

Regresión Linear Múltiple

Estimar una variable dado varias variables
independiente

Regresión lineal con múltiples variables (ejemplo)

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
325	5	1	45	4,200
247	4	2	30	3,500
128	3	2	25	3,100
210	3	2	28	2,000
89	2	2	16	2,100
75	2	1	9	1,800

Modelo de Regresión Lineal Múltiple

Modelo: $y = h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

n : número de variables independientes

Parámetros del modelo : $\theta_0, \theta_1, \theta_2, \cdots, \theta_n$

Función de coste:

$$J(\theta_0, \theta_1, \theta_2, \cdots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Meta: $\min_{\theta_0, \theta_1, \theta_2, \cdots, \theta_n} J(\theta_0, \theta_1, \theta_2, \cdots, \theta_n)$

Simplificación


Considerando que el modelo

$$y = h_{\theta}(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n \quad (4)$$

Con $x_0 = 1$, tiene mayor facilidad de manejar los datos

Datos de entrada:

Todos son unos


$$\mathbf{x}^{(1)} = [x_0^{(1)}, x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}]^T$$

$$\mathbf{x}^{(2)} = [x_0^{(2)}, x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}]^T$$

$$\mathbf{x}^{(3)} = [x_0^{(3)}, x_1^{(3)}, x_2^{(3)}, \dots, x_n^{(3)}]^T$$

:

:

$$\mathbf{x}^{(m)} = [x_0^{(m)}, x_1^{(m)}, x_2^{(m)}, \dots, x_n^{(m)}]^T$$

Cada dato de entrada $\mathbf{x}^{(i)} \in \mathbb{R}^{n+1}$

Simplificación

La relación entre datos de entrada y la salida está dada:

$$\hat{y}^{(i)} = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} \quad (5)$$

Los parámetros de modelo serían

$$\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_n]^T$$

$$\text{y } \boldsymbol{\theta} \in \mathbb{R}^{n+1}.$$

La ecuación (5) se puede expresar como: $\hat{y}^{(i)} = h_{\theta}(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^T \mathbf{x}^{(i)}$

$$\hat{y}^{(i)} = h_{\theta}(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^T \mathbf{x}^{(i)} = [\theta_0, \theta_1, \dots, \theta_n] \times \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$$

Algoritmo de aprendizaje

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \dots, \theta_n), j = 0, 1, \dots, n$$

Donde

$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \dots, \theta_n)$ es la derivada parcial de la función de coste con respecto a θ_j , $j=0,1,\dots,n$

α es factor de aprendizaje que controla velocidad de aprendizaje y desajuste de error.

Derivada de la función de costo

Función de coste

$$J(\theta_0, \theta_1, \theta_2, \theta_3, \dots \theta_n) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_n]^T$$

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

(1) Derivada parcial de función de coste con respecto a θ_0

$$\begin{aligned}\frac{\partial}{\partial \theta_0} J(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_0} \frac{1}{2m} \sum_{i=1}^m (h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)})^2 = \frac{\partial}{\partial \theta_0} \frac{1}{2m} \sum_{i=1}^m (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots + \theta_n x_n^{(i)} - y^{(i)})^2 \\ &= \frac{1}{m} \sum_{i=1}^m (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots + \theta_n x_n^{(i)} - y^{(i)}) x_0^{(i)} \\ &= \frac{1}{m} \sum_{i=1}^m \underbrace{(h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)})}_{\text{Error}} \underbrace{x_0^{(i)}}_1\end{aligned}$$

(2) Derivada parcial de función de costo con respecto a θ_1

$$\begin{aligned}\frac{\partial}{\partial \theta_1} J(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_1} \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 = \frac{\partial}{\partial \theta_1} \frac{1}{2m} \sum_{i=1}^m (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots + \theta_n x_n^{(i)} - y^{(i)})^2 \\ &= \frac{1}{m} \sum_{i=1}^m (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots + \theta_n x_n^{(i)} - y^{(i)}) x_1^{(i)} \\ &= \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_1^{(i)}\end{aligned}$$

¿Cuál es la función general de derivada parcial?

