Métodos numéricos
Sesión 4
(Aritmética computacional)

Universidad Externado de Colombia
Programa Ciencia de Datos

202410

Tabla de contenidos

Digitos significativos

Definición (Digito significativo)

Los **digitos significativos** son digitos que empiezan con un digito distinto de cero del extremo izquierdo y terminan con el digito correcto del extremo derecho, incluye los ceros finales que son exactos.

Ejemplo:

1. Digitos distintos de cero
   - 81: Dos digitos significativos
   - 127.85: Cinco digitos significativos
2. *Ceros* entre dos digitos significativos distintos de cero
   - 127.17001: Ocho digitos significativos
   - 101.820001: Nueve digitos significativos
3. *Ceros* a la derecha del último digito en decimal o entero
   - 2.700: Cuatro digitos significativos
   - 0.0270: Tres digitos significativos
   - 17500: Tres, cuatro, cinco digitos.

Digitos significativos

Definición (Digito significativo)

Los **digitos significativos** son digitos que empiezan con un digito distinto de cero del extremo izquierdo y terminan con el digito correcto del extremo derecho, incluye los ceros finales que son exactos.

Ejemplo:

1. Digitos distintos de cero
   - 81: Dos digitos significativos
   - 127.85: Cinco digitos significativos
2. *Ceros* entre dos digitos significativos distintos de cero
   - 127.17001: Ocho digitos significativos
   - 101.820001: Nueve digitos significativos
3. *Ceros* a la derecha del último digito en decimal o entero
   - 2.700: Cuatro digitos significativos
   - 0.0270: Tres digitos significativos
   - 17500: Tres, cuatro, cinco digitos.

Ejemplo 1:

$$\begin{cases} 0.2072x + 0.4248y = 1.4767 \\ 0.4166x + 0.8494y = 1.8656 \end{cases}$$

Digitos significativos

Definición (Digito significativo)

Los **digitos significativos** son digitos que empiezan con un digito distinto de cero del extremo izquierdo y terminan con el digito correcto del extremo derecho, incluye los ceros finales que son exactos.

Ejemplo:

1. Digitos distintos de cero
   - 81: Dos digitos significativos
   - 127.85: Cinco digitos significativos
2. *Ceros* entre dos digitos significativos distintos de cero
   - 127.17001: Ocho digitos significativos
   - 101.820001: Nueve digitos significativos
3. *Ceros* a la derecha del último digito en decimal o entero
   - 2.700: Cuatro digitos significativos
   - 0.0270: Tres digitos significativos
   - 17500: Tres, cuatro, cinco digitos.

Ejemplo 1:

$$\begin{cases} 0.2072x + 0.4248y = 1.4767 \\ 0.4166x + 0.8494y = 1.8656 \end{cases}$$

1. $$\begin{cases} 0.207x + 0.425y \approx 1.477 \\ 0.417x + 0.849y \approx 1.866 \end{cases}$$

   $$x = -311.014 \qquad y = 154.957$$

2. $$\begin{cases} 0.2072x + 0.4248y = 1.4767 \\ 0.4166x + 0.8494y = 1.8656 \end{cases}$$

Digitos significativos

Definición (Digito significativo)

Los **digitos significativos** son digitos que empiezan con un digito distinto de cero del extremo izquierdo y terminan con el digito correcto del extremo derecho, incluye los ceros finales que son exactos.

Ejemplo:

1. Digitos distintos de cero
   - 81: Dos digitos significativos
   - 127.85: Cinco digitos significativos
2. *Ceros* entre dos digitos significativos distintos de cero
   - 127.17001: Ocho digitos significativos
   - 101.820001: Nueve digitos significativos
3. *Ceros* a la derecha del último digito en decimal o entero
   - 2.700: Cuatro digitos significativos
   - 0.0270: Tres digitos significativos
   - 17500: Tres, cuatro, cinco digitos.

