Teoría de la Información: Modelado y Análisis

Arno Formella

13 de febrero de 2004

1. Curso

La página incial del curso es:

`http://www.ei.uvigo.es/~formella/doc/tc03`

Estos apuntes se acompaña con ilustraciones en pizarra dónde se explica las notaciones y algunos de los conceptos.

El texto es meramente una brevísima introducción (5 horas) a la teoría de la información.

2. Bibliografía

Libros:

- David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, (previsto para 2004), existe versión electrónica: `http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.html`

3. Tareas para una Presentación

digital fountain codes:

cuckoo hashing:

advanced information visualization:

4. Motivación

¿Qué es información?

una opinión:

- dado dos entidades: fuente y destino
- información es una
  - secuencia de símbolos
  - que se puede transmitir de la fuente al destino y
  - que permite a ambas partes reproducir una misma secuencia.

es decir, la información depende del estado de ambas partes.

intercambio de información con mutuo acuerdo, es decir, fuente y destino han acordado un protocolo para interpretar los símbolos intercambiados entre ellos

el canal de comunicación puede ser no–perfecto (p.ej., modem—línea telefónica—modem, satélite—propagación—tierra, memoria—disco—memoria)

la transmisión no se puede ver solamente en el sentido de espacio, sino también en el tiempo

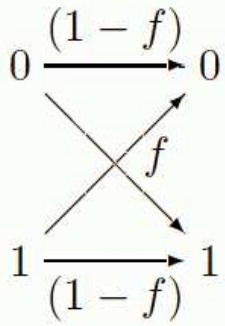
si no hay protocolo establecido entre fuente y destino, el destino puede intentar averiguar el proceso de la fuente como manda la información

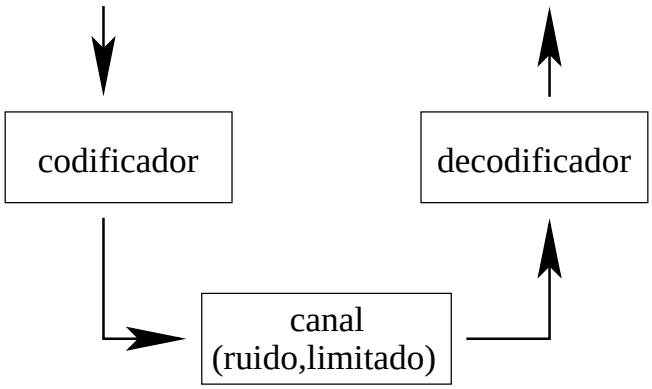
la teoría de información se dedica a los problemas de:

- la transmisión de símbolos sobre un canal no perfecto
- la compresión de una secuencia de símbolos sin pérdida de informacón (o sin pérdida relevante)
- la inferencia de la información emitida por una fuente desconocida

canal binario simétrico:

$$\begin{array}{lcllcl} P(y=0 \mid x=0) & = & 1-f; & P(y=0 \mid x=1) & = & f; \\ P(y=1 \mid x=0) & = & f; & P(y=1 \mid x=1) & = & 1-f. \end{array}$$





Ejemplo: disco duro

asumimos canal binario simétrico para leer y escribir los bits al disco con una probabilidad de fallo de 0.1 %