¿Cómo se puede escribir $\frac{\partial}{\partial \theta_k} J(\boldsymbol{\theta})$?

$$\frac{\partial}{\partial \theta_k} J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_k^{(i)}, \quad k=0, \dots, n$$

Algoritmo de aprendizaje

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \dots, \theta_n), j = 0, 1, \dots, n$$

$$\theta_k := \theta_k - \alpha \frac{\partial}{\partial \theta_k} J(\boldsymbol{\theta})$$

$$= \theta_k - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_k^{(i)}, k = 0, \dots, n$$

Donde α es parámetro de aprendizaje (learning rate)

Normalización y Estandarización

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
325	5	1	45	4,200
247	4	2	30	3,500
128	3	2	25	3,100
210	3	2	28	2,000
89	2	2	16	2,100
75	2	1	9	1,800

El rango de variables independientes es muy diferente entre sí



- (1) No se puede determinar el valor de factor de aprendizaje adecuadamente.
- (2) Para evitar divergencia del valor de la función de coste, tenemos que usar un valor muy pequeño para el factor de aprendizaje (tasa de aprendizaje) → Se tarda mucho tiempo para converger.

El valor de factor de aprendizaje

Cuando no está normalizado o estandarizado los variables independientes, la superficie de error es un ovalo muy pronunciada.

Requiere un factor de aprendizaje muy pequeño para que el algoritmo converja.

$$\alpha < \frac{2}{\|x_{max}\|^2}$$

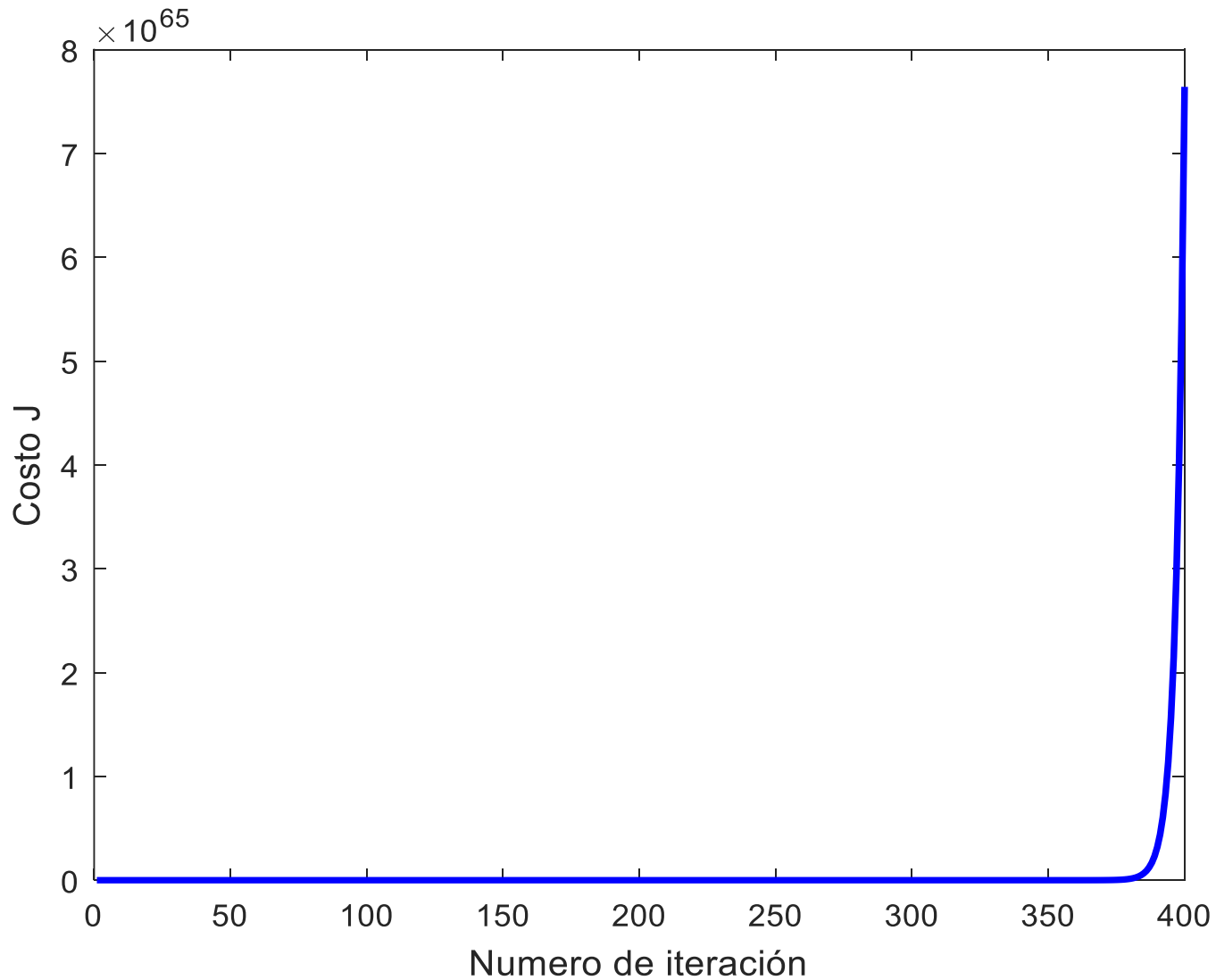
Ejercicio 1: ¿Cuál es el valor de α máximo para que el sistema converja?