Ejemplo 1:

$$\begin{cases} 0.2072x + 0.4248y = 1.4767 \\ 0.4166x + 0.8494y = 1.8656 \end{cases}$$

1. $$\begin{cases} 0.207x + 0.425y \approx 1.477 \\ 0.417x + 0.849y \approx 1.866 \end{cases}$$

   $$x = -311.014 \qquad y = 154.957$$

2. $$\begin{cases} 0.2072x + 0.4248y = 1.4767 \\ 0.4166x + 0.8494y = 1.8656 \end{cases}$$

   $$x = -473.158 \qquad y = 234.263$$

Error absoluto y relativo

Definición

Suponga que p^* es una aproximación a p. El **error real** es $p - p^*$, el **error absoluto** es

$$|p - p^*|$$

, y el **error relativo** es

$$\frac{|p - p^*|}{|p|}$$

, siempre y cuando $p \neq 0$

Ejemplo:

- $p_1 = 1.277 \qquad p_1^* = 1.278$

Error absoluto y relativo

Definición

Suponga que p^* es una aproximación a p. El **error real** es $p - p^*$, el **error absoluto** es

$$|p - p^*|$$

, y el **error relativo** es

$$\frac{|p - p^*|}{|p|}$$

, siempre y cuando $p \neq 0$

Ejemplo:

- $p_1 = 1.277 \qquad p_1^* = 1.278$

$$\text{Error absoluto} = |1.277 - 1.278| = 0.001$$

$$\text{Error relativo} = \frac{|1.277 - 1.278|}{|1.277|} = 0.000783$$

Error absoluto y relativo

Definición

Suponga que p^* es una aproximación a p. El **error real** es $p - p^*$, el **error absoluto** es

$$|p - p^*|$$

, y el **error relativo** es

$$\frac{|p - p^*|}{|p|}$$

, siempre y cuando $p \neq 0$

Ejemplo:

- $p_1 = 1.277 \qquad p_1^* = 1.278$

$$\text{Error absoluto} = |1.277 - 1.278| = 0.001$$

$$\text{Error relativo} = \frac{|1.277 - 1.278|}{|1.277|} = 0.000783$$

- $p_2 = 0.007 \qquad p_2^* = 0.008$

Error absoluto y relativo

Definición

Suponga que p^* es una aproximación a p. El **error real** es $p - p^*$, el **error absoluto** es

$$|p - p^*|$$

, y el **error relativo** es

$$\frac{|p - p^*|}{|p|}$$

, siempre y cuando $p \neq 0$

Ejemplo:

- $p_1 = 1.277 \qquad p_1^* = 1.278$

$$\text{Error absoluto} = |1.277 - 1.278| = 0.001$$

$$\text{Error relativo} = \frac{|1.277 - 1.278|}{|1.277|} = 0.000783$$

- $p_2 = 0.007 \qquad p_2^* = 0.008$

$$\text{Error absoluto} = |0.007 - 0.008| = 0.001$$

$$\text{Error relativo} = \frac{|0.007 - 0.008|}{|0.007|} = 0.\widehat{142857}$$

Exactitud y precisión

Definición (Exactitud)

1. n cifras decimales: Cuando se puede confiar en n digitos a la derecha del lugar decimal
2. n digitos significativos: Cuando se puede confiar en un total de n digitos que sean importantes a partir del digito distinto de cero del extremo izquierdo.

Exactitud y precisión

Definición (Exactitud)

1. *n* cifras decimales: Cuando se puede confiar en *n* digitos a la derecha del lugar decimal
2. *n* digitos significativos: Cuando se puede confiar en un total de *n* digitos que sean importantes a partir del digito distinto de cero del extremo izquierdo.