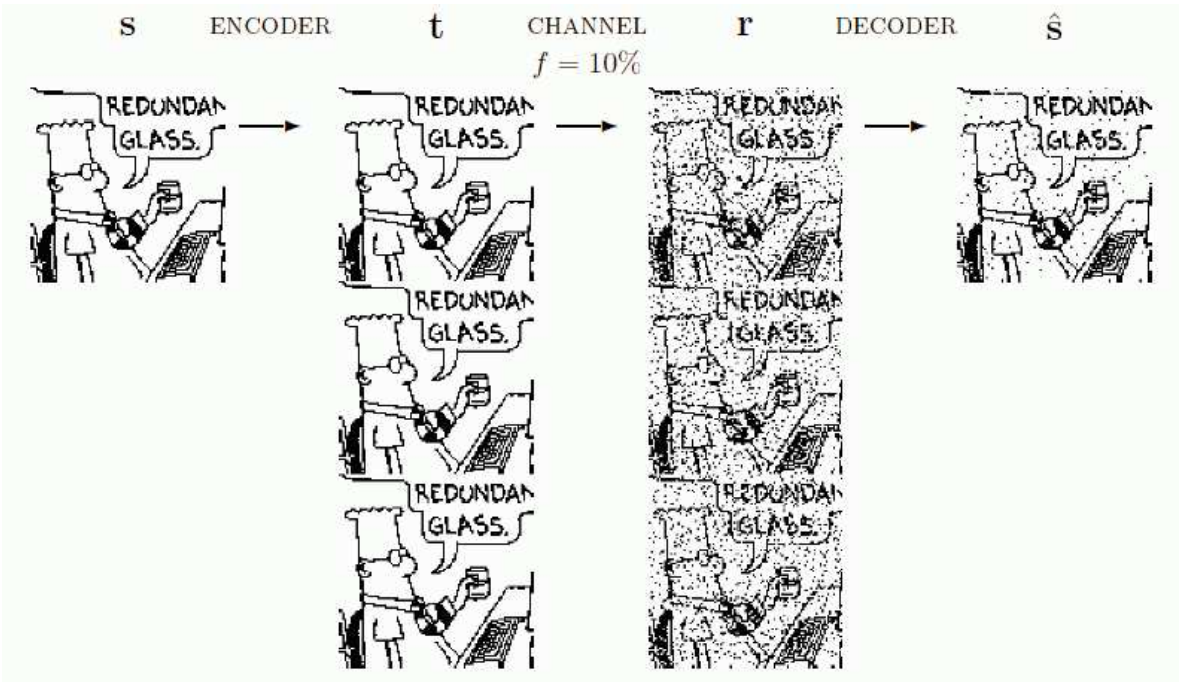
para que el disco sea útil, necesitamos una probabilidad de 10^{-15} intercambiando 1 GByte cada día durante 10 años.

códigos de repetición

usamos para cada bit tres bits y decidimos por voto mayoritario

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{\mathbf{s}}$	0	0	1	0	0	1	0
fallos corregidos		$\star$					
fallos no corregidos					$\star$		

- la probabilidad de fallo se ha reducido de 0.1 a aprox. 0.03 ($\simeq \sqrt{f}$)
- se ha aumentado el número de discos (o el espacio usado) por 3
- para llegar a 10^{-15} se necesitaría alrededor de 60 discos



(7,4)–Hamming códigos

se manda en vez de 4 bits 7 bits

codificación:

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

se observa: cada pareja se distingue por lo menos en 3 bits

¿Cómo se ha construido la table?

se puede describir la construcción del código también con una multiplicación con una matriz:

$$\mathbf{t} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \cdot \mathbf{s}$$

descodificación:

se podría buscar la palabra que menos se destingue de la palabra transmitida (según su distancia Hamming)

o mejor

se calcula el síndrome que es el vector de bits que indica donde se ha violado la paridad

síndrome **z**	000	001	010	011	100	101	110	111
cambia	*nada*	r_7	r_6	r_4	r_5	r_1	r_2	r_3