X1	X2	Y
2104	3	399900
1600	3	329900
2400	3	369000
1416	2	232000
3000	4	539900
1985	4	299900
1534	3	314900
1427	3	198999
1380	3	212000
1494	3	242500
1940	4	239999
2000	3	347000
1890	3	329999
4478	5	699900
1268	3	259900

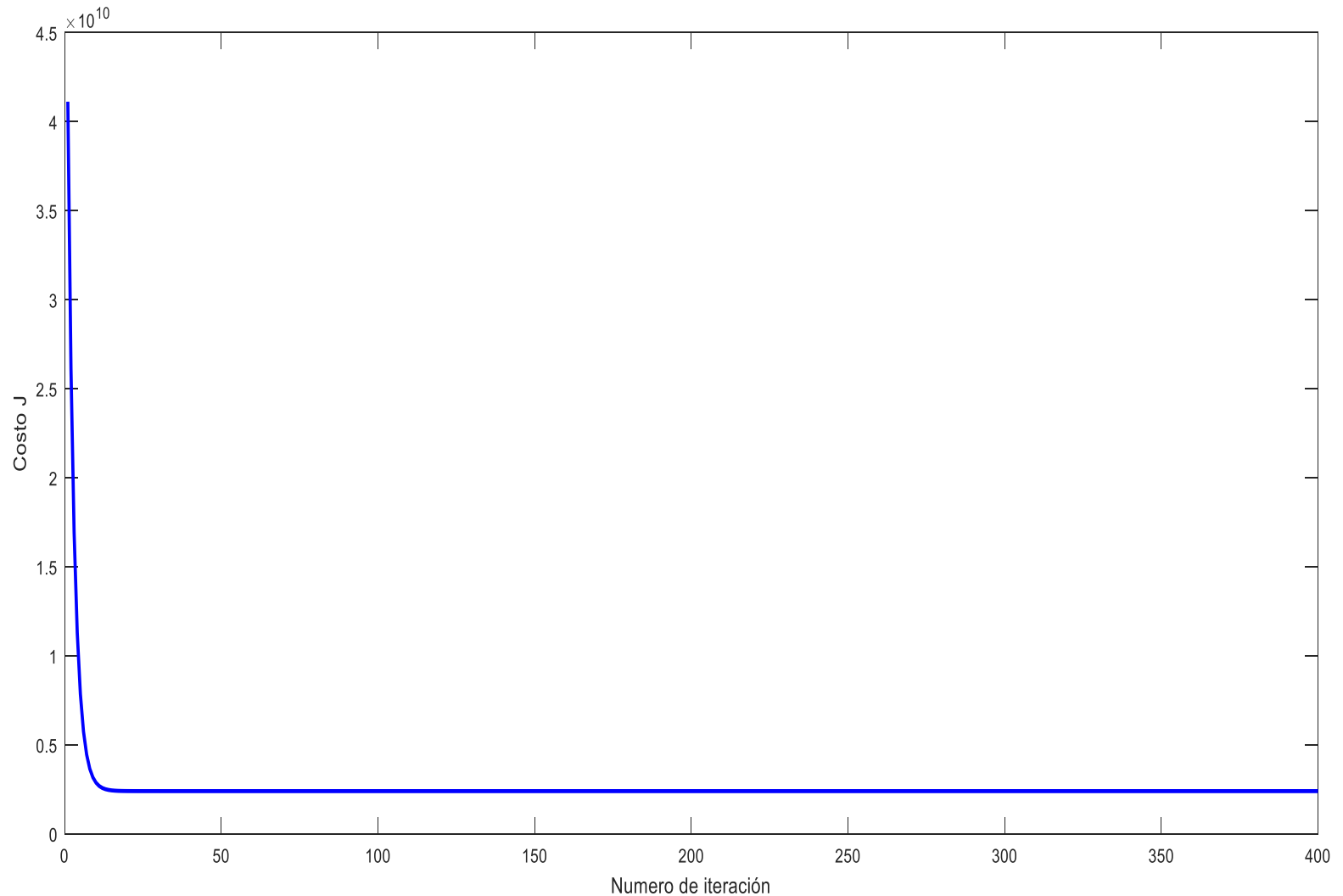
Respuesta (Ejercicio 1)