1. Suponga que tiene una regla graduada en milimetros para medir un terreno cualquiera. Esto significa que las medidas serán exactas a un milimetro o por 0.001m, por esta razón, una medida como 12.271m tendrá una **exactitud** de tres cifras decimales, mientras que, una medida como 12.2712365m no tendría sentido, dado que podría tomar una medida como la anterior u otra, como la siguiente: 12.272m.
2. Para el caso de la medida 12.271m se tendría una exactitud de cinco(5) digitos significativos.
3. $$2.4 + 6.91 = 9.31$$

Definición (Redondeo y truncamiento

1. **Redondeo**: Se utiliza para

   reducir los digitos significativos

   en un número
   - Un número x está **truncado** a *n* digitos cuando todos los digitos que siguen al n-ésimo digito son descartados y ninguno de los *n* restantes se cambia. (i.e. 0.147 0.14, 0.185 0.18).
   - Un número x está **redondeado** a *n* digitos cuando x se reemplaza por un n-dígito que se aproxime a x con un error mínimo. (i.e. 0.147 0.15, 0.185 0.19).

Puntos flotantes

$$\mathbb{R}_+(\text{decimal}) = \begin{cases} \text{Parte entera} \\ \text{Parte fraccionaria} \end{cases}$$

$$\mathbb{R}_+(\text{decimal}) = \{\text{Notación cientifica (normalizada)}$$

i.e.

- $12.131 = 0.12131 \times 10^2$
- $0.0012 = 0.12 \times 10^{-2}$

Representación (Punto flotante normalizada)

Cualquier número real $x \in \mathbb{R}$, $x \neq 0$ se puede representar de la forma *punto flotante normalizada* como:

$$x = \pm 0.d_1 d_2 d_3 \cdots \times 10^n$$

donde $d_1 \neq 0$ y $n \in \mathbb{Z}$ y $d_i \in \mathbb{D}$, es decir:

$$x = \pm r \times 10^n \qquad \left(\frac{1}{10} \leq r < 1\right)$$

r: Mantisa normalizada

n: Exponente

Punto flotante Binario

Representación (Punto flotante binario)

Si $x \neq 0$ se puede representar de la forma *punto flotante normalizada* como:

$$x = \pm q \times 2^m \qquad \left(\frac{1}{2} \leq q < 1\right)$$

q: Mantisa normalizada: $q = (0, b_1 b_2 ...)_2$ donde $b_1 \neq 0$
n: Exponente

Ejercicio

1. Encuentre los números de punto flotante que se pueden expresar dado que:

$$x = \pm(0.b_1 b_2 b_3)_2 \times 2^{\pm k} \qquad (k, b_i \in \{0, 1\})$$

Punto flotante Binario

Representación (Punto flotante binario)

Si $x \neq 0$ se puede representar de la forma *punto flotante normalizada* como:

$$x = \pm q \times 2^m \qquad \left(\frac{1}{2} \leq q < 1\right)$$

q: Mantisa normalizada: $q = (0, b_1 b_2 ...)_2$ donde $b_1 \neq 0$
n: Exponente

Ejercicio

1. Encuentre los números de punto flotante que se pueden expresar dado que:

$$x = \pm(0.b_1 b_2 b_3)_2 \times 2^{\pm k} \qquad (k, b_i \in \{0, 1\})$$

2. Encuentre los números de punto flotante **normalizados** (Agujero en cero)

Nota

Los números reales que se pueden representar en una computadora se llaman **números de máquina**

Representación bits

Máquina 32 bits (digitos binarios)

Suponderemos que la máquina (computadora) almacena números en palabras de 32 bits de la forma:

$$\pm q \times 2^m$$

Representación bits

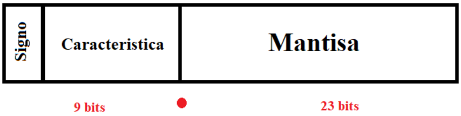
Máquina 32 bits (digitos binarios)

Suponderemos que la máquina (computadora) almacena números en palabras de 32 bits de la forma:

$$\pm q \times 2^m$$

¿Cómo asignar ese espacio?

