Modelos Estadísticos para la Regresión y la Clasificación

Práctico 5 - Regresión y clasificación

Micaela Long

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

7 de octubre de 2024

Práctico 5
Regresión y clasificación

Material para hacer práctico 5 (disponible en EVA):

- Teóricos regresión lineal (Tema 5)
- Teóricos clasificación (Tema 6)

Será de utilidad tener a mano el libro :
"An Introduction to Statistical Learning with Applications in R"

o su versión en Python:
"An Introduction to Statistical Learning with Applications in Python"

Ambos pueden ser descargados aquí: `https://www.statlearning.com/`

Objetivo: establecer una relación entre una variable dependiente Y y una variable independiente x para poder hacer predicciones sobre Y cuando se conoce x.

$$y = f(x) + \epsilon$$

Modelo regresión lineal simple:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Donde:

- y es la variable dependiente.
- x es la variable independiente.
- β_0, β_1 son parámetros desconocidos.
- ϵ es el error aleatorio.

Dado un conjunto de datos (x_i, y_i) para $i = 1, 2, \ldots, n$, el objetivo es encontrar los coeficientes β_0 y β_1 de la mejor recta:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

¿Cómo definimos mejor? Vamos a medir qué tan lejos está nuestra predicción $\hat{y}_i$ del valor verdadero y_i

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\beta_0 + \beta_1 x_i) \end{aligned}$$

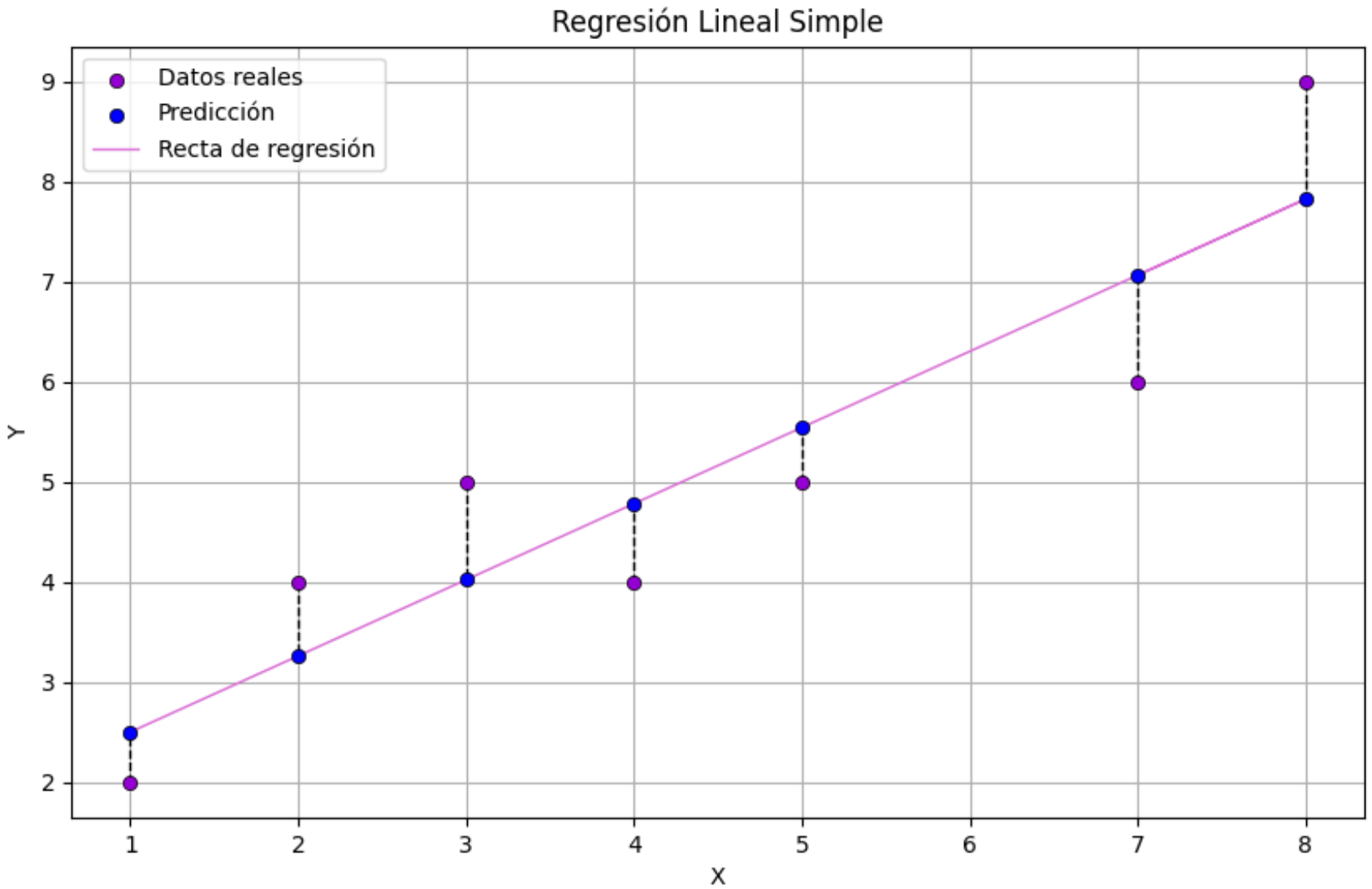
Queremos a **minimizar la suma de los cuadrados residuales (SCR)**

