 2{,}12$$

La regla SF nos indica que debemos escoger y_1 el menor entero que cumpla

$$2^{y_1} \geq \tfrac{1}{0{,}4} = 2{,}5$$

cuyo resultado es, $y_1 = 2$.

De forma similar podemos calcular los demás casos:

$$y_2 = 3(2^{y_2} \geq \frac{1}{0,2} = 5), y_3 = 3(2^{y_3} \geq \frac{1}{0,2} = 5),$$
$$y_4 = 4(2^{y_4} \geq \frac{1}{0,1} = 10), y_5 = 4(2^{y_5} \geq \frac{1}{0,1} = 10.$$

Por tanto, (n_i = Número $(y_j = i)$) los parámetros requeridos son $n_2 = 1, n_3 = 2, n_4 = 2$.
Utilizando el método del árbol, es fácil construir un código PF con esos parámetros, por ejemplo:

00, 010, 011, 1000, 1001.

El tamaño medio sería:

$$\text{L} = 0.4 \text{ x } 2 + 0.2 \text{ x } 3 + 0.2 \text{ x } 3 + 0.1 \text{ x } 4 + 0.1 \text{ x } 4 = 2.8,$$

que verifica (como garantiza el teorema 4) que es menor que $H(p) + 1 \approx 3,12$.

Ejercicio 7

Utiliza la regla de codificación Shannon-Fano para construir un código binario PF para una fuente con una probabilidad de distribución $p = [0.25, 0.10, 0.15, 0.05, 0.20, 0.25]$. Encuentra la longitud media de palabra L, y verificar que L está ente $H(p)$ y $H(p) + 1$.

Ejercicio 7

Utiliza la regla de codificación Shannon-Fano para construir un código binario PF para una fuente con una probabilidad de distribución p = [0.25, 0.10, 0.15, 0.05, 0.20, 0.25]. Encuentra la longitud media de palabra L, y verificar que L está ente H(p) y H(p) + 1.

Solución:

$H_2(p) = 0{,}25 \times log_2(\frac{1}{0{,}25}) + 0{,}1 \times log_2(\frac{1}{0{,}1}) +$
$0{,}15 \times log_2(\frac{1}{0{,}15}) + 0{,}05 \times log_2(\frac{1}{0{,}05}) + 0{,}2 \times log_2(\frac{1}{0{,}2}) + 0{,}25 \times log_2(\frac{1}{0{,}25}) = 2{,}42$

$$2^{y_1} \geq \frac{1}{0{,}25} = 4; y_1 = 2 \; 2^{y_2} \geq \frac{1}{0{,}1} = 10; y_2 = 4$$

$$2^{y_3} \geq \frac{1}{0{,}15} = 6{,}66; y_3 = 3 \; 2^{y_4} \geq \frac{1}{0{,}05} = 20; y_4 = 5$$

$$2^{y_5} \geq \frac{1}{0{,}2} = 5; y_5 = 3 \; 2^{y_6} \geq \frac{1}{0{,}25} = 4; y_6 = 2$$

$n_2 = 2, n_3 = 2, n_4 = 1, n_5 = 1.$

$$\text{L} = 0.25 \text{ x } 2 + 0.25 \text{ x } 2 + 0.20 \text{ x } 3 + 0.15 \text{ x } 3 + 0.10 \text{ x } 4 + 0.05 \text{ x } 5 = 2.7.$$

Verifica que $2{,}42 \leq 2{,}7 \leq 3{,}42$.

Ejercicio 8

Utiliza la regla de codificación Shannon-Fano para construir un código ternario PF para una fuente con una probabilidad de distribución $p = [0.5,0.3,0.2]$. Muestra, construyendo un código mejor que este código no es óptimo.

Ejercicio 8

Utiliza la regla de codificación Shannon-Fano para construir un código ternario PF para una fuente con una probabilidad de distribución p = [0.5,0.3,0.2]. Muestra, construyendo un código mejor que este código no es óptimo.

Solución:

$H_3(p) = 0{,}5 \times log_3(\frac{1}{0{,}5}) + 0{,}3 \times log_3(\frac{1}{0{,}3}) + 0{,}2 \times log_3(\frac{1}{0{,}2}) = 0{,}93$

$$3^{y_1} \geq \frac{1}{0{,}5} = 2; y_1 = 1 \; 3^{y_2} \geq \frac{1}{0{,}3} = 3{,}33; y_2 = 2$$

$$3^{y_3} \geq \frac{1}{0{,}2} = 5; y_3 = 2$$

$n_1 = 1, n_2 = 2.$

$$\text{L} = 0.5 \text{ x } 1 + 0.3 \text{ x } 2 + 0.2 \text{ x } 3 = 1.7.$$

Verifica que $0{,}93 \leq 1{,}7 \leq 1{,}93$.
Pero podríamos crear un código con 1 bit cada uno, L=1.