- Signo para q: 1 bit
- Entero —m—: 8 bits (característica)
- Número 1: 23 bits (mantisa)



Representación bits

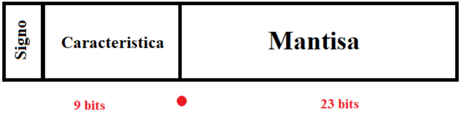
Máquina 32 bits (digitos binarios)

Suponderemos que la máquina (computadora) almacena números en palabras de 32 bits de la forma:

$$\pm q \times 2^m$$

¿Cómo asignar ese espacio?

- Signo para q: 1 bit
- Entero —m—: 8 bits (característica)
- Número 1: 23 bits (mantisa)



Nota

Con este esquema se podrían representar números reales con —m— tan grande como $2^7 - 1 = 127$. Por el exponente, se podrían tener numeros del -127 al 128.

Punto flotante de simple precisión

$$(-1)^s \times 2^{c-127} \times (1.f)_2$$

1. Signo de la mantisa:

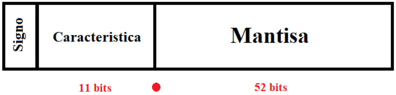
$$s = \begin{cases} 0 & \text{Positivo} \\ 1 & \text{Negativo} \end{cases}$$

2. La característica se utiliza para representar al número c en el exponente 2^{c-127} (código en exceso)
3. Los siguientes 23 bits representan f de la parte fraccionaria de la mantisa en la forma "uno más" $(1.f)_2$

Punto flotante de doble precisión

"...en este caso cada número de punto flotante de doble precisión se almacena en la memoria de dos palabras..."

$$(-1)^s \times 2^{c-1023} \times (1.f)_2$$



1. Signo de la mantisa:

$$s = \begin{cases} 0 & \text{Positivo} \\ 1 & \text{Negativo} \end{cases}$$

2. La característica se utiliza para representar al número c en el exponente 2^{c-1023} (código en exceso)
3. Los siguientes 52 bits representan f de la parte fraccionaria de la mantisa en la forma "uno más" $(1.f)_2$

Ejemplos

1. $[1 \;\; 10000100 \;\; 10100001111000000000000]_2$

$$-52.234375$$

$$[C250F000]_{16}$$

2. $[1 \;\; 10000000100 \;\; 101000011110000\cdots 00]_2$

$$-52.234375$$

Aritmética de dígitos finitos

Suponga que, $fl(x)$ y $fl(y)$ son las representaciones de punto flotante para $x, y \in \mathbb{R}$ y que se definen las operaciones de máquina, suma, resta, multiplicación y división con lo siguientes simbolos $\oplus$, $\ominus$, $\odot$ y $\oslash$, tales que:

- $x \oplus y = fl(fl(x) + fl(y))$
- $x \ominus y = fl(fl(x) - fl(y))$
- $x \odot y = fl(fl(x) \cdot fl(y))$
- $x \oslash y = fl(fl(x) \div fl(y))$

Ejemplo

1. Trunque a cinco digitos para calcular las operaciones anteriores para los números $x = \dfrac{5}{7}$ e $y = \dfrac{1}{3}$

Ejemplo

1. Trunque a cinco digitos para calcular las operaciones anteriores para los números $x = \dfrac{5}{7}$ e $y = \dfrac{1}{3}$

$fl(x) = 0.71428 \times 10^0$ $\qquad$ $fl(y) = 0.33333 \times 10^0$

Ejemplo

1. Trunque a cinco digitos para calcular las operaciones anteriores para los números $x = \dfrac{5}{7}$ e $y = \dfrac{1}{3}$

$fl(x) = 0.71428 \times 10^0$ $\qquad$ $fl(y) = 0.33333 \times 10^0$

1. $x \oplus y = fl(0.71428 \times 10^0 + 0.33333 \times 10^0) = fl(1.04761 \times 10^0)$

$$x \oplus y = 0.10476 \times 10^1$$

Ejercicio

Operación	p	p^*	Error absoluto	Error relativo
$x \oplus y$	$\frac{22}{21}$	0.10476×10^1	0.190×10^{-4}	0.182×10^{-4}
$x \ominus y$	$\frac{8}{21}$	0.38095×10^0	0.238×10^{-5}	0.625×10^{-5}
$x \odot y$	$\frac{5}{21}$	0.23809×10^0	0.524×10^{-5}	0.220×10^{-4}
$x \oslash y$	$\frac{15}{7}$	0.21428×10^1	0.571×10^{-4}	0.267×10^{-4}

