Tarea 3

Roberto Gutiérrez-Sánchez, Jimena Salas-Alfaro
rogutierrez@estudiantec.cr jimenamaria598@gmail.com
Instituto Tecnológico de Costa Rica
Escuela de Ingeniería Electrónica
Introducción al Reconocimiento de Patrones

I. Relaciones empleadas en las demostraciones

Sea s un escalar:

$$\mathbf{AI} = \mathbf{A} = \mathbf{IA} \tag{1}$$

$$tr(s) = s \tag{2}$$

$$\nabla_{\mathbf{A}^T} tr(\mathbf{ABA}^T\mathbf{C}) = \mathbf{B}^T\mathbf{A}^T\mathbf{C}^T + \mathbf{BA}^T\mathbf{C} \tag{3}$$

Si $\mathbf{AB}$ es cuadrada

$$\nabla_{\mathbf{A}} tr(\mathbf{BA}) = \nabla_{\mathbf{A}} tr(\mathbf{AB}) = \mathbf{B}^T \tag{4}$$

$$\nabla_A f(\mathbf{A}) = (\nabla_{A^T} f(\mathbf{A}))^T \tag{5}$$

II. Demostraciones matemáticas

II-A. *Ejercicio 1.1*

Si se sabe que con los vectores $\underline{\mathbf{u}}$ y $\underline{\mathbf{v}}$ del mismo tamaño se cumple que

$$\underline{\mathbf{u}}^T\underline{\mathbf{v}} = \sum_{i=1}^{n} u_i v_i \tag{6}$$

Se puede convertir la siguiente suma

$$\sum_{i=1}^{n} w^{(i)}(\underline{\theta}^T\underline{\mathbf{x}}^{(i)} - y^{(i)})^2 = \sum_{i=1}^{n} u_i w^{(i)} u_i = \underline{\mathbf{u}}^T\underline{\mathbf{v}} \tag{7}$$

Con $u_i = \underline{\theta}^T\underline{\mathbf{x}}^{(i)} - y^{(i)}$ y $v_i = w^{(i)}u_i$
Generalizando para el vector $\underline{\mathbf{u}}$, se tiene que

$$\underline{\mathbf{u}} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix} \underline{\theta} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \mathbf{X}\underline{\theta} - \underline{\mathbf{y}} \tag{8}$$

Ahora para el vector $\underline{\mathbf{v}}$, para escalar cada elemento de un vector se debe multiplicar por una matriz diagonal con las escalas

$$\underline{\mathbf{v}} = \begin{bmatrix} w^{(1)}u^{(1)} \\ w^{(2)}u^{(2)} \\ \vdots \\ w^{(n)}u^{(n)} \end{bmatrix} = \begin{bmatrix} w^{(1)} & 0 & \dots & 0 \\ 0 & w^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w^{(n)} \end{bmatrix} \begin{bmatrix} u^{(1)} \\ u^{(2)} \\ \vdots \\ u^{(n)} \end{bmatrix} = \mathbf{W}\underline{\mathbf{u}} = \mathbf{W}(\mathbf{X}\underline{\theta} - \underline{\mathbf{y}}) \tag{9}$$

De esta forma, se tiene entonces que

$$\sum_{i=1}^{n} w^{(i)}(\underline{\theta}^T\underline{\mathbf{x}}^{(i)} - y^{(i)})^2 = \sum_{i=1}^{n} u_i w^{(i)} u_i = \underline{\mathbf{u}}^T\underline{\mathbf{v}} = (\mathbf{X}\underline{\theta} - \underline{\mathbf{y}})^T\mathbf{W}(\mathbf{X}\underline{\theta} - \underline{\mathbf{y}}) \tag{10}$$

II-B. Ejercicio 1.2

Sea la matriz de diseño $\mathbf{X} \in \mathbb{R}^{m\times n}$, el vector $\underline{\theta}$ de tamaño n, el vector $\underline{\mathbf{y}}$ de tamaño m, y la matrix diagonal $\mathbf{W} \in \mathbb{R}^{m\times m}$

