

Notas de Clases :

Introducción al Machine
Learning aplicado a Audio

Notación :

1) x, x_i, y_i : escalares

2) \mathbf{x}, \mathbf{y} : vectores

$\mathbf{x} = [x_1, x_2, \dots, x_m]$ (Ejemplo de vector fila)

3) X, Y : Conjuntos o Matrices

4) X^T : matriz traspuesta

5) Las ecuaciones las numeraré con parentesis,
Por ejemplo :

$$E(x) = -\sum_{i=1}^m F(\langle u_i, x \rangle) \quad (0)$$



indicando
que esta
es la "eqn"
equación (0)

6) \mathbb{R} : números reales

\mathbb{Z}^+ : enteros positivos

7) \Leftrightarrow ó $\hat{\Leftrightarrow}$: Equivalente

Regresión Lineal

(Clase 02)

22/09/2023

Definición del problema : Dado un conjunto de puntos $(x_1, y_1), \dots, (x_m, y_m)$ queremos encontrar el parámetro θ de una función $g_\theta(\cdot)$ tal que queremos minimizar la función de pérdida \mathcal{L} :

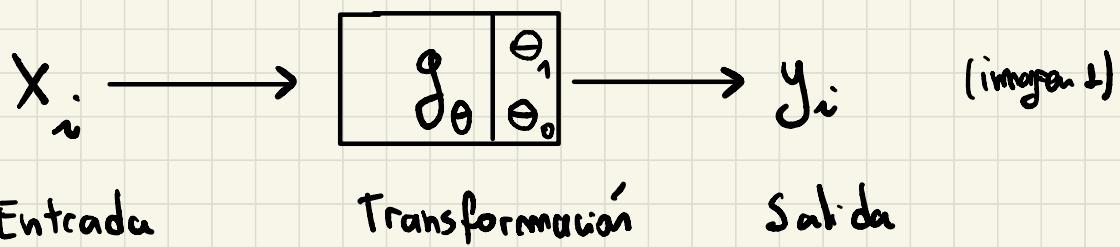
$$\hat{\theta} : \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^m \mathcal{L}(y_i, g_\theta(x_i)) \quad (1)$$

Observemos el siguiente ejemplo para entender la ecuación (1) y desglosar los conceptos. Ej. 1 :

Datos de entrada	Datos de Salida	
x_1	y_1	
x_2	y_2	
\vdots	\vdots	
x_m	y_m	

Podríamos pensar esto como mediciones!

Entonces, dados un modelo $g \in \mathcal{G}^{(1)}$, queremos lograr lo siguiente:

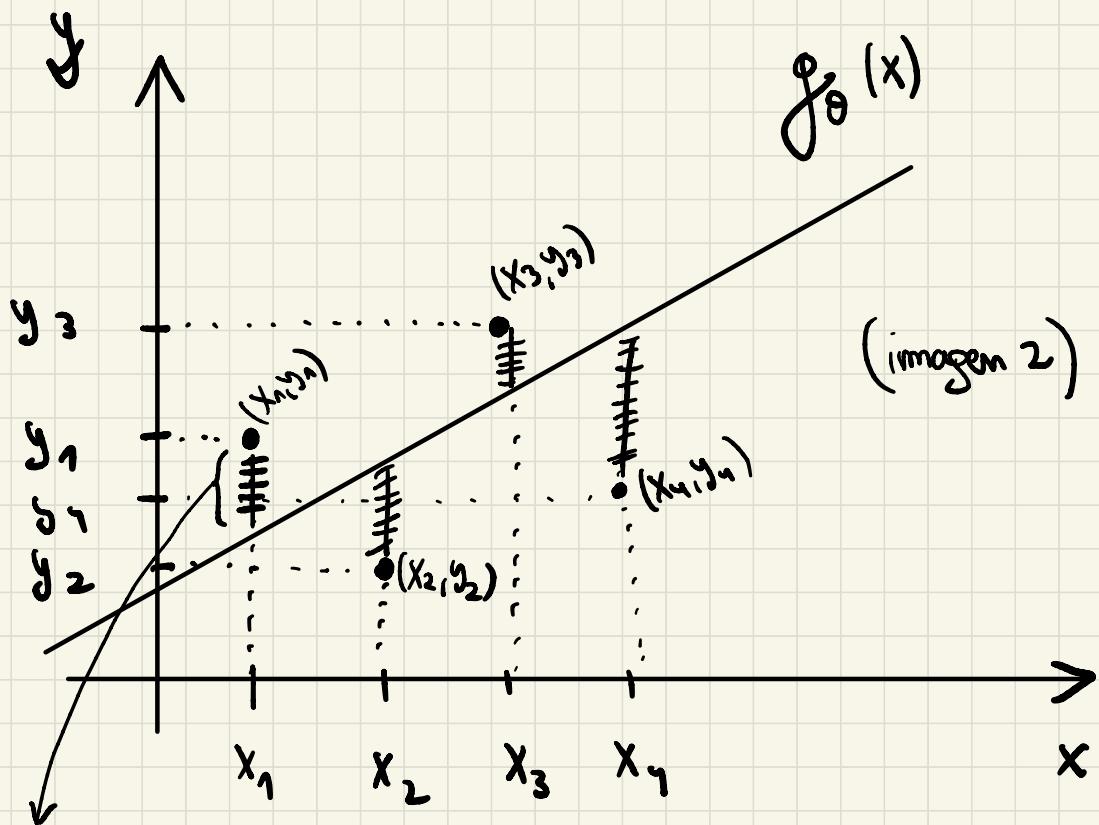


Observe que $f_{\theta}(x) = \theta_1 x + \theta_0$ tiene dos parámetros θ_1 , θ_0 que podemos pensar como "perillas" que debemos ajustar para que $f_{\theta}(x_i) = y_i$ la mayor cantidad de veces si es posible, θ también es una recta!

De alguna forma debemos medir el error que comete el modelo. Una forma común de medir error es utilizando el "error cuadrático medio" dado por :

$$\frac{1}{2m} \sum_{i=1}^m \left(y_i - g_\theta(x_i) \right)^2 \quad (2)$$

Es decir, en (2) compararemos la salida ideal " y_i " con la aproximación de nuestro modelo " $g_\theta(x_i)$ ". Gráficamente, si (x_i, y_i) estuvieran en 2 dimensiones y tuviéramos 4 mediciones :



la distancia $\| \cdot \|$
es el error cometido por el modelo.

Dado el ejemplo, desglosemos la notación:

$$\hat{\theta} : \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^m l(y_i, g_\theta(x_i)) \quad (1)$$

Queremos encontrar

$\hat{\theta}$ en \mathbb{R}^2 pues $\theta = [\theta_0, \theta_1]$

que minimice la función

de costo (2)

Recumiendo:

θ : Vector de parámetros

d : dimensión del vector de parámetros

l : pérdida y $\sum_{i=1}^m l$ es el costo.

$\hat{\theta}$: solución del problema

Para resolver este problema usaremos 2 caminos diferentes. La finalidad es comparar ambas perspectivas.

1) A través de cálculo :

- Calculamos las derivadas parciales de la función de costo e igualamos a cero.
- Observe que si cambiarmos variables para (2) de la siguiente forma :

$$\frac{1}{2m} \sum_{i=1}^m \left(y_i - g_{\theta}(x_i) \right)^2 \quad (3)$$

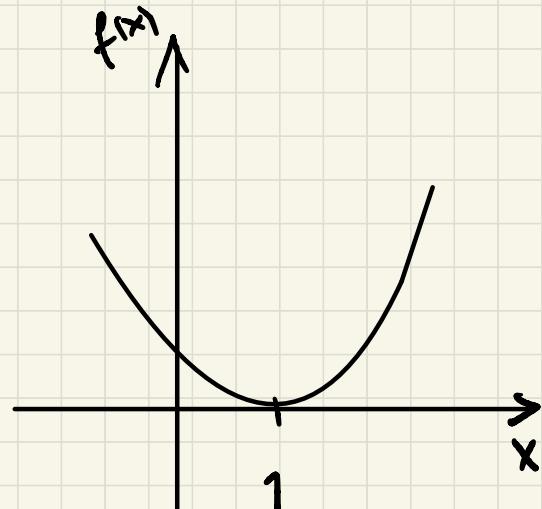
↓

$$\frac{L}{2m} \sum_{i=1}^m \mathcal{Z}^2$$

donde $\mathcal{Z} = (y_i - g_{\theta}(x_i))$ j estamos ante una suma de parábolas.

