Instituto Tecnolgico de Costa Rica — EL-5852 Introduccin al Reconocimiento de Patrones
Escuela de Ingeniera Electrnica — Sebastian Vargas Zuiga 2016138236
Profesor: Dr. Pablo Alvarado Moya — Luis Alonso Vega Badilla 2017110213
II Semestre 2019 — March 31, 2020

# Tarea 3

1. Considere un problema de regresin lineal en el que queremos ponderar de forma distinta cada ejemplo de entrenamiento, tal y como vimos en clase con el mtodo de regresin ponderada localmente. Especficamente, queremos minimizar

$$J(\underline{\theta}) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\underline{\theta}^T \underline{x}^{(i)} - y^{(i)})^2$$

   (a) Demuestre que para el caso general $J(\underline{\theta})$ se puede reescribir como

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

   para una matriz diagonal W apropiada donde X es la matriz de diseo e $\underline{y}$ el vector de salidas, tal y como lo definimos en clase.
   Solucin:
   Para esta parte lo que se har ser desarrollar la expresin de $J(\underline{\theta})$ con X y W como:

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,n} \end{bmatrix}$$

   donde, cada fila de la matriz representa un vector de valores de entrada.

$$W = \begin{bmatrix} W_{1,1} & 0 & \cdots & 0 \\ 0 & W_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_{n,n} \end{bmatrix}$$

   Para el desarrollo $\mathrm{W}_{1,1} = W_1, W_{2,2} = W_2, \cdots, W_{n,n} = W_n$.

Desarrollando:

$$J(\underline{\theta}) = \frac{1}{2}\left(\begin{bmatrix} X_{1,:} \\ X_{2,:} \\ \vdots \\ X_{n,:} \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}\right)^T \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_n \end{bmatrix}\left(\begin{bmatrix} X_{1,:} \\ X_{2,:} \\ \vdots \\ X_{n,:} \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}\right)$$

$$J(\underline{\theta}) = \frac{1}{2}\left(\begin{bmatrix} X_{1,:}\cdot\underline{\theta} - y_1 \\ X_{2,:}\cdot\underline{\theta} - y_2 \\ \vdots \\ X_{n,:}\cdot\underline{\theta} - y_n \end{bmatrix}\right)^T \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_n \end{bmatrix}\left(\begin{bmatrix} X_{1,:}\cdot\underline{\theta} - y_1 \\ X_{2,:}\cdot\underline{\theta} - y_2 \\ \vdots \\ X_{n,:}\cdot\underline{\theta} - y_n \end{bmatrix}\right)$$

$$J(\underline{\theta}) = \frac{1}{2}\left(\begin{bmatrix} W_1(X_{1,:}\cdot\underline{\theta} - y_1) \\ W_2(X_{2,:}\cdot\underline{\theta} - y_2) \\ \vdots \\ W_n(X_{n,:}\cdot\underline{\theta} - y_n) \end{bmatrix}\right)^T \left(\begin{bmatrix} X_{1,:}\cdot\underline{\theta} - y_1 \\ X_{2,:}\cdot\underline{\theta} - y_2 \\ \vdots \\ X_{n,:}\cdot\underline{\theta} - y_n \end{bmatrix}\right)$$

$$J(\underline{\theta}) = \frac{1}{2}[W_1(X_{1,:}\cdot\underline{\theta} - y_1)^2 + W_2(X_{2,:}\cdot\underline{\theta} - y_2)^2 + \cdots + W_n(X_{n,:}\cdot\underline{\theta} - y_n)^2]$$

$$\therefore J(\underline{\theta}) = \frac{1}{2}\sum_{i=1}^{n} W_i(X_{i,:}\cdot\underline{\theta} - y_i)^2$$

(b) Si todos los pesos $w^{(i)}$ son iguales a 1, entonces en clase vimos que la ecuacin normal es simplemente

$$X^T X\underline{\theta} = X^T\underline{y}$$

y el valor de $\underline{\theta}$ que minimiza $J(\underline{\theta})$ est dado por $(X^TX)^{-1}X^Ty$.
Encuentre el gradiente $\nabla_{\underline{\theta}}J(\underline{\theta})$ e igualelo a cero para encontrar una versin generalizada de las ecuaciones normales en este contexto con ponderacin. Encuentre una forma cerrada del valor de $\underline{\theta}$ que minimiza a J($\underline{\theta}$), en funcion de X, W e $\underline{\text{y}}$.
Solucion:
El mnimo se encuentra buscando $\nabla_{\underline{\theta}}J(\underline{\theta}) = 0$

$$\nabla_{\underline{\theta}}J(\underline{\theta}) = \frac{1}{2}\nabla_{\underline{\theta}}(X\underline{\theta} - \underline{y})^T W(X\underline{\theta} - \underline{y}) = \underline{0}$$

$$\nabla_{\underline{\theta}}J(\underline{\theta}) = \nabla_{\underline{\theta}}\frac{1}{2}(\underline{\theta}^T X^T W X\underline{\theta} - \underline{\theta}^T X^T W\underline{y} - \underline{y}^T W X\underline{\theta} + \underline{y}^T W\underline{y}) = \underline{0}$$