Se define entonces $J(\underline{\theta}) = \frac{1}{2}(\mathbf{X}\underline{\theta} - \underline{\mathbf{y}})^T\mathbf{W}(\mathbf{X}\underline{\theta} - \underline{\mathbf{y}})$, y se debe de calcular $\nabla_{\underline{\theta}}J(\underline{\theta}) = 0$

Procedimiento:

$$J(\underline{\theta}) = \frac{1}{2}(\underline{\theta}^T\mathbf{X}^T - \underline{\mathbf{y}}^T)(\mathbf{W}\mathbf{X}\underline{\theta} - \mathbf{W}\underline{\mathbf{y}}) \tag{11}$$

$$J(\underline{\theta}) = \frac{1}{2}(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta} - \underline{\mathbf{y}}^T\mathbf{W}\mathbf{X}\underline{\theta} - \underline{\theta}^T\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}} + \underline{\mathbf{y}}^T\mathbf{W}\underline{\mathbf{y}}) \tag{12}$$

De esta manera se tiene que el gradiente es la suma de gradientes

$$\nabla_{\underline{\theta}}J(\underline{\theta}) = \frac{1}{2}(\nabla_{\underline{\theta}}(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta}) - \nabla_{\underline{\theta}}(\underline{\mathbf{y}}^T\mathbf{W}\mathbf{X}\underline{\theta}) - \nabla_{\underline{\theta}}(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}) + \nabla_{\underline{\theta}}(\underline{\mathbf{y}}^T\mathbf{W}\underline{\mathbf{y}})) \tag{13}$$

Para el primer término se usa (2) y (3), ya que el valor interno tiene tamaño $(1xn)(nxm)(mxm)(mxn)(nx1) = (1x1)$, o sea es un escalar. Y con $\mathbf{A} = \underline{\theta}^T$, $\mathbf{B} = \mathbf{X}^T\mathbf{W}\mathbf{X}$, $\mathbf{C} = \mathbf{I}$, se obtiene entonces

$$\nabla_{\underline{\theta}}(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta}) = \nabla_{\underline{\theta}}tr(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta}) = \mathbf{X}^T\mathbf{W}^T\mathbf{X}\underline{\theta} + \mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta} \tag{14}$$

Para el segundo término, se utiliza la ecuación (4), junto a la (2) ya que el resultado es escalar

$$-\nabla_{\underline{\theta}}(\underline{\mathbf{y}}^T\mathbf{W}\mathbf{X}\underline{\theta}) = -\nabla_{\underline{\theta}}tr(\underline{\mathbf{y}}^T\mathbf{W}\mathbf{X}\underline{\theta}) = -\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}} \tag{15}$$

Para el tercer término se usan las ecuaciones (4) y la (5), junto a la (2) ya que el resultado es escalar

$$-\nabla_{\underline{\theta}}(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}) = -\nabla_{\underline{\theta}}tr(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}) = -(\nabla_{\underline{\theta}^T}tr(\underline{\theta}^T\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}))^T = -(\mathbf{y}^T\mathbf{W}^T\mathbf{X})^T = -\mathbf{X}^T\mathbf{W}^T\underline{\mathbf{y}} \tag{16}$$

El cuarto término no depende de $\underline{\theta}$, por lo tanto el gradiente es 0

De esta manera se obtiene que

$$\nabla_{\underline{\theta}}J(\underline{\theta}) = \frac{1}{2}(\mathbf{X}^T\mathbf{W}^T\mathbf{X}\underline{\theta} + \mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta} - \mathbf{X}^T\mathbf{W}\underline{\mathbf{y}} - \mathbf{X}^T\mathbf{W}^T\underline{\mathbf{y}}) \tag{17}$$

Dado que $\mathbf{W}$ es diagonal entonces $\mathbf{W}^T = \mathbf{W}$ y esa ecuación se convierte en

$$\nabla_{\underline{\theta}}J(\underline{\theta}) = \frac{1}{2}(2\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta} - 2\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}) = \mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\theta} - \mathbf{X}^T\mathbf{W}\underline{\mathbf{y}} \tag{18}$$