Es decir, si pensamos simplificadamente el problema, todas ellas tienen un valor mínimo!. Ejemplo 2:

$$f(x) = (x - 1)^2$$



Derivaremos :

$$\dot{f}(x) = 2(x - 1)$$

$$\ddot{f}(x) = 2$$

- Como $\ddot{f}(x) > 0$ la función posee mínimo
 - Luego, $2(x - 1) = 0$
- $$2x = 2 \quad (=) \quad \boxed{x = 1}$$
- La función alcanza su mínimo en $x=1$.

Volviendo a nuestro problema : Queremos encontrar Θ que tiene 2 variables.
 \therefore Debemos derivar en 2D (Gradiente)

O de forma equivalente :

$$\frac{\partial}{\partial \theta_0} \left(\frac{1}{2m} \sum_{i=1}^m (y_i - g_{\theta}(x_i))^2 \right) = 0$$

$$\frac{\partial}{\partial \theta_1} \left(\frac{1}{2m} \sum_{i=1}^m (y_i - g_{\theta}(x_i))^2 \right) = 0$$

Comentario !!! la expresión anterior es equivalente a:

$$\nabla_{\theta} \left(\frac{1}{2m} \sum_{i=1}^m (y_i - g_{\theta}(x_i))^2 \right) = 0$$

Calculemos (después de algunas wentas) :

$$\theta_1 \sum_{i=1}^m x_i^2 + \theta_0 \sum_{i=1}^m x_i = \sum_{i=1}^m x_i y_i$$

$$\theta_1 \sum_{i=1}^m x_i + \theta_0 m = \sum_{i=1}^m y_i$$

Matricialmente :

$$\left[\begin{array}{cc} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{array} \right] \underbrace{\left[\begin{array}{c} \theta_1 \\ \theta_0 \end{array} \right]}_{\Theta} = \left[\begin{array}{c} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{array} \right] \quad (4)$$

D : matriz relacionada a nuestros datos

y : vector relacionado con nuestros datos

• La matriz D es invertible !

es simétrica con columnas linealmente independientes.

• Luego :

$$D \Theta = y \quad / D^{-1} \text{ multiplicamos por la inversa a la izquierda}$$

$$D^{-1} D \Theta = D^{-1} y \quad , \text{ Solución } \perp$$

$$I \Theta = D^{-1} y \quad \Leftrightarrow \quad \boxed{\Theta = D^{-1} y} \quad (5)$$

2) Otra forma de ver el problema
es a través del álgebra lineal:

Dado el modelo g_θ con 2 parámetros
sustituimos en g_θ los x_i medidos:

$$g_\theta(x_i) = \theta_1 x_i + \theta_0$$

y obtenemos lo siguiente:

$$g_\theta(x_1) = \theta_1 x_1 + \theta_0$$

$$g_\theta(x_2) = \theta_1 x_2 + \theta_0$$

:

:

:

$$g_\theta(x_m) = \theta_1 x_m + \theta_0$$

Sin embargo nosotros queremos que

$$g_\theta(x_i) = y_i + \varepsilon_i$$

↑

Salida
del
Modelo

↑

Salida
Ideal

↑

Error

donde ε_i es el error
cometido por el

modelo en el ejemplo
(x_i, y_i)

Es decir :

$$\begin{aligned}y_1 &= \theta_1 x_1 + \theta_0 + \epsilon_1 \\y_2 &= \theta_1 x_2 + \theta_0 + \epsilon_2 \\&\vdots && = \\y_m &= \theta_1 x_m + \theta_0 + \epsilon_m\end{aligned}$$

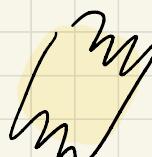
Matricialmente :

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (6)$$

O equivalentemente :

$$E = y - X\theta$$

 : al final de
estas notas
están algunas demás.

