Formato float32 (binary32)

Objetivos. Estudiar el formato **float32** (llamado también **binary32**) que ocupa 32 bits y se usa para representar números reales con cierta precisión finita.

Requisitos. Notación científica, sistema binario, sistema hexadecimal.

1. El formato float32 (conocido también binary32) es uno de los formatos descritos en los estándares técnicos IEEE 754-1985, IEEE 754-2008, IEEE 854-1987 y ISO/IEC/IEEE 60559:2011.

Ligas útiles:
`http://www.fdi.ucm.es/profesor/mozos/AEC/aritm_pf.PDF`
`http://babbage.cs.qc.edu/IEEE-754/`
`http://en.wikipedia.org/wiki/Single_precision`

2. Descripción breve. Cada cadena de 32 bits se divide en tres partes: signo (ocupa 1 bit), exponente (ocupa 8 bits) y la parte fraccionaria de la mantisa (ocupa 23 bits):

$$\underbrace{1 \text{ bit}}_{s}\ \underbrace{8 \text{ bits}}_{e}\ \underbrace{23 \text{ bits}}_{f}.$$

El número real correspondiente es

$$a = (-1)^s \cdot 1.f_2 \cdot 2^{e_2 - 127},$$

Notemos que la parte entera de la mantisa siempre es 1 y no se guarda (para ahorrar un bit). Por brevedad vamos a escribir el formato float32 con dígitos hexadecimales.

3. Ejemplo. Transformar en el formato float32 el número

$$a = -7.975 \cdot 10^1.$$

Solución. Primero escribimos a en la notación científica binaria:

$$a = -79.75 = -1001111.11_2 = (-1)^1 \cdot 1.00111111_2 \cdot 2^6.$$

Denotamos por f la cadena 00111111 y notamos que $s = 1$. La definición del formato float32 dice que $e - 127 = 6$; de aquí calculamos el exponente e:

$$e = 127 + 6 = 133 = 10000101_2.$$

Juntamos estas tres partes y obtenemos la representación de a en el formato float32:

$$a = \underbrace{1}_{s}\ \underbrace{10000101}_{e}\ \underbrace{00111111000000000000000}_{f}.$$

Agrupamos los bits en tetradas y escribimos a en el formato float32 con dígitos hexadecimales:

$$a = \underbrace{1100\ 0010\ 1001\ 1111\ 1000\ 0000\ 0000\ 0000}_{\text{float32}} = \underbrace{C29F8000_{16}}_{\text{float32}}. \qquad \square$$

4. Ejemplo. Transformar a la notación científica decimal el siguiente número dado en el formato float32 con cifras hexadecimales:

$$a = \underbrace{3EC00000_{16}}_{\text{float32}}.$$

Solución. Primero escribimos a con dígitos binarios y los separamos en tres partes de longitudes 1, 8 y 23:

$$a = \underbrace{0011\ 1110\ 1100\ 0000\ 0000\ 0000\ 0000\ 0000}_{\text{float32}} = \underbrace{0}_{s}\ \underbrace{01111101}_{e}\ \underbrace{100\ldots}_{f}.$$

De aquí $e = 01111101_2 = 125$,

$$\begin{aligned} a &= (-1)^0 \cdot 2^{125-127} \cdot 1.100\ldots_2 = 2^{-2} \cdot 1.100\ldots_2 = 0.01100\ldots_2 \\ &= \frac{1}{4} + \frac{1}{8} = \frac{3}{8} = 0.375 = 3.75 \cdot 10^{-1}. \end{aligned} \qquad \square$$