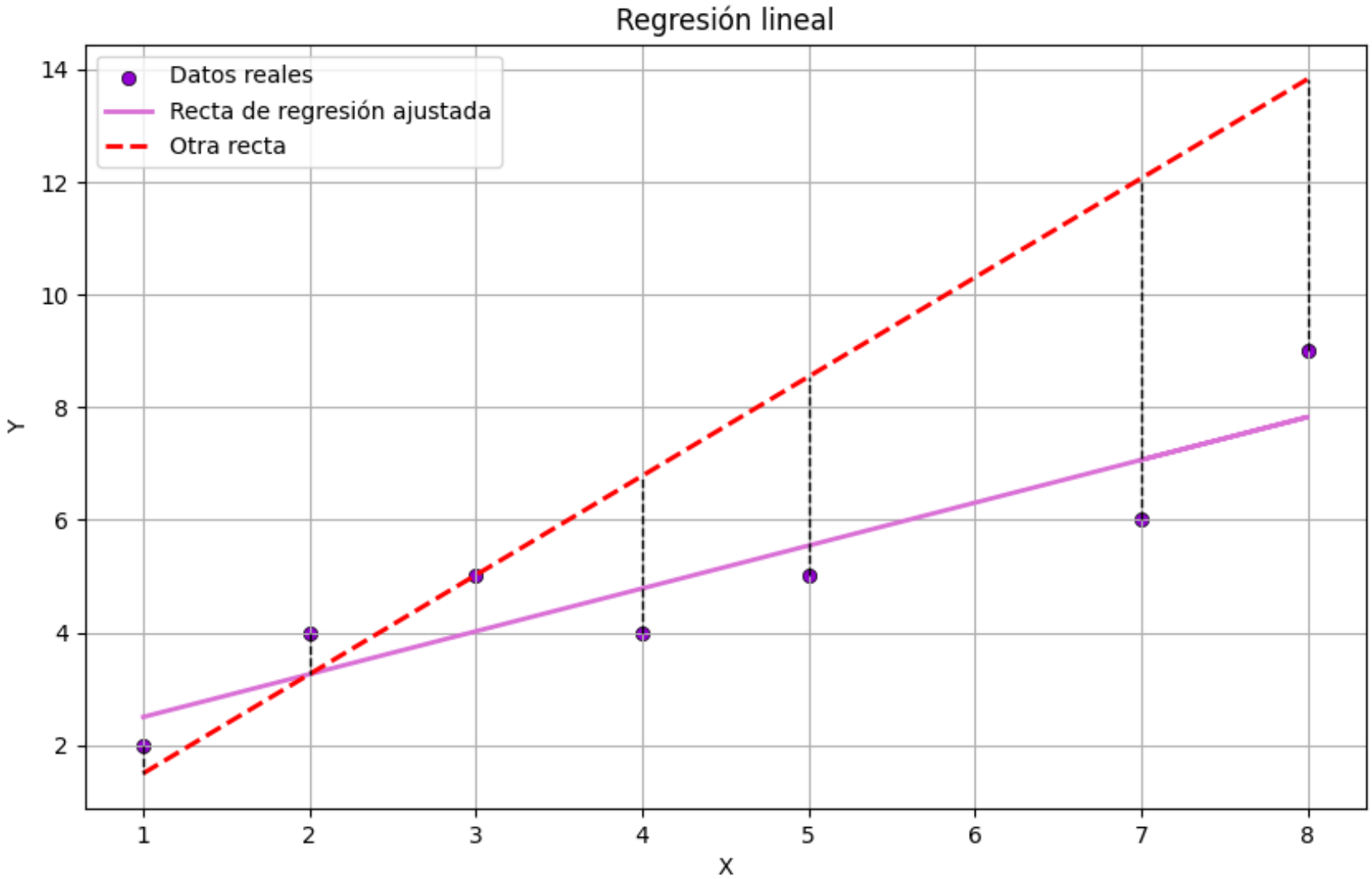
$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Aplicando el **método de mínimos cuadrados** obtenemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$


Regresión Lineal Simple
Datos reales
Predicción
Recta de regresión
Y
X
1
2
3
4
5
6
7
8
9


Regresión lineal
Datos reales
Recta de regresión ajustada
Otra recta
Y
X
14
12
10
8
6
4
2
1
2
3
4
5
6
7
8

Ejercicio 1

Justificar la expresión del intervalo de confianza para la recta de regresión.