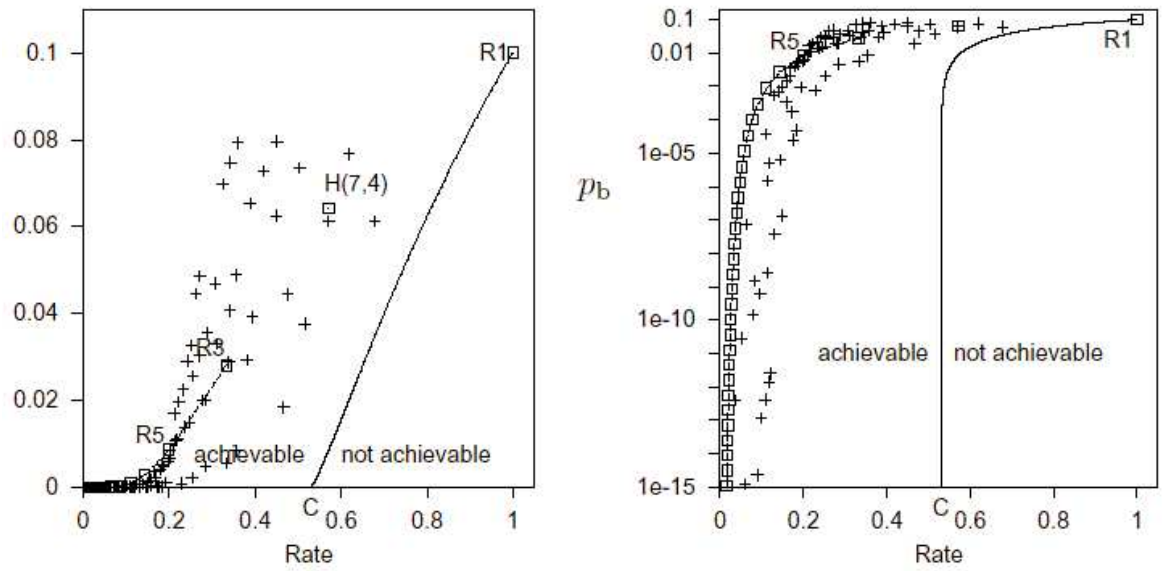
el proceso de codificación solamente funciona adecuadamente si se asume que se ha cambiado un solo bit durante la transmisión

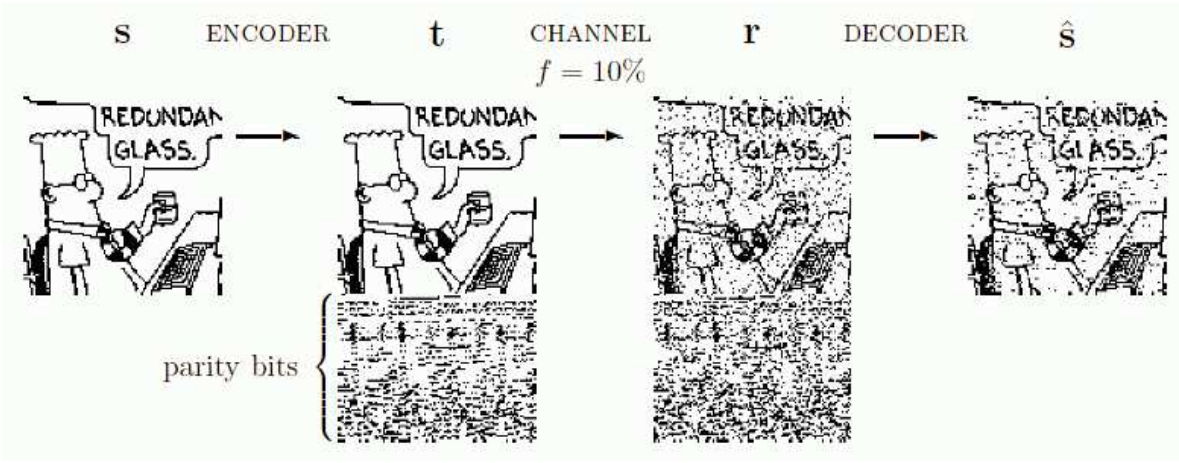
también se puede expresar el cálculo del síndrome con una multiplicación con una matriz:

$$\mathbf{z} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{r}$$

- la probabilidad de fallo se ha reducido de 0.1 a aprox. 0.07
- una generalización de Hamming códigos se llama BCH–códigos
- se ha aumentado el número de discos (o el espacio usado) por 7/4=1.75
- la relación *bits–de–datos/bits–transmitidos* se llama tasa del código