$$\alpha < \frac{2}{\|x_{max}\|^2} = \frac{2}{(4478)^2} \approx 9.9738e - 08$$

$$\alpha = 0.0000001 = 1.0 \times 10^{-7} > 9.9738e-08$$

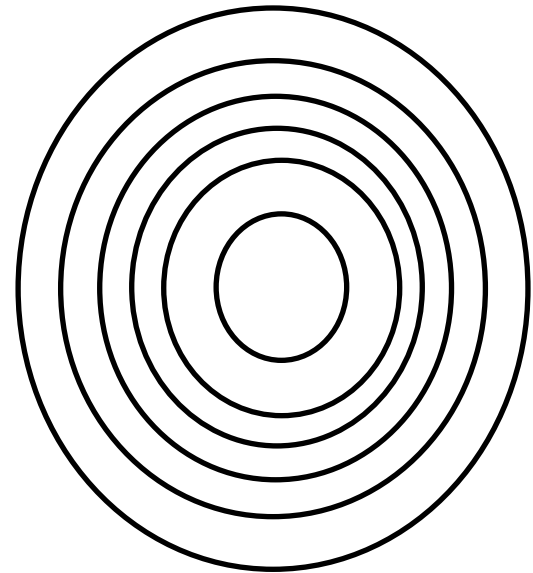
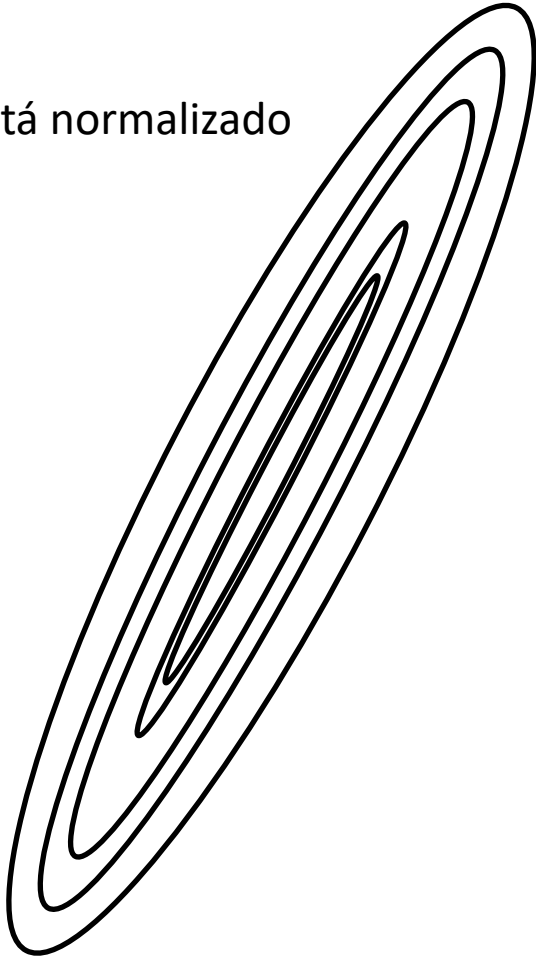


