Métodos Numéricos - Curso 2001

Instituto de Matemática (IMERL)

Examen 1 de febrero del 2002

1. Se dispone de un conjunto de observaciones $\{(t_i, y_i), i = 1, ..., n\}$, las que se quieren aproximar por una función de la forma $f(t) = \alpha + \beta t$.

   (a) Explique en que consiste el criterio de aproximación de mínimos cuadrados.

   (b) Deduzca las ecuaciones normales.

   (c) Dados los datos:

t_i	0	1	2
y_i	1.0	2.8	3.9

   encuentre los valores de α y β.

   (d) Para mejorar la aproximación se plantea usar la función $g(t) = \alpha + \beta t + \sin(\omega t)$. Si se sabe que el valor óptimo de los parámetros (α, β, ω) es cercano a $(\alpha_0, \beta_0, \omega_0)$, explique el algoritmo de Gauss-Newton para resolver el problema. Escriba un código Matlab. (Obs.: no se pide resolver el problema numéricamente.)

2. Considere una tabla $\{(x_i, y_i), i = 1, ..., n\}$ de los valores de una función.

   (a) Explique en que consiste la interpolación polinómica y un método para hallar el polinomio interpolante. Compare con la interpolación mediante una función lineal a trozos.

   (b) Deduzca la expresión para el error de interpolación.

   (c) Para la siguiente tabla de valores de la función $f(x) = \sin(x)$:

x_i	0	1	2	3
y_i	0	0.84	0.91	0.14

   halle la expresión del error de interpolación, y una buena cota superior para el mismo en el intervalo [0 3].

3. Suponga que se desea calcular la derivada segunda de la función $f(x)$ en un valor x dado usando la expresión:

$$D_2 f(h) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

   (a) Halle el error cometido, distinguiendo los errores de truncamiento y de redondeo (debido a las operaciones en punto flotante.) Obs.: suponga que x, $x-h$, $x+h$, h^2 y $f(\cdot)$ tienen representación exacta en punto flotante y que el cociente se calcula en forma exacta.

   (b) Discuta el valor de h óptimo.

   (c) Si ya calculó los valores de $D_2 f(h)$ y $D_2 f(h/2)$, ¿cómo usa estos valores para obtener una mejor aproximación de la derivada segunda de f? (Suponga que para el valor de h elegido el error de redondeo es despreciable frente al error de truncamiento).

1. (a) Si se desea aproximar un conjunto de n observaciones $\{(t_i, y_i) : i = 1, \ldots, n\}$ mediante una función del tipo $f(\lambda_1, \ldots, \lambda_m; t)$, donde $\lambda_1, \ldots, \lambda_m$ son m parámetros a ajustar usando el criterio de mínimos cuadrados, lo que se hace es hallar un juego de parámetros $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_m$ que minimice la expresión

$$E(\lambda_1, \ldots, \lambda_m) := \sum_{i=1}^{n} |f(\lambda_1, \ldots, \lambda_m; t_i) - y_i|^2$$

Esta cantidad es una medida del error que se comete al aproximar los datos mediante la función $t \longmapsto f(\lambda_1, \ldots, \lambda_m; t)$.

Para el caso en cuestión, $f(\alpha, \beta; t) = f(t) = \alpha + \beta t$, se deben hallar α_0 y β_0 de forma que $E(\alpha, \beta) = \sum_{i=1}^{n} |\alpha + \beta t_i - y_i|^2$ sea mínimo.

(b) Es fácil ver que se cumple la igualdad $E(\alpha, \beta) = \|Ax - y\|_2^2$, donde

$$A = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

y $\|\cdot\|_2$ es la norma euclídea de $\mathbb{R}^n$. En este caso (que es lineal), el problema de minimizar $E(x)$ es equivalente a las ecuaciones normales $A^t A x = A^t y$.

Se tiene

$$y = \begin{bmatrix} 1,0 \\ 2,8 \\ 3,9 \end{bmatrix}, A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, A^t A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}$$

$$A^t y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2,8 \\ 3,9 \end{bmatrix} = \begin{bmatrix} 7,7 \\ 10,6 \end{bmatrix}$$

Luego $A^t A x = A^t y$ tiene solución $\alpha \cong 1,12$ y $\beta = 1,55$.