Ejemplos

1. $[1\ \ 10000100\ \ 10100001111000000000000]_2$

$$-52.234375$$

- Primer bit: 1 $\rightarrow$ Signo del número [Negativo]
- Siguientes ocho bits: 1000 0100 $\rightarrow$ Característica (exponente sesgado c)

$$c = 1 \cdot 2^7 + 0 \cdot 2^6 + \cdots + 1 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0 = 132$$

Como:

$$(-1)^s \times 2^{c-127} \times (1.f)_2$$

Entonces

$$132 - 127 = 5, \text{ luego, la parte exponencial es:} 2^{132-127} = 2^5$$

- Siguientes veintitres bits: 101 0000 1111 0000 0000 0000$\rightarrow$ Mantisa

$$f = 1 \cdot \left(\frac{1}{2}\right)^1 + 0 \cdot \left(\frac{1}{2}\right)^2 + 1 \cdot \left(\frac{1}{2}\right)^3 + \cdots + 1 \cdot \left(\frac{1}{2}\right)^8 + 1 \cdot \left(\frac{1}{2}\right)^9 + 1 \cdot \left(\frac{1}{2}\right)^{10} + 1 \cdot \left(\frac{1}{2}\right)^{11} + \cdots$$

$$f = \frac{1}{2} + \frac{1}{8} + \frac{1}{256} + \frac{1}{512} + \frac{1}{1024} + \frac{1}{2048}$$

$$(1.f) = (1 + f) = \left(1 + \left[\frac{1}{2} + \frac{1}{8} + \frac{1}{256} + \frac{1}{512} + \frac{1}{1024} + \frac{1}{2048}\right]\right) = \frac{3343}{2048}$$

Errores de computo

$$(-1)^s \times 2^{c-1023} \times (1.f)_2$$

1. Número positivo normalizado más pequeño que se puede representar, cuenta con:
   - $s = 0$
   - $c = 1$
   - $f = 0$

$$2^{-1022} \times (1.0)$$

2. Número positivo normalizado más grande que se puede representar:
   - $s = 0$
   - $c = 2046$
   - $f = 1 - 2^{-52}$

$$2^{1023} \times (2 - 2^{-52})$$

Tenga en cuenta que...

- Los $x < 0.22251 \times 10^{-307} \rightarrow$ Subordinamiento (Subflujo) (tendencia cero)
- Los $x > 0.17977 \times 10^{309} \rightarrow$ Desbordamiento (Sobreflujo) (Se detiene)

$$x = q \times 2^m \qquad \left(\frac{1}{2} \leq q < 1, -126 \leq m \leq 127\right)$$

1. Redondeo correcto: Proceso de reemplazar x por el número de máquina más cercano
2. Error de redondeo: Error implicado

$$x = (0.1b_2b_3 \ldots b_{25}b_{26}b_{27} \ldots)_2 \times 2^m$$

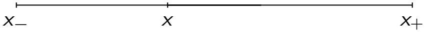
`Redondeo hacia abajo`

$$x_- = (0.1b_2b_3 \ldots b_{24})_2 \times 2^m$$

`Redondeo hacia arriba`

$$x_+ = \left[(0.1b_2b_3 \ldots b_{24})_2 + 2^{-24}\right] \times 2^m$$

$$|x - x_-| \leq \frac{1}{2}|x_+ - x_-| = 2^{-25+m}$$



`Error relativo`

$$\frac{|x - x_-|}{|x|} \leq 2^{-24} = \mathbf{u} \rightarrow \texttt{Error de redondeo unitario}$$