Informalmente si (forzamos)
(asumimos) $E = 0$

(donde 0 = el vector cero!) ; el problema
se reduce a :

$$y = X\theta \quad (7)$$

La matriz X es una matriz especial,
llamada matriz de Vandermonde que tiene
sus columnas siempre linealmente independientes.

Dado el siguiente teorema 2 :

Si X tiene columnas linealmente independientes
entonces $X^T X$ es invertible.

Resolvemos la ecuación (7) de la siguiente
forma :

Multiplicamos por la izquierda X^T :

$$X^T X \theta = X^T y$$

Invertimos $(X^T X)$:

$$\theta = (X^T X)^{-1} X^T y \quad (8)$$

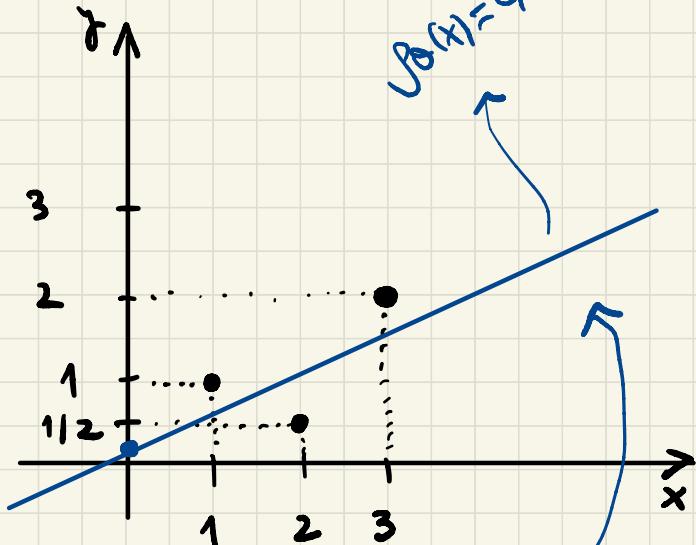
(Ecuación Normal)

- Observe que la solución al problema depende de las mediciones y el modelo $g(\cdot)$ que escogimos.

Ejemplo Regresión Lineal

i	x_i	y_i
1	1	1
2	2	$\frac{1}{2}$
3	3	2

↑ ↑ ↑
indice input output



Usando la matriz 4 :

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

$$\begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.26 \end{bmatrix}$$

∴ El modelo ideal usando 2 parámetros es:

$$f_{\theta}(x) = 0.5x + 0.26$$

Ejemplos notebook de audios (Zero)

El ejemplo con audio es más complejo de lo que parece. Es una generalización de problema (7) recordar que :

$$y = \begin{pmatrix} x \\ \theta \end{pmatrix}$$

es un vector

es una matriz

Ahora, quisieramos generalizar al caso :

$$Y = X\theta$$

↑ ↑ ↗

matrix matrix matrix

(9)

Informalmente si asumimos que X es una matriz con columnas linealmente independientes, podemos usar la ecación normal:

$$X^T X \theta = X^T Y$$

$$\theta = (X^T X)^{-1} X^T Y \quad (9)$$

donde la única diferencia es que θ e Y son Matrices!

Ahora, volvamos al audio :

3087 muestras de salida de un audio

1000 muestras
de entrada de
un audio

$$\begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,3087} \\ y_{2,1} & y_{2,2} & \dots & y_{2,3087} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_{49,1} & \dots & y_{49,3087} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,1000} \\ 1 & x_{21} & x_{22} & \dots & x_{2,1000} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & \dots & \dots & \dots & x_{49,1000} \end{bmatrix} + \theta + \epsilon$$

- Es decir, dado un conjunto de audios con una misma persona diciendo "zero" quebramos el audio en 2 partes

Nuevamente
 $\epsilon = 0$

$$[x_1, \dots, x_{1000}, x_{1001}, \dots, x_{4087}]$$

- Cada parte seguirá la entrada y salidas del modelo.
 - Un único audio es almacenado para testear el modelo (El modelo de Regresión completa el audio).
-

Precímbulo : ideas de lineal

- Una matriz inversa debe cumplir con

$$A^{-1} A = I \quad y \quad A^T A = I$$

donde I es la matriz identidad, con
 $A \in \mathbb{R}^{m \times m} \Rightarrow I \in \mathbb{R}^{m \times m}$ (cuadradas).