(c) Al aproximar los datos con la función $g(\alpha, \beta, \omega; t) = g(t) = \alpha + \beta t + sen(\omega t)$ se obtiene un problema de mínimos cuadrados no lineal, es decir la correspondencia $(\alpha, \beta, \omega) \longmapsto \alpha + \beta t + sen(\omega t)$ no es lineal. Por lo tanto $E(\alpha, \beta, \omega)$ no puede expresarse de la forma $\|Ax - y\|_2^2$ siendo A e y constantes. El método de Gauss-Newton es un método iterativo en el cual se comienza con un juego inicial de parámetros $x^{(0)} = (\alpha_0, \beta_0, \omega_0)$ a partir del cual se obtienen sucesivamente nuevos juegos de parámetros $x^{(1)}, x^{(2)}, \ldots$ de la siguiente forma: Si ya tengo $x^{(k)}$ aproximo $E(x) = \sum_{i=1}^{n} |g(x; t_i) - y_i|^2$ sustituyendo $g(x; t_i) - y_i$ por su desarrollo de Taylor de primer orden en torno a $x^{(k)}$; obteniendo así:

$$E(x) \cong \widetilde{E}(x) = \sum_{i=1}^{n} |g(x^{(k)}; t_i) + \nabla(g|_{t=t_i})(x - x^{(k)}) - y_i|^2 = \|A(x - x^{(k)}) - b\|_2^2$$

donde

$$A = \begin{bmatrix} D_\alpha g(x^{(k)}, t_1) & D_\beta g(x^{(k)}, t_1) & D_\omega g(x^{(k)}, t_1) \\ D_\alpha g(x^{(k)}, t_2) & D_\beta g(x^{(k)}, t_2) & D_\omega g(x^{(k)}, t_2) \\ \vdots & \vdots & \vdots \\ D_\alpha g(x^{(k)}, t_n) & D_\beta g(x^{(k)}, t_n) & D_\omega g(x^{(k)}, t_n) \end{bmatrix}, b = \begin{bmatrix} y_1 - g(x^{(k)}, t_1) \\ y_2 - g(x^{(k)}, t_2) \\ \vdots \\ y_n - g(x^{(k)}, t_n) \end{bmatrix}$$

Ahora se minimiza $\widetilde{E}(x)$, lo que corresponde a un problema de mínimos cuadrados lineal. Se halla así $x^{(k+1)} = x^{(k)} + h$ donde h es la solución del problema $\|Ah - b\|_2^2 \sim 0$.
La iteración se interrumpe, por ejemplo, si $\|x^{(k+1)} - x^{(k)}\|$ es pequeño.

Programa:
Admito ya implementadas dos funciones matlab:

i. $ge(x,t)$ que recibe $x = [\alpha; \beta; \omega]$ y $t = [t_1; \ldots; t_n]$; y devuelve $y = [g(x,t_1); \ldots; g(x,t_n)]$
ii. $Jge(x,t)$ que recibe $x = [\alpha; \beta; \omega]$ y $t = [t_1; \ldots; t_n]$; y devuelve la matriz A antes mencionada.

```
program x=mc(x,y,tol)
h=inf;
while norm(h)>tol
A=Jge(x,t);
b=y-ge(x,t);
h=A\b;
x=x+h;
end
end
```

2. (a) Dadas $n+1$ parejas de reales $\{(x_i, y_i) : i = 0, \ldots, n\}$, con $x_i \neq x_j$ si $i \neq j$, la interpolación polinómica consiste en hallar el (único) polinomio p de grado $\leq n$ que satisface $p(x_i) = y_i$ para todo $i = 0, \ldots, n$.
Dos de las formas de hallar p:

i. Si $p(x) = a_n x^n + \ldots + a_1 x + a_0$, imponiendo las condiciones $p(x_i) = y_i$ para todo $i = 0, \ldots, n$ se obtiene un sistema lineal $(n+1) \times (n+1)$ que permite hallar los coeficientes $a_n, \ldots, a_0$; a saber:

$$\begin{bmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

ii. Considérense los $n+1$ polinomios:

$$L_k(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_0)(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}$$

con $k=0,\dots,n$. Se tiene $L_k(x_l) = \delta_{kl}$ y $gr(L_k) = n$ para todo $k,l = 0,\dots,n$. Luego el polinomio

$$p(x) = \sum_{k=0}^{n} y_k L_k(x)$$

cumple $p(x_i) = y_i$ para todo $i = 0,\dots,n$ y $gr(p) \leq n$, por lo tanto éste es el polinomio buscado.