Recordar que un intervalo de confianza al nivel $1-\alpha$ *para la respuesta media* $\mu = \beta_0 + \beta_1 x_0 = \mathbb{E}(Y|x_0)$ *es*

$$\left[\hat{y}_0 - t_{\alpha/2,n-2} \cdot \sqrt{\frac{SCR}{n-2}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_x^2}\right)}, \hat{y}_0 + t_{\alpha/2,n-2} \cdot \sqrt{\frac{SCR}{n-2}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_x^2}\right)}\right]$$

Sugerencia: Ver teórico Regresión lineal (parte 2).

Práctico 5

Ejercicio 2:
Considerar la siguiente tabla de datos sobre el rendimiento de cultivos de papas y el registro de lluvias acumuladas en el período de duración del cultivo:

Datos de entrenamiento

x = Lluvia (mm)	y = Rendimiento (ton/ha)
206	29
188	25
219	31
372	25
345	29
231	30
203	26
170	23
55	12
91	15
292	28
141	24
129	23
170	22
324	30

Datos de validación

x = Lluvia (mm)	y = Rendimiento (ton/ha)
213	30
80	16
391	25
250	26
57	9
303	28
263	28
157	25
72	13
157	23
188	26
216	25
362	28
283	33
308	30

1. Correr una regresión lineal para predecir el rendimiento y en función de la lluvia x.
2. Comparar el MSE en entrenamiento, validación y CV.
3. Determinar el grado óptimo en caso de aplicar una regresión polinomial.
4. Hallar el valor de λ óptimo para la regresión polinomial de grado 5 con regularización.

Ejercicio 2

https://colab.research.google.com/drive/1njtEKbMLdVuZWVm_nr9nGvTzTmi5qJgW#scrollTo=CWxreDBNeC2K

Ejercicio 3
El objetivo del ejercicio es implementar el algoritmo de descenso de gradiente para estimar los parámetros óptimos de la regresión logística a partir de los siguientes datos:

Paciente	Glucosa	Tiene diabetes (Si/No)
A	90	No
B	160	Si
C	100	No
D	200	Si
E	130	Si

1. Mostrar que la función de perdida de la regresión logística es

$$L(\beta) = \frac{1}{N}\sum_{i=1}^{N} \ln\left(1 + e^{-y_i \beta' x_i}\right)$$

donde $\beta = (\beta_0, \beta_1)$ es el parámetro que se busca.

2. Mostrar que el gradiente de L respecto de β está dado por

$$\frac{\partial L(\beta)}{\partial \beta_j} = \frac{1}{N}\sum_{i=1}^{N} -y_i x_{i_j} \frac{e^{-y_i \beta' x_i}}{1 + e^{-y_i \beta' x_i}}$$

3. Implementando la fórmula del descenso por gradiente

$$\beta_{k+1} = \beta_k - \alpha \nabla L(\beta_k)$$

calcular el óptimo $\hat{\beta}$ para este ejemplo.

Ejercicio 5

Considere la siguiente tabla:

x_1	x_2	x_3	y
0	3	0	Rojo
2	0	0	Rojo
0	1	3	Rojo
0	1	2	Azul
1	0	1	Azul
1	1	1	Rojo

Donde x_1, x_2, y x_3 son las características y y es la etiqueta (Rojo o Azul).

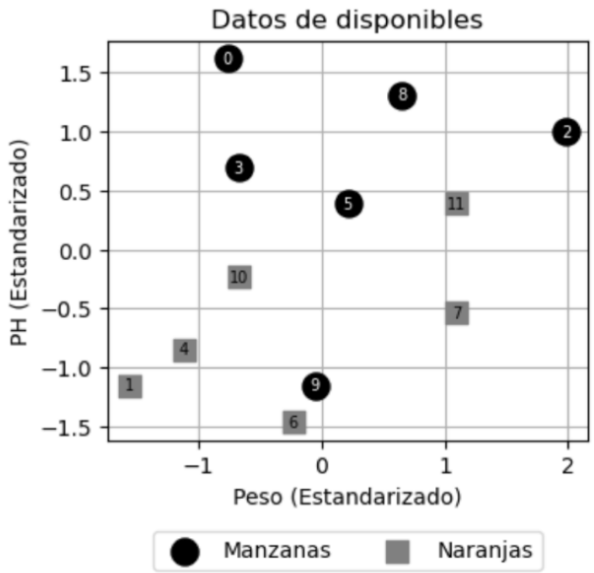
1. Con la distancia euclídea, ¿cuál es la predicción con $k = 1$ y con $k = 3$ para la observación de prueba $(0, 0, 0)$?
2. Si la frontera de decisión de Bayes en este problema es altamente no lineal, entonces, ¿esperaríamos que el mejor valor para k fuera grande o pequeño? ¿Por qué?