A partir de esta ecuación se puede despejar para $\nabla_{\underline{\theta}}J(\underline{\theta}) = 0$

$$\underline{\theta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}} \tag{19}$$

II-C. Ejercicio 1.3

La verosimilitud se define como

$$L(\underline{\theta}) = p(\mathbf{Y}|\mathbf{X};\underline{\theta}) \tag{20}$$

En donde contodos los m datos independientes se convierte en

$$L(\underline{\theta}) = p(\mathbf{Y}|\mathbf{X};\underline{\theta}) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\underline{\theta}) \tag{21}$$

Realizando la suposición que

$$p(y^{(i)}|x^{(i)};\underline{\theta}) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\Bigg(- \frac{(y^{(i)} - \underline{\theta}^T\underline{\mathbf{x}})^2}{2(\sigma^{(i)})^2}\Bigg) \tag{22}$$

Se obtiene

$$L(\underline{\theta}) = p(\mathbf{Y}|\mathbf{X};\underline{\theta}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\Bigg(- \frac{(y^{(i)} - \underline{\theta}^T\underline{\mathbf{x}})^2}{2(\sigma^{(i)})^2}\Bigg) \tag{23}$$

Debido a que se le debe encontrar el $\underline{\theta}^* = arg\max_{\underline{\theta}} L(\underline{\theta})$

Esto es equivalente a encontrar $\underline{\theta}^* = arg\max_{\underline{\theta}} ln(L(\underline{\theta}))$, el log-likelihood, el cual tiene ecuación

$$\ell(\underline{\theta}) = ln(L(\underline{\theta})) = ln\Bigg(\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\Bigg(- \frac{(y^{(i)} - \underline{\theta}^T\underline{\mathbf{x}})^2}{2(\sigma^{(i)})^2}\Bigg)\Bigg) = \sum_{i=1}^{m} ln\Bigg(\frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\Bigg(- \frac{(y^{(i)} - \underline{\theta}^T\underline{\mathbf{x}})^2}{2(\sigma^{(i)})^2}\Bigg)\Bigg) \tag{24}$$

Desarrollando esta suma

$$\ell(\underline{\theta}) = \sum_{i=1}^{m} ln\Bigg(\frac{1}{\sqrt{2\pi}\sigma^{(i)}}\Bigg) - \Bigg(\frac{(y^{(i)} - \underline{\theta}^T\underline{\mathbf{x}})^2}{2(\sigma^{(i)})^2}\Bigg) \tag{25}$$

El primer término de la suma no afecta el máximo de $\ell(\underline{\theta})$, por lo tanto

$$\underline{\theta}^* = arg\max_{\underline{\theta}} \Bigg(\frac{-1}{2} \sum_{i=1}^{m} \frac{1}{(\sigma^{(i)})^2} (\underline{\theta}^T\underline{\mathbf{x}} - y^{(i)})^2 \Bigg) \tag{26}$$

De aquí se observa que esta ecuación es igual a la ponderada considerando que $w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$, solamente para minimizar el error J, maximizando la verosimilitud.

III. Gráficas Generadas

Aquí para la figura 2 se puede apreciar que conforme aumenta el valor de τ, la curva se va linealizando, ya que el término $w^{(i)} = exp(-\frac{||\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(i)}||^2}{2\tau^2})$ tiende a 1 y por lo tanto la regresión se va convirtiendo en lineal.