diagrama que relaciona tasa de transmisión con probabilidad de fallo

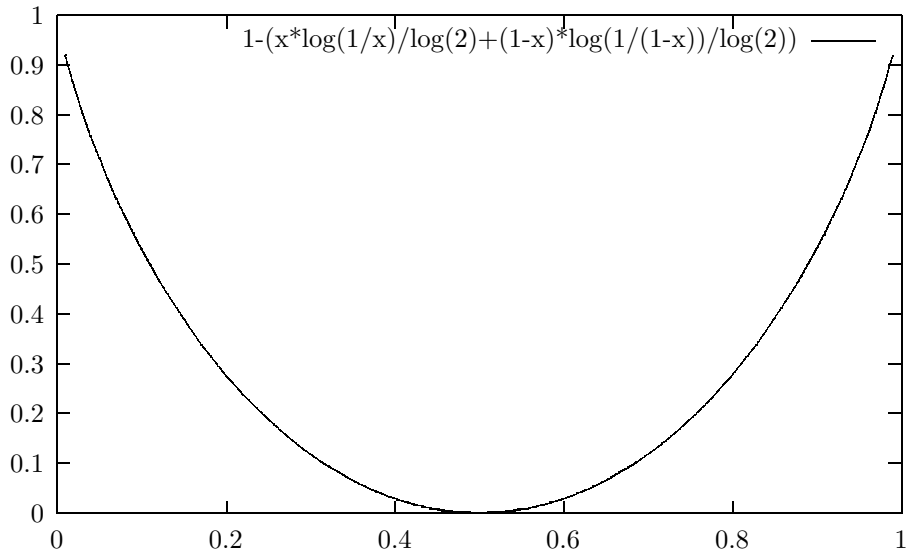




Teorema (Shannon):

La tasa máxima (=capacidad) de un canal binario simétrico que permite la transmisión con error arbitrariamente pequeño se calcula como

$$C(f) = 1 - H_2(f) = 1 - \left[f \log_2 \frac{1}{f} + (1-f) \log_2 \frac{1}{1-f} \right]$$



entonces, con el código apropriado basta con 2 discos duros (que sean lo suficientemente grande para almacenar un bloque) para garantizar una probabilidad de error tan pequeña como se desea, si la probabilidad de fallo en un bit es 0.1

5. Nociones básicas

Notaciones:

$\mathcal{A} = \{a_0, a_1, \ldots, a_{I-1}\}$	alfabeto con I símbolos
$\mathcal{P} = \{p_0, p_1, \ldots, p_{I-1}\}$	distribución de probabilidades discretas con $\sum_i p_i = 1$ y $p_i \geq 0$
x	salida (o valor) de una variable aleatoria
$X = (x, \mathcal{A}_X, \mathcal{P}_X)$	*ensemble*, si $x \in \mathcal{A}_X$, $P(x = a_i) = p_i$

entonces:

$$\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$$

abbreviación: $P(x = a_i) = P(x) = P(a_i)$ dependiendo del contexto

Ejemplo:

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Notaciones:

$P(T)$	probabilidad de un subconjunto $T \subset \mathcal{A}_X$ $P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i)$
XY	*ensemble* unido