Ejercicio 6

Se desea implementar el algoritmo de K vecinos más cercanos para clasificar Manzanas y Naranjas en base a su Peso y su PH.

El gráfico a continuación (derecha) muestra los datos disponibles estandarizados. En el lado izquierdo, se presenta el conjunto de datos segmentado en tres folds. Cada fila simboliza una observación, y su ID correspondiente está alineado con el gráfico situado a la derecha. Las columnas, por otro lado, exhiben los IDs de los 8 puntos que son parte de los dos folds que no contienen la observación de la fila en cuestión, todos ellos ordenados por su proximidad a dicha observación. Por ejemplo, en la primera fila se visualizan las observaciones de los folds 2 y 3. Están dispuestas en un orden que va desde la más cercana a la más lejana respecto a la observación con ID=0.

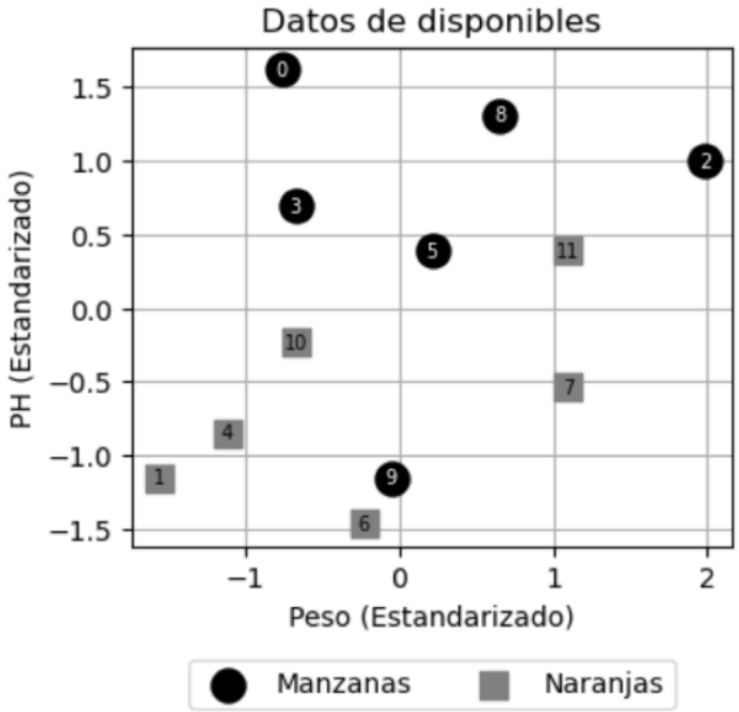
	ID	1ero	2do	3ro	4to	5to	6to	7mo	8vo
Fold 1	**0**	8	5	10	11	4	7	9	6
	1	4	10	6	9	5	7	11	8
	2	11	8	7	5	10	9	6	4
	3	10	5	8	4	11	9	7	6
Fold 2	**4**	1	10	9	3	0	11	8	2
	5	11	3	8	10	9	0	2	1
	6	9	10	1	3	11	8	0	2
	7	11	9	2	10	8	3	1	0
Fold 3	**8**	5	2	0	3	7	4	6	1
	9	6	4	7	1	5	3	0	2
	10	4	3	5	1	6	7	0	2
	11	5	7	2	3	0	6	4	1



Calcular el error de 3-fold cross-validation para los valores impares de K. ¿Cuál de estos valores de K elegiría?

Ejercicio 6

	ID	1ero	2do	3ro	4to	5to	6to	7mo	8vo
Fold 1	**0**	8	5	10	11	4	7	9	6
	1	4	10	6	9	5	7	11	8
	2	11	8	7	5	10	9	6	4
	3	10	5	8	4	11	9	7	6
Fold 2	**4**	1	10	9	3	0	11	8	2
	5	11	3	8	10	9	0	2	1
	6	9	10	1	3	11	8	0	2
Fold 3	**7**	11	9	2	10	8	3	1	0
	8	5	2	0	3	7	4	6	1
	9	6	4	7	1	5	3	0	2
	10	4	3	5	1	6	7	0	2
	11	5	7	2	3	0	6	4	1



Hacerlo para $K = 1, 3, 5, 7$ vecinos