- Sea $X \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^{n \times 1}$, tal que :

$$Xx = 0 \quad \text{con } x \neq \vec{0}$$

Son elementos del espacio nulo de la matriz $N(X)$
 Por ejemplo :

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \quad u = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 + (-1) \\ 2 + (-2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Cuando una matriz tiene columnas dependientes (o que no aportan información) el espacio nulo,

o también llamado Kernel. A veces escriben

$$u \in N(X)$$

"u pertenece al espacio nulo de X"

- Supongamos, A es invertible y en $\mathbb{R}^{m \times m}$

Sea $u \in N(A)$ entonces $Au = 0$,

luego :

$$A^{-1}A u = 0 \Leftrightarrow I \cdot u = 0 \Leftrightarrow u = 0$$

\therefore el espacio nulo de A es $\{0\}$ solo el uno!

- Si a_1, a_2, \dots, a_m son las columnas de A entonces :

$$Ax = \begin{bmatrix} | & | & | \\ a_1 & a_2 & \dots & a_m \\ | & | & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = a_1 x_1 + \dots + a_m x_m$$

* Piensen en la multiplicación de matrices - Vector (V)ita en clases -

- Si $Ax = 0 \Leftrightarrow a_1 x_1 + \dots + a_m x_m = 0$
- Si además las columnas son Línealmente independientes $\Rightarrow x_1 = x_2 = \dots = x_m = 0.$



Equivalent a decir: No hay forma de combinar las columnas de A con op. elementales si no es solo eligiendo uno!

Prueba

Pensemos entonces, dado el premábito, en lo siguiente :

$$X^T(Xu) = 0$$

Aquí el vector u es un vector transformado por X , luego el nuevo vector entre paréntesis Xu formalmente "en la imagen" de X es a la vez un vector en el espacio nulo de X^T , puesto que la expresión se cera.

Pero X tiene columnas linealmente independientes!

Luego, si $Xu = 0 \Rightarrow u = 0$.

∴ $Xu = 0$ es decir,

el espacio nulo de $X^T X$ es solo $\{0\}$

Y en tal caso $X^T X$ es invertible. \square

Ecaciones Normales

1) Formalmente, queremos minimizar el error:

$$\epsilon = y - X\theta \quad (1)$$

Recordando que $\|X\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_m^2}$ es la norma-2 de vectores (una forma de medir un vector como si tuviéramos una regla o cinta para medir). Tomemos la norma-2 del error al cuadrado:

$$\|\epsilon\|_2^2 = \|y - X\theta\|_2^2$$

Recordando que $\|X\|_2^2 = x_1^2 + x_2^2 + \dots + x_m^2$ con $x \in \mathbb{R}^n$

$$\|X\|_2^2 = X^T X = [x_1, x_2, \dots, x_m] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Entonces :

$$\begin{aligned}\|\epsilon\|_2^2 &= (y - X\theta)^T (y - X\theta) \\ &= y^T y - y^T X \theta - \theta^T X^T y + \theta^T X^T X \theta\end{aligned}$$

Ahora queremos derivar el error en relación a los parámetros θ , asumimos del cálculo matricial las siguientes derivadas :

$$\begin{aligned}\nabla_{\theta} (\theta^T X^T y) &= X^T y \quad \nabla_{\theta} (\theta^T X^T X \theta) = 2X^T X \theta \\ \nabla_{\theta} (y^T X \theta) &= X^T y\end{aligned}$$

Aplicando el gradiente y utilizando las fórmulas:

$$\nabla_{\theta} \left(\cancel{y^T y} - y^T X \theta - \theta^T \cancel{X^T y} + \theta^T X^T X \theta \right) = 0$$

Entonces:

$$- X^T y - X^T y + 2 X^T X \theta = 0$$

Luego:

$$\cancel{2} X^T X \theta = \cancel{2} X^T y$$

$$(X^T X) \theta = X^T y \quad | (X^T X)^{-1}$$

$$\theta = (X^T X)^{-1} X^T y \quad \square$$