Puesto que la expresin entre parntesis es un escalar real

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

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

y utilizando $\nabla_{A^T}ABA^TC = B^TA^TC^T + BA^TC$, con $A = \underline{\theta}^T$,$B = XX^T$ y C=W

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

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

$$\nabla_{\underline{\theta}}J(\underline{\theta}) = X^T W X\underline{\theta} - X^T W\underline{y}$$

Recordando que para minimizar $\underline{\theta}$ el gradiente $\Delta_{\underline{\theta}} J(\underline{\theta}) = \underline{0}$

$$X^T W X \underline{\theta} = X^T W \underline{y}$$

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

(c) Suponga que tenemos un conjunto de entrenamiento $\{(\underline{x}^{(i)}, y^{(i)}); i = 1 \cdots m\}$ de m ejemplos independientes, pero en el que los $y^{(i)}$ se observaron con varianzas distintas. Especficamente, suponga que

$$p(y^{(i)}, \underline{x}^{(i)}; \underline{\theta}) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\left(-\frac{(y^{(i)} - \underline{\theta}^T \underline{x}^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

o en otras palabras, $y^{(i)}$ tiene media $\underline{\theta}^T x^{(i)}$ y varianza $(\sigma^{(i)})^2$, donde las $\sigma^{(i)}$ son constantes fijas, conocidas. Demuestre que encontrar el estimado de mxima verosimilitud de  se reduce a resolver un problema de regresin lineal ponderada. Establezca claramente que los $w^{(i)}$ se calculan en trminos de $\sigma^{(i)}$.

Solucin:

La mxima verosimilitud se basa en elegir $\underline{\theta}$ de modo que los datos sean lo mas probables posibles.

Para esto, se utiliza el logaritmo natural(ln), es decir, verosimilitud logartmica.

$$\ell(\underline{\theta}) = lnL(\underline{\theta})$$

$$\ell(\underline{\theta}) = ln \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\left(-\frac{(y^{(i)} - \underline{\theta}^T \underline{x}^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

$$\ell(\underline{\theta}) = \sum_{i=1}^{m} ln \frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp\left(-\frac{(y^{(i)} - \underline{\theta}^T \underline{x}^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

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

Donde el primer trmino es una constante, sacando el $\frac{1}{2}$ por distributividad y haciendo a $w^{(i)} = \frac{1}{\sigma^{(i)2}}$ se puede observar que el segundo trmino tiene la forma de la funcin de error en una RL ponderada por lo que se puede concluir que para maximizar la verosimilitud se requiere minimizar "la funcin de error" al igual que la misma RL.

2. Visualizacin de los datos

   (a) Use las ecuaciones normales para implementar la regresin lineal (no ponderada) $y = \underline{\theta}^T \underline{x}$ en el primer ejemplo de entrenamiento (esto es, la primera fila que no es el encabezado). En una figura, grafique tanto los datos crudos como la lnea recta resultante del ajuste. Indique el $\underline{\theta}$ ptimo resultante de la regresin lineal