$$\alpha = 0.000000001 = 1.0 \times 10^{-9} < 9.9738e - 08$$



Superficie de error

No está normalizado



Normalizado

Normalización y Estandarización

(1) Normalización

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

(2) Estandarización

$$X_{stand} = \frac{X - \text{mean}(X)}{S(X)}$$

Donde $S(X)$: desviación estándar.

Ejemplo: Normalizar en el rubro del tamaño x_1

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
325	5	1	45	4,200
247	4	2	30	3,500
128	3	2	25	3,100
210	3	2	28	2,000
89	2	2	16	2,100
75	2	1	9	1,800

$$\mathbf{x}_{1_norm}^{(1)} = \frac{325 - 75}{325 - 75} = 1$$

$$\mathbf{x}_{1_norm}^{(2)} = \frac{247 - 75}{325 - 75} = 0.688$$

$$\mathbf{x}_{1_norm}^{(2)} = \frac{128 - 75}{325 - 75} = 0.212$$

:

Ejercicio 2: Normalizar antigüedad de la inmueble

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
1.0	5	1	45	4,200
0.688	4	2	30	3,500
0.212	3	2	25	3,100
0.54	3	2	28	2,000
0.056	2	2	16	2,100
0	2	1	9	1,800

Ejercicio 2: Respuesta

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
1.0	1.0	0	1.0	4,200
0.688	0.667	1	0.583	3,500
0.212	0.333	1	0.444	3,100
0.54	0.333	1	0.528	2,000
0.056	0	1	0.194	2,100
0	0	0	0	1,800

Cuando están normalizados los datos, $\|x_{max}\|^2=1$, por lo tanto $\alpha < \frac{2}{\|x_{max}\|^2}=2$

Ejemplo: Estandarizar en el rubro del tamaño x_1

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
325	5	1	45	4,200
247	4	2	30	3,500
128	3	2	25	3,100
210	3	2	28	2,000
89	2	2	16	2,100
75	2	1	9	1,800

$$\bar{x}_1^{(1)} = \frac{\mathbf{x}_1^{(1)} - \mu_1}{s_1} = \frac{325 - 178.67}{89.58} = 1.634$$

$$\bar{x}_1^{(2)} = \frac{\mathbf{x}_1^{(2)} - \mu_1}{s_1} = \frac{247 - 178.67}{89.58} = 0.740$$

$$\bar{x}_1^{(1)} = \frac{\mathbf{x}_1^{(1)} - \mu_1}{s_1} = \frac{128 - 178.67}{89.58} = -0.566$$

⋮

Ejercicio-3: Obtener valores estandarizados en antigüedad

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
1.634	5	1	45	4,200
0.740	4	2	30	3,500
-0.566	3	2	25	3,100
0.350	3	2	28	2,000
-1.001	2	2	16	2,100
-1.1572	2	1	9	1,800

Respuesta

Tamaño x_1 (m2)	# de habitaciones x_2	# de pisos x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
1.634	1.712	-1.414	1.772	4,200
0.740	0.781	0.707	0.397	3,500
-0.566	-0.156	0.707	-0.044	3,100
0.350	-0.156	0.707	0.221	2,000
-1.001	-1.093	0.707	-0.839	2,100
-1.1572	-1.093	-1.414	-1.457	1,800