Regla de Huffman

- Si un código UD existe, entonces es posible construir un código PF con los mismos parámetros. Por tanto, debemos restringir la búsqueda de códigos óptimos a códigos que cumpla la propiedad PF.
- La regla de Shannon-Fano produce códigos satisfactorios, pero por regla general no nos ofrece un código óptimo.
- La regla de Huffman, descrita a continuación, nos garantiza un código óptimo. (regla similar puede ser para $b > 2$.)

Lema 1

Un código óptimo PF $c : S \to \mathbb{B}^*$ para una fuente (S,p) tiene las siguientes propiedades:

1. *si la palabra codificada c(s') es mayor que c(s) entonces $p_s \geq p'_s$*
2. *entre las palabras codificadas de máxima longitud hay dos de la siguiente forma x0 y x1, para algún $x \in \mathbb{B}^*$*

La regla de Huffman emplea dos construcciones basadas en estas propiedades:

H1 Dada una fuente (S,p), sea s'y s" dos símbolos con las menores probabilidades. Construir una nueva fuente (S^*, p^*) remplazando s'y s" por un símbolo s^*, con probabilidad $p^*_{s*} = p_{s'} + p_{s''}$. Todos los demás símbolos se mantienen con las mismas probabilidades.

H2 Si tenemos un código binario h^* para (S^*, p^*) con $h^*(s^*) = w$, entonces definimos un código binario PF h para (S,p) mediante las reglas $h(s') = w0$, $h(s'') = w1$, y $h(u) = h^*(u)$ para todos $u \neq s', s''$.

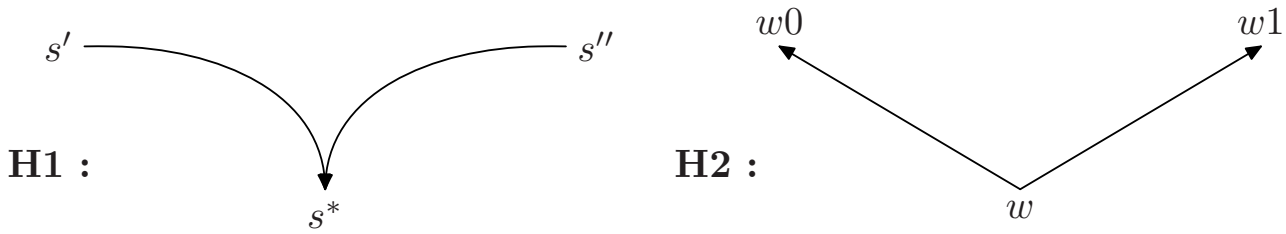

Figura 2 : Las dos reglas de Huffman

- Supongamos que tenemos una fuente con m símbolos.
- La regla H1 puede ser utilizada para construir una secuencia de fuentes cada una con un símbolo menos que la anterior, por lo que el proceso finalizaría en la fuente m*th*, la cual solamente posee un símbolo.
- El código óptimo para la última fuente es la asignación de la palabra vacía a ese símbolo único.
- Una vez aplicado lo anterior, la H2 puede ser utilizada para construir los códigos para cada una de las fuentes en la secuencia.

Ejemplo 5

Utilizar la regla de Huffman para construir un código optimo para una fuente con distribución $p = [0.4, 0.2, 0.2, 0.1, 0.1]$.

Ejemplo 5

Utilizar la regla de Huffman para construir un código optimo para una fuente con distribución p = [0.4, 0.2, 0.2, 0.1, 0.1].

Solución:
Comenzando con $p^{(1)} = p$ y definiendo $p^{(i+1)} = p^{(i)*}$ para i = 1, 2, 3, 4, la regla H1 produce la secuencia de fuentes mostrada en la siguiente figura.