(b) Error de interpolación:

Sea $[a,b]$ un intervalo que contenga a $x_0,\dots,x_n$ y a un punto x. Supongan además que $y_i = f(x_i)$ para todo $i = 0,\dots,n$, siendo $f : [a,b] \longrightarrow \mathbb{R}$ continua y con $n+1$ derivadas en (a,b). Si p es el polinomio interpolante de f en los puntos $x_0,\dots,x_n$ se tiene:

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x-x_0)(x-x_1)\dots(x-x_n)$$

para un cierto punto $\xi_x \in (a,b)$.

Prueba:

Sea $F : [a,b] \longrightarrow \mathbb{R}$ dada por:

$$F(t) = [f(x) - p(x)]A(t) - [f(t) - p(t)]A(x)$$

donde $A(t) = (t-x_0)(t-x_1)\dots(t-x_n)$. F es continua y con $n+1$ derivadas en (a,b). Se puede suponer que $x \neq x_i$, $i = 0,\dots,n$, en cuyo caso la tesis es obvia. F se anula en los puntos $x, x_0,\dots,x_n$; estos $n+2$ puntos determinan $n+1$ subintervalos consecutivos de $[a,b]$ en cuyos extremos F es 0. Aplicando el teorema de Rolle a cada intervalo obtenemos $n+1$ puntos de (a,b) en donde F' se anula. Considerando ahora los $n+1$ intervalos determinados por estos puntos y aplicando el teorema de Rolle a F' se obtienen n puntos de (a,b) en donde F'' se anula. Repitiendo este proceso se llega finalmente a que existe un punto ξ en (a,b) tal que $F^{(n+1)}(\xi) = 0$. Derivando $n+1$ veces F en la fórmula dada antes, teniendo en cuenta que $A^{(n+1)}(t) = (n+1)!$ y que $p^{(n+1)}(t) = 0$ (ya que $gr(p) \leq n$) se llega a:

$$0 = F^{(n+1)}(\xi) = [f(x) - p(x)](n+1)! - f^{(n+1)}(\xi)A(x)$$

De aquí resulta la tesis.

(c) Según la parte anterior se tiene:

$$E(x) := f(x) - p(x) = \frac{sen(\xi_x)}{24}x(x-1)(x-2)(x-3)$$

Acotando $|sen(\xi_x)| \leq 1$ se obtiene:

$$|E(x)| \leq \frac{|x||x-1||x-2||x-3|}{24}$$

Ahora hay que acotar el término $|x||x-1||x-2||x-3|$ para $x \in [0,3]$:

i. Si se acota a cada factor por 3 y se obtiene: $|E(x)| \leq \frac{3^4}{24} = \frac{27}{8} \cong 3,3$.

ii. Mirando la pareja de factores $|x|$, $|x-3|$ se ve que cualquiera sea $x \in [0,3]$ al menos uno de ellos es $\leq 1,5$. Asimismo uno de los dos factores $|x-1|$ o $|x-2|$ es ≤ 1. Luego se obtiene $|E(x)| \leq \frac{3 \cdot 1,5 \cdot 3 \cdot 1}{24} = \frac{9}{16} \cong 0,56$.

iii. Otra forma de acotar es hallar los extremos de $A(x) := x(x-1)(x-2)(x-3)$ en $[0,3]$. Se tiene $A'(x) = 4x^3 - 18x^2 + 22x - 6$. Se ve fácilmente que función $A(x)$ es par respecto a la abscisa $x = 1,5$ por tanto $1,5$ es raíz de $A'(x)$. Bajando $A'(x)$ por Ruffini (usando $1,5$) resulta un polinomio de segundo grado, y se pueden hallar todas las raíces de $A'(x)$. Estas son: $1,5$ y $(3 \pm \sqrt{5})/2$; calculando $A(1,5) \cong 0,56$ y $A((3 \pm \sqrt{5})/2) = -1$ se concluye que $|A(x)| \leq 1$ y por lo tanto $|E(x)| \leq \frac{1}{24} \cong 0,04$.