Implementación (1) cálculo de errores

Cálculo secuencial

$$\begin{aligned}\hat{y}^{(i)} &= h_{\theta}(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^T \mathbf{x}^{(i)} \\ &= [\theta_0, \theta_1, \dots, \theta_n] \times \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}\end{aligned}$$

$$e^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

$$i=1,2,\dots,m$$

Vectorizado

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix}$$

$$\boldsymbol{\theta} = [\theta_0 \quad \theta_1 \quad \dots \quad \theta_n]^T$$

$$\mathbf{y} = [y^{(1)} \quad y^{(2)} \quad \dots \quad y^{(m-1)} \quad y^{(m)}]^T$$

$$\hat{\mathbf{y}} = \mathbf{X} \times \boldsymbol{\theta}$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

n : dimensión de cada dato

m : numero de datos

Implementación (2) –Función de coste y derivada

Cálculo secuencial

$$\begin{aligned} J(\theta_0, \theta_1, \theta_2, \theta_3, \dots \theta_n) \\ = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) &= \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \\ &= \frac{1}{m} \sum_{i=1}^m e^{(i)} x_j^{(i)} \end{aligned}$$

$$j=0, \dots, n$$

Vectorizado

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum e^2$$

$$\frac{\partial}{\partial \theta} J(\boldsymbol{\theta}) = \frac{1}{m} \sum eX$$

Implementación (3) Actualización de parámetros

Cálculo secuencial

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \dots, \theta_n)$$

$$j = 1 \dots n$$

Vectorizado

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Datos categóricos

No todas las variables independientes son numéricas. Algunas son categóricas.

R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77

Variables categóricas



Pregunta: Cuál de las variables independiente es categórica

Tamaño x_1 (m2)	# de habitaciones x_2	Código postal x_3	Antigüedad x_4 (años)	Precio y (mil pesos)
325	5	1112	45	4,200
247	4	1234	30	3,500
128	3	9876	25	3,100
210	3	3548	28	2,000
89	2	2189	16	2,100
75	2	2341	9	1,800

Variables categóricas

- Generalmente está expresada con texto
- No tiene relación “mayor”, “menor” entre ellas
“New York” < “California” --- no tiene sentido
- No se puede aplicar operaciones matemáticas, tales como suma, promedio, diferencia, etc.

Sin embargo, variables categóricas tiene información importante en análisis de datos, que no se puede ignorar.

Manejo de variables categóricas

(1) Asignar etiqueta a cada valor de variable categórica (“Label Encoding”)

ejemplo:

“New York” \rightarrow 0, “California” \rightarrow 1, “Florida” \rightarrow 2

Generar valor numérico temporalmente, para el proceso posterior

(2) Convertir cada etiqueta en un vector de 0 o 1 (“One-Hot Encoding”)

ejemplo:

“New York” \rightarrow 0 \rightarrow [1 0 0]

“California” \rightarrow 1 \rightarrow [0 1 0]

“Florida” \rightarrow 2 \rightarrow [0 0 1]

Manejo de variables categóricas

(3) Evitar colinealidad

Colinealidad: Correlación lineal entre variables independiente.

Por ejemplo:

Tres variables independientes: x_1, x_2, x_3

Regresión lineal múltiple: $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

Si la relación $x_1 = \alpha x_2 + \beta x_3 + \gamma$, existe colinealidad entre tres variables independientes.

- Se puede eliminar x_1 del modelo, ya que x_1 se puede obtener con x_2 y x_3
- x_1 no está contribuyendo \rightarrow estorbando

Resultado de One-Hot Encoding : Pregunta

R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12

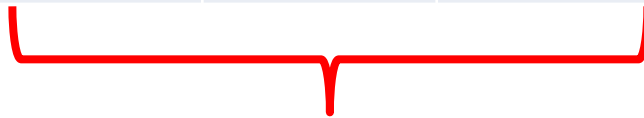


R&D Spend	Administration	Marketing Spend	OneHot1	OneHot2	OneHot3	Profit
165349.2	136897.8	471784.1	0	0	1	192261.83
162597.7	151377.59	443898.53	1	0	0	191792.06
153441.51	101145.55	407934.54	0	1	0	191050.39
144372.41	118671.85	383199.62	0	0	1	182901.99
142107.34	91391.77	366168.42	0	1	0	166187.94
131876.9	99814.71	362861.36	0	0	1	156991.12

¿ Datos tiene colinealidad?