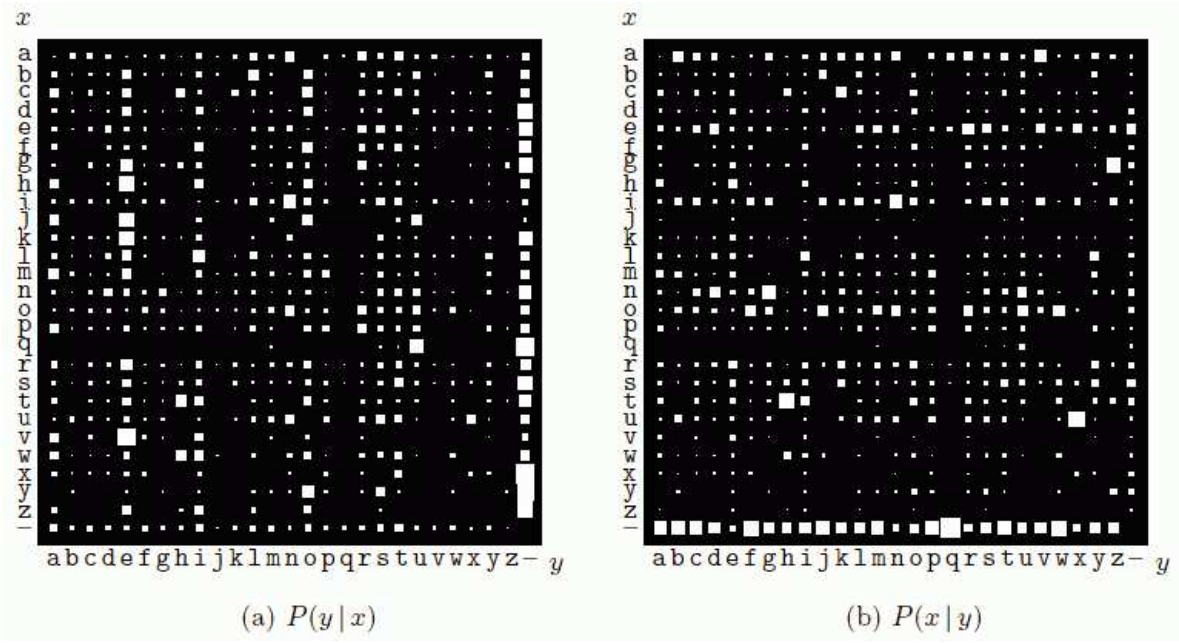
(x, y)	salida (o valor) de una variable aleatoria de un *ensemble* unido $x \in \mathcal{A}_X$ y $y \in \mathcal{A}_Y$ las dos variables no son necesariamente independiente
$P(x, y)$	probabilidad unida $P(x, y) = P(x = a_i, y = b_i)$

dado un *ensemble* unido:

Vocabulario:

probabilidad marginal	$P(x = a_i) = \sum_{y \in \mathcal{A}_Y} P(x = a_i, y)$
probabilidad condicional	$P(x = a_i \mid y = b_i) = \frac{P(x = a_i, y = b_i)}{P(y = b_i)}$ si $P(y = b_i) \neq 0$ probabilidad que x sea a_i dado que y es b_i

Ejemplo:



(a) $P(y \mid x)$ (b) $P(x \mid y)$

6. Reglas de cálculo

$\mathcal{H}$ denota bajo que condición se asume las probabilidades

Regla de producto (o regla de cadena):

$$P(x, y \mid \mathcal{H}) = P(x \mid y, \mathcal{H})P(y \mid \mathcal{H}) = P(y \mid x, \mathcal{H})P(x \mid \mathcal{H})$$

Regla de suma:

$$\begin{aligned} P(x \mid \mathcal{H}) &= \sum_y P(x, y \mid \mathcal{H}) \\ &= \sum_y P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H}) \end{aligned}$$

Regla de BAYES:

$$\begin{aligned} P(y \mid x, \mathcal{H}) &= \frac{P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H})}{P(x \mid \mathcal{H})} \\ &= \frac{P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H})}{\sum_{y'} P(x \mid y', \mathcal{H}) P(y' \mid \mathcal{H})} \end{aligned}$$

Vocabulario:

independencia	dos variables aleatorias x e y son independientes si: $P(x, y) = P(x)P(y)$

Ejemplo:

- Fulanito se hace la prueba de una enfermedad.
- La prueba es 95 % fiable, es decir, en 95 % de los casos que alguien tenga la enfermedad, dice *si*, y en 95 % de los casos que alguien no tenga la enfermedad, dice *no*.
- Se sabe que 1 % de la población está enferma.
- ¿Cuál es la probabilidad que Fulanito esté enfermo?

Solución:

$a = 1$	Fulanito está enfermo
$a = 0$	Fulanito no está enfermo

$b = 1$	La prueba es positiva
$b = 0$	La prueba es negativa

probabilidades condicionales:

$$P(b=1 \mid a=1) = 0{,}95 \qquad P(b=1 \mid a=0) = 0{,}05$$
$$P(b=0 \mid a=1) = 0{,}05 \qquad P(b=0 \mid a=0) = 0{,}95$$

probabilidades marginales (de a):

$$P(a=1) = 0{,}01 \qquad P(a=0) = 0{,}99$$

probabilidad marginal (de $b=1$);

