Matematica Applicata
- Numeri in virgola mobile -

Laboratorio didattico

A.A 2011/2012

Stefano Vena

I numeri Reali

- Essendo impossibile rappresentare su una macchina (le cui risorse sono necessariamente finite) l'infinità dei numeri reali (**R**) viene impiegato un sottoinsieme di dimensione finita che indicheremo con (**F**) e chiameremo insieme dei numeri floating-point.
- Ogni singolo numero reale x viene rappresentato dalla macchina con un numero arrotondato, che si indica con fl(x) e viene detto numero macchina, che non coincide necessariamente con il numero x di partenza

Differenze fra R e F

- Consideriamo il numero razionale **x = 1/7**, la cui rappresentazione decimale è $\mathbf{0,\overline{142857}}$.
- Tale rappresentazione è infinita, nel senso che esistono infinite cifre non nulle dopo la virgola.
- Impiegando un calcolatore, tale numero viene rappresentato come **0,1429**
- Cioè un numero costituito apparentemente da sole 4 cifre decimali, l'ultima delle quali inesatta rispetto alla quarta cifra del numero reale

Differenze tra R e F

- Il numero razionale 1/3 è valutato come 0,3333, nel quale anche la quarta cifra è esatta.
- Questo comportamento è dovuto al fatto che i numeri reali sul calcolatore vengono **arrotondati**.
- Viene memorizzato solo un numero fissato a priori di cifre decimali.
- L'ultima cifra decimale memorizzata risulta incrementata di 1 rispetto alla corrispondente cifra decimale del numero originario qualora la cifra successiva in quest'ultimo risulti maggiore od uguale a 5.

Memorizzazione dei numeri Reali

- $x=(-1)^s \cdot(0,a_1a_2...a_t)\cdot\beta^e = (-1)^s \cdot m\cdot\beta^{e-t}, a_1 \neq 0$
- dove **s** vale 0 o 1.
- **β** (un numero intero positivo maggiore od uguale a 2) é la base.
- **m** è un intero detto mantissa di lunghezza è t
- **t** è il numero massimo di cifre $\mathbf{a_i}$ (con $0 \le a_i \le \beta - 1$) memorizzabili.
- **e** è un numero intero detto esponente.

I Floating-point

- $x=(-1)^s \cdot(0,a_1a_2...a_t)\cdot\beta^e = (-1)^s \cdot m\cdot\beta^{e-t}$, $a_1 \neq 0$
- I numeri di macchina nel formato sono detti numeri floating-point essendo variabile la posizione del punto deci- male.
- Le cifre $a_1a_2 \ldots a_p$ (con $p \leq t$) vengono generalmente chiamate le prime **p** cifre significative di **x**.
- La condizione $a_1 \neq 0$ impedisce che lo stesso numero possa avere più rappresentazioni.
- Ad esempio, senza questa condizione, 1/10 in base 10 potrebbe essere rappresentato come 0.1 · 100 o 0.01 · 101 e così via.
- L'insieme F è dunque completamente caratterizzato dalla base β, dal numero di cifre significative t e dall'intervallo (L,U) (con L < 0 ed U > 0) di variabilità dell'esponente e. Viene perciò anche indicato con F(β,t,L,U)

Floating point – Errore di arrotondamento

- Il numero 0 non appartiene a F, poiché per esso $a_1 = 0$ quindi viene trattato a parte.
- L'errore di arrotondamento che si commette sostituendo ad un numero reale $x \neq 0$ il suo rappresentante fl(x) in F, è generalmente piccolo.

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\epsilon M$$

Floating point- Errore di arrotondamento

- Dove $\varepsilon M = \beta^{1-t}$ rappresenta la distanza fra 1 ed il più vicino numero floating-point maggiore di 1.
- Si osservi che εM dipende da β e da t.
- Ad esempio, in MATLAB, F(2,53,-1021,1024) si ha $\varepsilon M = 2^{-52} \simeq 2.22 \cdot 10^{-16}$
- Il numero $u = \frac{1}{2}\varepsilon M$ rappresenta dunque il massimo errore relativo che la macchina può commettere. Ed è detta **unità di arrotondamento**.

Floating point - Limiti

- L ed U sono valori finiti e delimitano il più piccolo ed il più grande numero positivo.
- $xmin = \beta^{L-1}$, $xmax = \beta^{U} (1 - \beta^{-t})$
- Valori positivi minore di xmin producono errori di underflow e considerati come 0
- Valori positivi maggiori di xmax danno origine a origine a errori di overflow e vengono considerati con il valore speciale **Inf**
- Il fatto che xmin e xmax siano gli estremi di un intervallo molto vasto della retta reale non deve trarre in inganno: i numeri di F sono molto addensati vicino a xmin, diventando sempre più radi all'avvicinarsi di xmax.

Esercizi

1. Da quanti numeri è costituito l'insieme $F(2, 2, -2, 2)$?
2. Quanto vale εM per tale insieme?
3. Si verifichi che in generale l'insieme $F(\beta, t, L, U)$ contiene $2(\beta - 1)\beta^{t-1}(U - L + 1)$ numeri.

Soluzioni

1. Stanno in F(2,2,-2,2) tutti i numeri della forma $\pm\, 0.1\, a_2\, 2^e$ con a_2=0,1 ed e intero compreso fra -2 e 2. Fissato l'esponente, si possono rappresentare i soli numeri 0.10 e 0.11, a meno del segno; di conseguenza, in F(2,2,-2,2) sono contenuti 20 numeri.
2. $\varepsilon M = ½$
3. Fissato l'esponente, abbiamo a disposizione β posizioni per le cifre $a_2,\ldots,a_t$ e β-1 per la cifra a_1(che non può assumere il valore 0). In tutto avremo perciò $(\beta-1)\ \beta^{t-1}$ numeri rappresentabili a meno del segno e per esponente fissato. L'esponente può assumere U-L+1 valori e quindi, complessivamente, l'insieme F(β,t,L,U) è costituito da 2 $(\beta-1)\ \beta^{t-1}(U-L+1)$ elementi