Resultado de One-Hot Encoding : Pregunta

R&D Spend	Administration	Marketing Spend	OneHot1	OneHot2	OneHot3	Profit
165349.2	136897.8	471784.1	0	0	1	192261.83
162597.7	151377.59	443898.53	1	0	0	191792.06
153441.51	101145.55	407934.54	0	1	0	191050.39
144372.41	118671.85	383199.62	0	0	1	182901.99
142107.34	91391.77	366168.42	0	1	0	166187.94
131876.9	99814.71	362861.36	0	0	1	156991.12



Genera colinealidad

$$\text{OneHot1} = 1 - (\text{Onehot2} + \text{Onehot3})$$

$$x_1 = \alpha x_2 + \beta x_3 + \gamma \quad \alpha = \beta = -1, \gamma = 1$$

Eliminación de Colinealidad

“New York” $\rightarrow 0 \rightarrow [1 \ 0 \ 0] \rightarrow [0 \ 0]$

“California” $\rightarrow 1 \rightarrow [0 \ 1 \ 0] \rightarrow [1 \ 0]$

“Florida” $\rightarrow 2 \rightarrow [0 \ 0 \ 1] \rightarrow [0 \ 1]$

R&D Spend	Administration	Marketing Spend	OneHot1	OneHot2	OneHot3	Profit
165349.2	136897.8	471784.1	0	0	1	192261.83
162597.7	151377.59	443898.53	1	0	0	191792.06
153441.51	101145.55	407934.54	0	1	0	191050.39
144372.41	118671.85	383199.62	0	0	1	182901.99
142107.34	91391.77	366168.42	0	1	0	166187.94
131876.9	99814.71	362861.36	0	0	1	156991.12



R&D Spend	Administration	Marketing Spend	OneHot1	OneHot2	Profit
165349.2	136897.8	471784.1	0	0	192261.83
162597.7	151377.59	443898.53	1	0	191792.06
153441.51	101145.55	407934.54	0	1	191050.39
144372.41	118671.85	383199.62	0	0	182901.99
142107.34	91391.77	366168.42	0	1	166187.94
131876.9	99814.71	362861.36	0	0	156991.12

Práctica 1

(programar sin usar funciones)

- Dos variables independientes y un variable dependiente
- Operación vectorizada
- Estandarización de datos

Uso de funciones de Sklearn

- StandardScaler
- MinMaxScaler

Tarea (opcional)

Generar una función de estandarización y normalización sin usar Sklearn

Práctica 2 (Datos reales)

Precio de inmueble en ciudad de Boston.

(The Boston house-price data)

- 13 variables independientes que incluyen tasa de crimen, proporción entre estudiantes y maestro, concentración de nitrógeno, etc.
- Usando estos 13 datos, predecir precio de inmueble
- Desde los resultados (coeficientes ajustados), se puede analizar la contribución (positiva/negativa) de cada variable independiente.

Boston Housing Dataset

13 variables independientes

	Sigla	Descripción
1	CRIM	Tasa de crimen por capita
2	ZN	proporción de terreno residencial zonificado para lotes de más de 25,000 pies cuadrados.
3	INDUS	proporción de acres de negocios no minoristas por ciudad.
4	CHAS	Variable ficticia de Rio Charles (1 si el tramo limita con el río; 0 en caso contrario)
5	NOX	Concentración de óxidos nítricos (partes por 10 millones)
6	RM	Promedio de número de habitación
7	AGE	proporción de unidades ocupadas por sus propietarios construidas antes de 1940
8	DIS	Distancias ponderadas a cinco centros de empleo de Boston
9	RAD	índice de accesibilidad a carreteras radiales
10	TAX	tasa de impuesto a la propiedad sobre el valor total por cada \$10,000
11	PTRATIO	Relación alumno-profesor por ciudad
12	B	$1000(B_k - 0,63)^2$ donde B_k es la proporción de negros por ciudad
13	LSTAT	% lower status of the population

1 variable dependiente : MEDV (Precio de vivienda en miles de dólares)