$$P(b=1) = P(b=1 \mid a=1)P(a=1) + P(b=1 \mid a=0)P(a=0)$$

probabilidad que Fulanito esté enfermo:

$$\begin{aligned} P(a=1 \mid b=1) &= \frac{P(b=1 \mid a=1)P(a=1)}{P(b=1 \mid a=1)P(a=1) + P(b=1 \mid a=0)P(a=0)} \\ &= \frac{0{,}95 \times 0{,}01}{0{,}95 \times 0{,}01 + 0{,}05 \times 0{,}99} \\ &= 0{,}16 \end{aligned}$$

es decir: solamente un 16 %

si $P(a=1) = 0{,}001$ la probabilidad baja a un 1.87 %

Recomiendo leer la sección 2.3 del libro.

Notaciones:

$h(x)$ contenido de información de una salida x (dado un *ensemble*)
$h(x) = \log_2 \frac{1}{P(x)}$

$H(X)$ entropía de un *ensemble*
es el contenido de información medio de una salida de un *ensemble*
$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}$

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
	$\sum_i p_i \log_2 \frac{1}{p_i}$		4.1

Reglas para la entropía:

positivo:

$$H(X) \geq 0$$

$$H(X) = 0 \qquad \iff \qquad p_i = 1 \quad \text{para un } i$$

confinado:

$$H(X) \leq \log(|\mathcal{A}_X|)$$

$$H(X) = \log(|\mathcal{A}_X|) \qquad \iff \qquad \forall i : p_i = 1/|X|$$

ensembles unidos X e Y:

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)}$$

$$P(x,y) = P(x)P(y) \qquad \Longrightarrow \qquad H(X,Y) = H(X) + H(Y)$$

recursividad simple:

$$H(\mathbf{p}) = H(p_1, 1-p_1) + (1-p_1) H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_I}{1-p_1}\right)$$

recursividad general:

$$\begin{aligned} H(\mathbf{p}) \;=\; & H\left[(p_1 + p_2 + \cdots + p_m), (p_{m+1} + p_{m+2} + \cdots + p_I)\right] \\ & + (p_1 + \cdots + p_m) H\left(\frac{p_1}{(p_1 + \cdots + p_m)}, \ldots, \frac{p_m}{(p_1 + \cdots + p_m)}\right) \\ & + (p_{m+1} + \cdots + p_I) H\left(\frac{p_{m+1}}{(p_{m+1} + \cdots + p_I)}, \ldots, \frac{p_I}{(p_{m+1} + \cdots + p_I)}\right). \end{aligned}$$

entropía relativa (o KULLBACK-LEIBLER divergencia)

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \qquad (\geq 0)$$

siendo $P(x)$ y $Q(x)$ dos distribuciones de probabilidades sobre el mismo alfabeto

7. Compresión

compresión de datos, modelado de datos, y eliminación de ruído tienen mucho en común

¿Cómo se puede medir información?

Ejemplo con las 12 bolas...

según SHANNON medimos el contenido de información de un símbolo emitido por una fuente, con el logaritmo de base 2 de su probabilidad de aparencia.

¿Porqué el logaritmo?

porque para variables aleatorias independientes, es decir, con $P(x,y) = P(x)P(y)$ tenemos:

$$
\begin{aligned}
h(x,y) &= \log \frac{1}{P(x,y)} \\
&= \log \frac{1}{P(x)P(y)} \\
&= \log \frac{1}{P(x)} + \log \frac{1}{P(y)} \\
&= h(x) + h(y)
\end{aligned}
$$

es decir, justamente *sumamos* sus contenidos de información individuales para obtener la información si ambos eventos ocurren a la vez.

el valor obtenido se puede interpretar como el número de *bits* necesarios para codificar la información.