3. Sea

$$D_2 f(h) := \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

(a) Error de truncamiento:

Efectuando el desarrollo de Taylor de f en torno a x se tiene:

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2} + f'''(x)\frac{h^3}{6} + f^{(IV)}(\xi)\frac{h^4}{24}$$

$$f(x-h) = f(x) - f'(x)h + f''(x)\frac{h^2}{2} - f'''(x)\frac{h^3}{6} + f^{(IV)}(\zeta)\frac{h^4}{24}$$

donde $\xi \in (x, x+h)$ y $\zeta \in (x-h, x)$. Sumando y operando se obtiene:

$$D_2 f(h) = f''(x) + (f^{(IV)}(\xi) + f^{(IV)}(\zeta))\frac{h^2}{24} = f''(x) + E_T(h)$$

Si M es una cota de $f^{(IV)}$ en $(x-h, x+h)$ se obtiene la siguiente cota para el error de truncamiento:

$$|E_T(h)| \leq \frac{Mh^2}{12}$$

Error de redondeo: Al evaluar la expresión $D_2 f(h)$ se efectúan diversos errores:

i. Representación en PF de x y h.

ii. Representación en PF de $x+h$, $x-h$ y h^2.

iii. Evaluación y representación en PF de $f(x)$, $f(x+h)$ y $f(x-h)$.

iv. Representación en PF de la suma $f(x+h) - 2f(x) + f(x-h)$.

v. Representación en PF del cociente.

De todos estos errores sólo se considerarán los mencionados en (iv). Por lo tanto se tiene:

$$PF[D_2 f(h)] = \frac{PF[PF(f(x+h) + f(x-h)) - PF(2f(x))]}{h^2} =$$

$$\frac{[(f(x+h) + f(x-h))(1+\delta_1) - 2f(x)(1+\delta_2)](1+\delta_3)}{h^2} \cong$$

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2}(1+2\delta) = D_2f(h) + E_R(h)$$

donde $|\delta| \leq \epsilon_{match}$.(Se está haciendo una suposición sobre como se asocian los sumandos. Asociaciones diferentes llevan a cotas que difieren en una constante que también se considerarán correctas). Si N es una cota de f se obtiene:

$$|E_R(h)| = \frac{|f(x+h) - 2f(x) + f(x-h)||2\delta|}{h^2} \leq \frac{8N\epsilon_{match}}{h^2}$$

El error total $E(h) = E_T(h) + E_R(h)$ resulta entonces:

$$|E(h)| \leq \frac{Mh^2}{12} + \frac{8N\epsilon_{match}}{h^2} =: c(h)$$

(b) El valor de h óptimo es aquel que minimiza la cota, $c(h)$, del error. Se tiene:

$$c'(h) = \frac{Mh}{6} - \frac{16N\epsilon_{match}}{h^3}$$

Luego:

$$c'(h) = 0 \iff h = h_{opt} = \sqrt[4]{\frac{48N\epsilon_{match}}{M}} \cong \sqrt[4]{\epsilon_{match}}$$

(c) Se tiene:

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2} + f'''(x)\frac{h^3}{6} + f^{(IV)}(x)\frac{h^4}{24} + f^{(V)}(x)\frac{h^5}{120} + o(h^5)$$

$$f(x-h) = f(x) - f'(x)h + f''(x)\frac{h^2}{2} - f'''(x)\frac{h^3}{6} + f^{(IV)}(x)\frac{h^4}{24} - f^{(V)}(x)\frac{h^5}{120} + o(h^5)$$

Sumando y operando resulta:

$$D_2f(h) = f''(x) + f^{(IV)}(x)\frac{h^2}{24} + o(h^3)$$

Sustituyendo aquí h por $h/2$ se tiene:

$$D_2f(h/2) = f''(x) + f^{(IV)}(x)\frac{h^2}{4{\cdot}24} + o(h^3)$$

En consecuencia:

$$\frac{4D_2f(h/2) - D_2f(h)}{3} = f''(x) + o(h^3)$$

Esta última expresión es un mejor estimador de $f''(x)$ ya que es de orden 3, en tanto que $D_2f(h)$ es de orden 1, por lo que cabe esperar que sea una mejor aproximación de $f''(x)$.