Analysis de resultado

- Una vez genera el modelo lineal que estima el precio de vivienda, se puede analizar los coeficientes de hiper-plano.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
-0.1	0.04	0.02	2.69	-17,77	3.81	0.0006	-1.48	0.31	-0.012	-0.952	0.009	-0.547

Práctica 3: variable categórica

Base de datos de “abalone” (oreja marina)



Type (categorical) M, F y I (infant)

Length (mm)

Diameter (mm)

Height (mm)

Whole weight (gramos)

Shucked weight [peso de cascara] (gramos)

Viscera weight (gramos)

Shell weight [peso de concha después de muerto

Rings (entero) que representa edad (año) de abalone

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994)

"The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait",

Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

Tarea-1

R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77

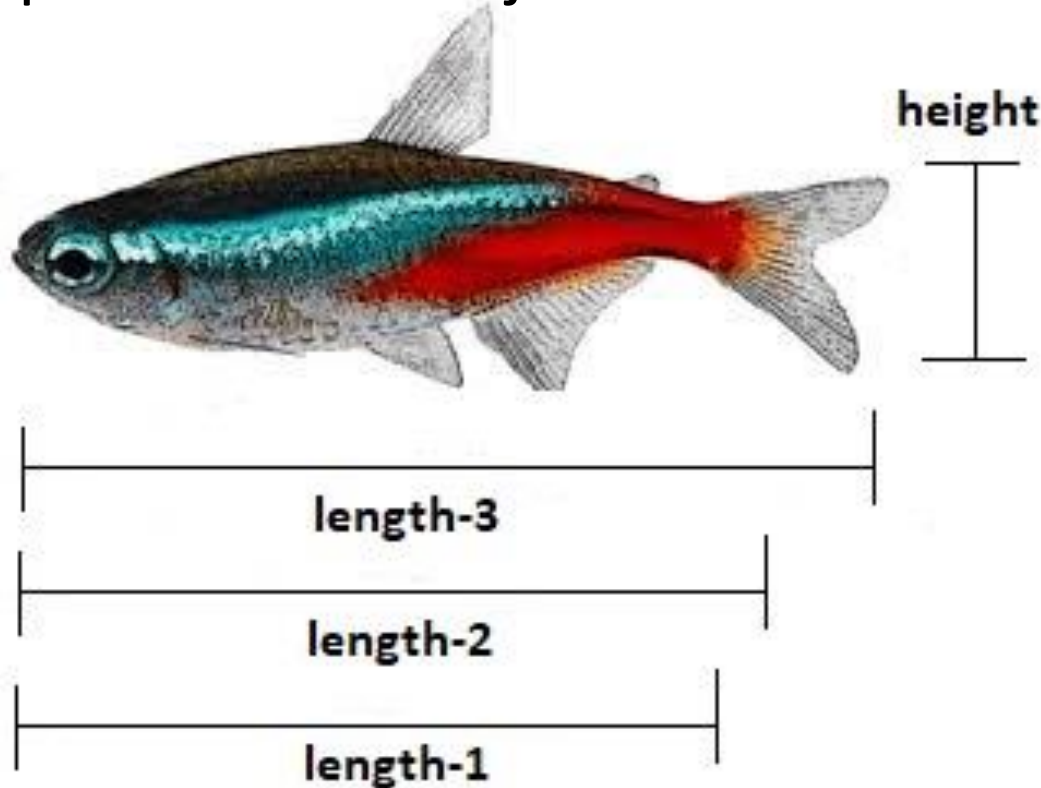
Obtener coeficientes de regresión lineal múltiple para predecir “profit”. Datos de 50 compañías que opera en tres ciudades de Estados Unidos.

Variables independientes : R&D Spend, Administration, Marketing Spend, State

Variable dependiente: Profit

Tarea-2

1. Usando regresión lineal múltiple, estimar peso de pez con los datos dados (longitudes, ancho y alto)
2. Analizar especie de pez es una variable que sirve para realizar mejor estimación o no.



7 especies de pez

1. Bream
2. Roach
3. Whitefish
4. Parkki
5. Pearch
6. Pike
7. Smelt