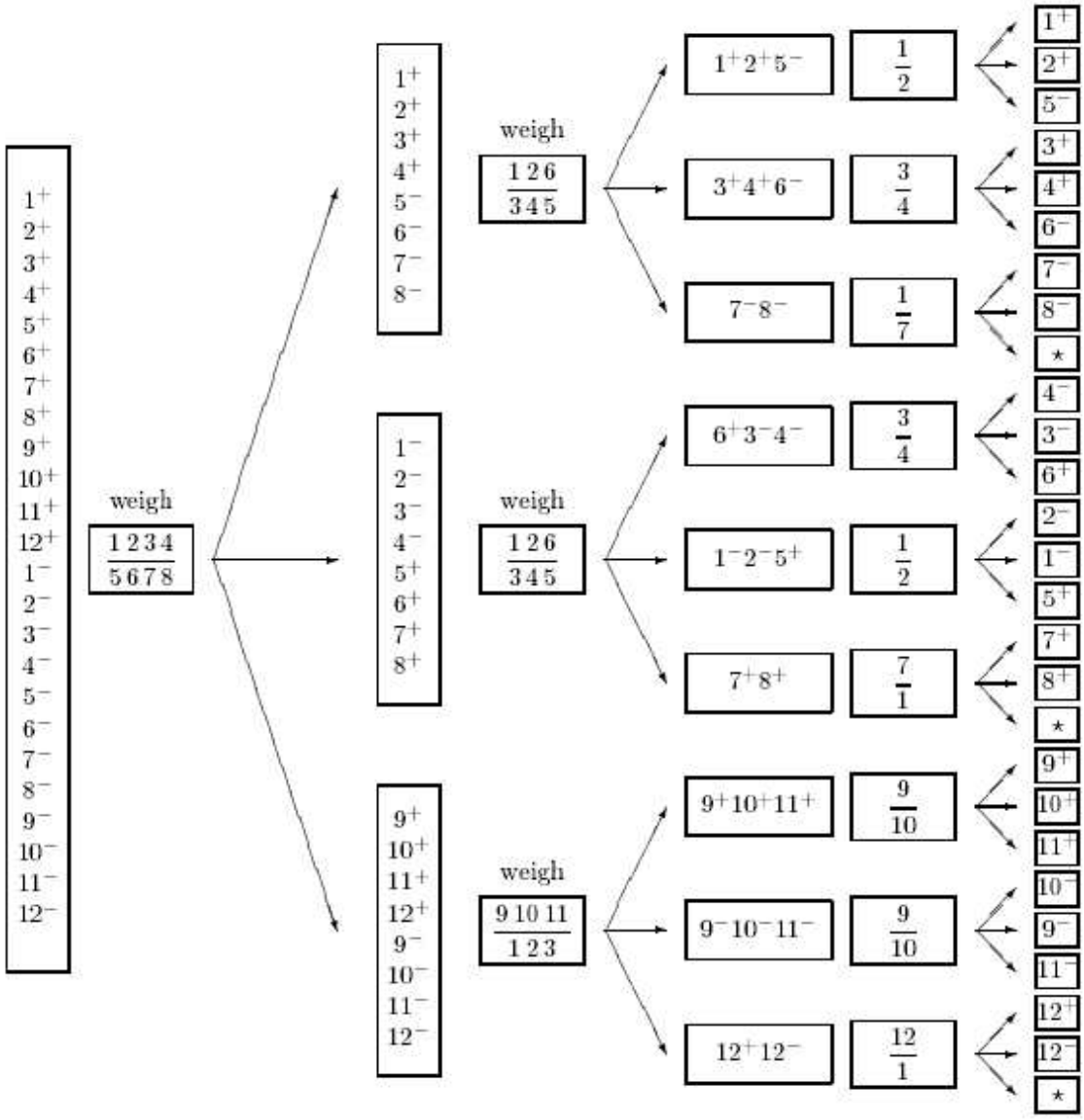
¿Cuántos bits de información están *escondidos* en el ejemplo con las bolas?

¿Cuántos bits de información se gana con un uso de la báscula?

¿Qué propriedad tiene una solución óptima?

en cada uso de la báscula hay que intentar equilibrar las probabilidades de todos los posibles salidas del experimento.

por ejemplo, 6 contra 6 bolas núnca puede dar el caso de equilibrio.



¿Hay otra estrategia?

la comprobación del teorema de SHANNON trata también del caso que las probabilidades no son uniformes.

siguimos directamente con el libro ...