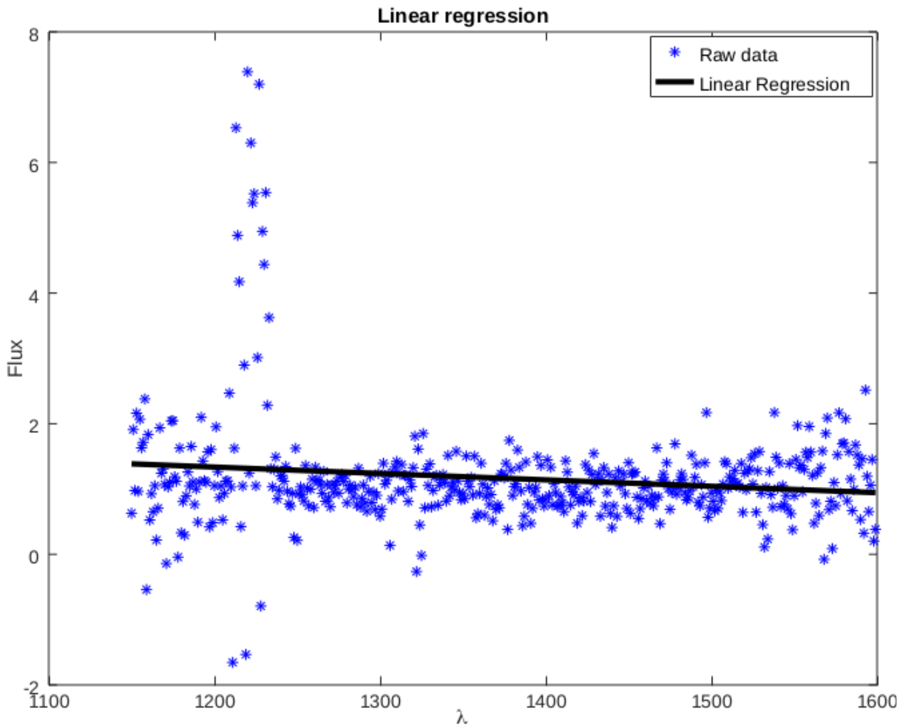

Figura 1: Regresión lineal

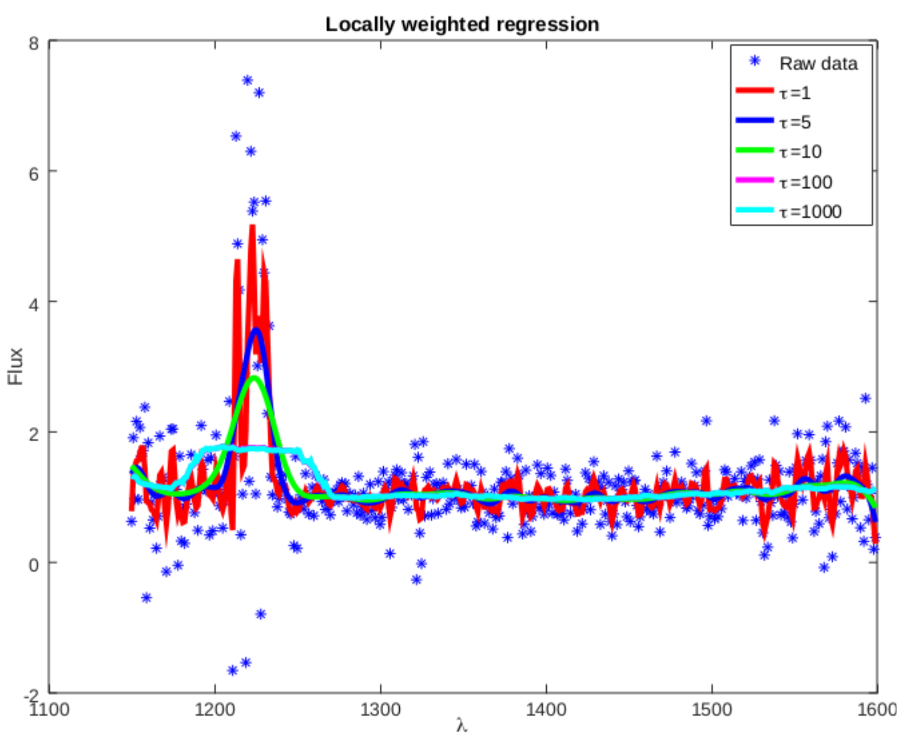

Figura 2: Regresión localmente ponderada