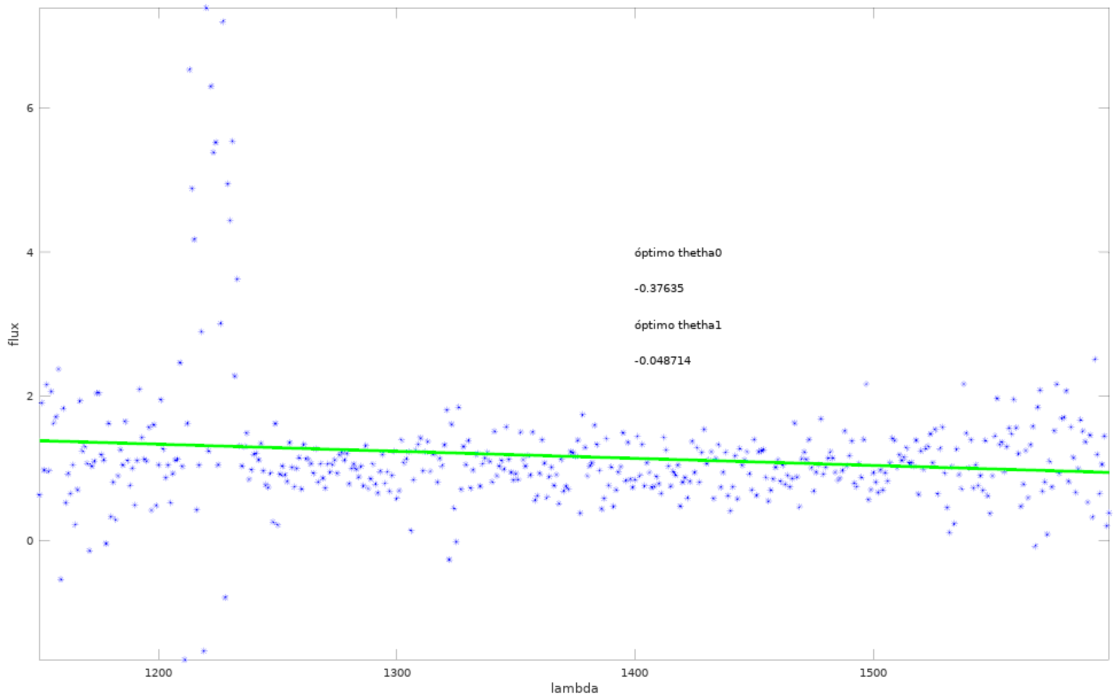


**Figura 1:** Regresión lineal no ponderada.

(b) Implemente la regresin lineal ponderada localmente en el primer ejemplo de entrenamiento. Use las ecuaciones normales que usted derivo en el punto 1.2. En una figura aparte, grafique tanto los datos crudos como la curva suave resultante de su regresin. Cuando evalue $h(\cdot)$ en un punto $\underline{x}$ use los pesos

$$w^{(i)} = exp\left(-\frac{\|\underline{x} - \underline{x}^{(i)}\|^2}{2\tau^2}\right)$$

con el parmetro de ancho de banda $\tau = 5$

(c) Repita el punto 2.2 cuatro veces ms con $\tau = 1$, 10, 100 y 1000. Grafique las curvas resultantes en una misma figura. Indique en una frase corta que ocurre a la curva de regresin conforme $\tau$ crece.

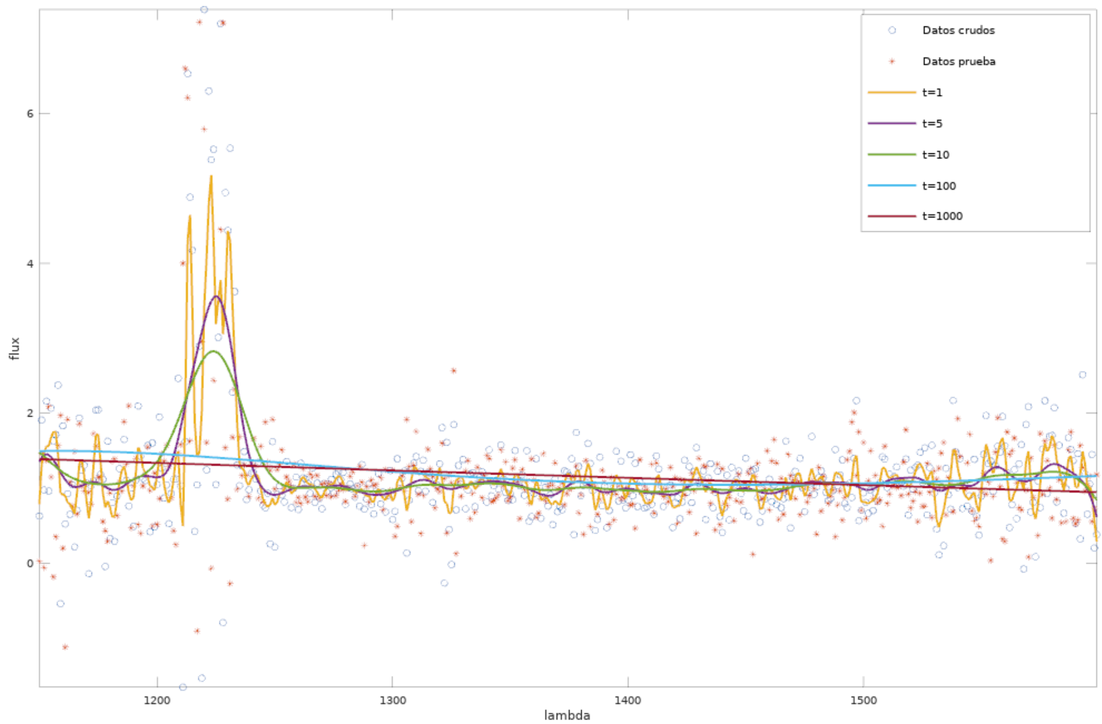


**Figura 2:** Regresión lineal ponderada para distintos valores de tau.

En la figura se puede notar que conforme el valor de tau dismuye, el modelo se aproxima a los datos de entrenamiento, mientras que si aumenta, la hiptesis se aproxima a la hiptesis de regresin lineal, por lo tanto, a valores altos de tau se genera subajuste y a valores pequeos provocan un sobreajuste de la hiptesis.