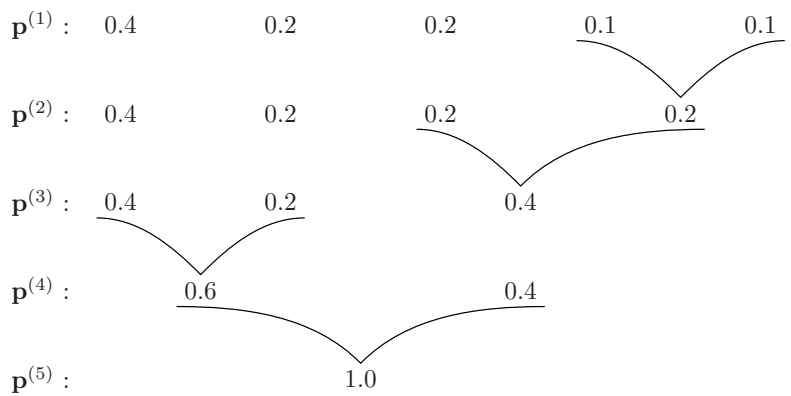

Figura 3 : Aplicación regla H1 Huffman

Para construir el código Huffman utilizamos la H2, comenzando del código que asigna la palabra vacía al símbolo único en la última fila. El proceso es mostrado en la siguiente figura.

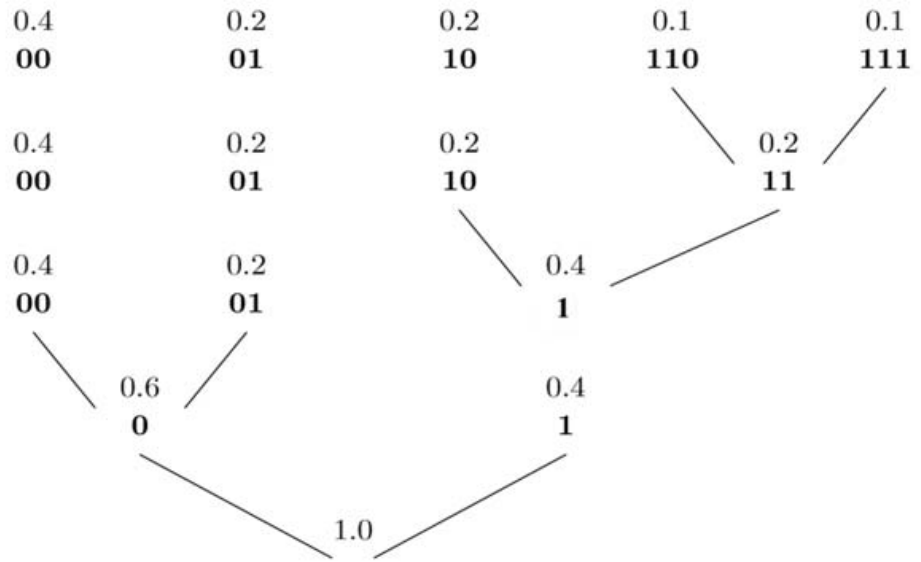

Figura 4 : Aplicación regla H2 Huffman

- Las palabras codificadas son 00, 01, 10, 110, 111.
- Para esta fuente la longitud media de palabra del código Huffman es:

$$L_{opt} = (0{,}4 + 0{,}2 + 0{,}2) \times 2 + (0{,}1 + 0{,}1) \times 3 = 2{,}2$$

- La entropía H = H(p) es aproximadamente 2.12.
- En el ejemplo anterior aplicando la regla SF a esta fuente obteniendo un código con una media de longitud de palabra de $L_{SF} = 2{,}8$.
- La relación entre los resultados obtenidos con la regla de huffman y los obtenidos anteriormente es:

$$H < L_{opt} < L_{SF} < H + 1.$$

Lema 2

Sea h y h^* definidas como en la regla H2. Entonces la longitud media de palabras de h y h^* satisfacen

$$L(h) = L(h^*) + p^*_{s^*}$$

Teorema 5

Si h^* es óptimo para (S^*, p^*) entonces h es óptimo para (S, **p***).*

Ejercicio 9

Construya un código óptimo para la fuente del Ejercicio 7 utilizando la regla de Huffman, y calcule su longitud media de palabra.

Ejercicio 10

Considere una fuente con probabilidad de distribución [0.4,0.3,0.2,0.1]. Compare la longitud media de palabra del código obtenido de la regla SF con la longitud media de palabra óptima obtenida mediante la regla de Huffman.

José A. Montenegro Montes
Dpto. Lenguajes y Ciencias de la Computación
ETSI Informática. Universidad de Málaga
monte@lcc.uma.es
twitter

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA