Machine Learning para Inteligencia Artificial

Modelos Lineales: Regresión Lineal

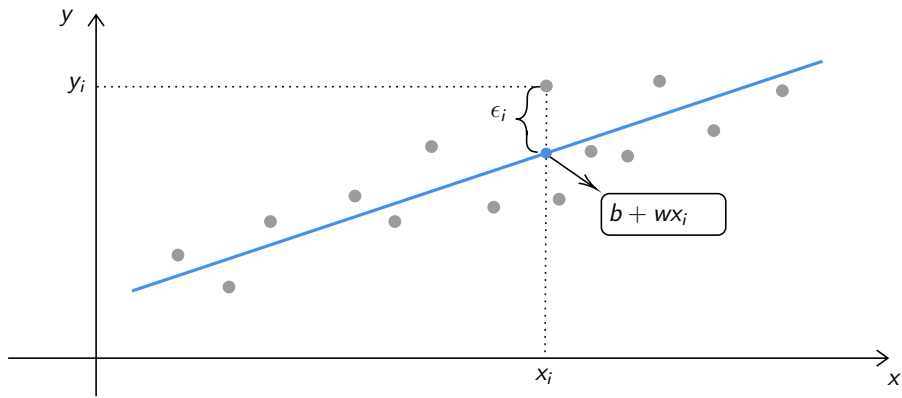
Universidad ORT Uruguay

28 de Mayo, 2025

Modelo lineal univariado

Consiste en aproximar la relación entre x e y mediante una recta

$$\widehat{y} = h(x;\boldsymbol{\theta}) = b + wx, \qquad \boldsymbol{\theta} = \begin{bmatrix} b \\ w \end{bmatrix} \in \mathbb{R}^2$$



Modelo lineal univariado

- El sesgo inductivo es entonces

$$\mathcal{H} = \{h_{\boldsymbol{\theta}} : x \mapsto b + wx\}$$

- El riesgo empírico se llama MSE (Mean Squared Error):

$$J_T(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - h(x_i;\boldsymbol{\theta}))^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - b - wx_i)^2$$

- Lo usamos para elegir los mejores parámetros:

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{b} \\ \widehat{w} \end{bmatrix} = \underset{\boldsymbol{\theta}=[b,w]^\top}{\arg\min} \Big\{ J_T(b,w) \Big\}$$

Cálculo de los coeficientes

- **Observación general**: dados números reales $A_1, \ldots, A_N$

$$\arg\min_a \left\{ \sum_{i=1}^{N} (A_i - a)^2 \right\} = \bar{A}$$

en donde $\bar{A}$ es el promedio $\frac{1}{N} \sum_{i=1}^{N} A_i$.

- Por la observación, si fijamos w, el mejor b es

$$\arg\min_b \left\{ \frac{1}{N} \sum_{i=1}^{N} (\underbrace{y_i - w x_i}_{A_i} - b)^2 \right\} = \bar{y} - w\bar{x}$$

Cálculo de los coeficientes

Luego basta encontrar $\widehat{w}$:

$$J_T(w) := J_T(\bar{y} - w\bar{x}, w) = \frac{1}{N}\sum_{i=1}^{N}(y_i - wx_i - \bar{y} + w\bar{x})^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}((y_i - \bar{y}) - w(x_i - \bar{x}))^2$$

$$= \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}_{\mathrm{var}(y)} - 2w\underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}_{\mathrm{cov}(x,y)} + w^2\underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}_{\mathrm{var}(x)}$$

$$= \mathrm{var}(y) - 2\,\mathrm{cov}(x,y)w + \mathrm{var}(x)w^2 \quad \text{(polinomio de grado 2 en } w\text{)}$$

Cálculo de los coeficientes

- Si derivamos e igualamos a cero vemos que el mínimo se alcanza en

$$\widehat{w} = \frac{\text{cov}(x, y)}{\text{var}(x)} = r\frac{\text{var}(y)^{1/2}}{\text{var}(x)^{1/2}}$$

en donde $r := (\text{coeficiente de correlación}) = \frac{\text{cov}(x,y)}{\text{var}(x)^{1/2}\,\text{var}(y)^{1/2}} \in [-1, 1]$

- Sean $\widehat{\epsilon}_i = y_i - (\widehat{b} + \widehat{w}x_i)$ los residuos, entonces obtenemos:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\epsilon}_i^2 = \text{var}(y)(1 - r^2)$$

En estadística se usa r^2 como medida de ajuste en lugar de la MSE.

Modelo lineal univariado en notación matricial

La función lineal puede escribirse como un producto de matrices

$$b + wx_i = [1, x_i] \begin{bmatrix} b \\ w \end{bmatrix} = \boldsymbol{x}_i^\top \boldsymbol{\theta} = [b, w] \begin{bmatrix} 1 \\ x_i \end{bmatrix} = \boldsymbol{\theta}^\top \boldsymbol{x}_i$$

Podemos juntar todos los valores de $\{b + wx_i\}_{i=1}^N$ en un solo producto matricial:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} b \\ w \end{bmatrix}}_{\boldsymbol{\theta}} = \underbrace{\begin{bmatrix} b + wx_1 \\ b + wx_2 \\ \vdots \\ b + wx_N \end{bmatrix}}_{\boldsymbol{X\theta}}$$

Modelo lineal univariado en notación matricial

- Con el vector de targets o etiquetas dado por $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$
- El riesgo empírico (MSE) queda

$$J_T(\boldsymbol{\theta}) = \frac{1}{N} \underbrace{\sum_{i=1}^{N} (y_i - b - wx_i)^2}_{(\text{Norma-2})^2} = \frac{1}{N} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2$$

- Y el vector de parámetros óptimo se puede calcular como

$$\boxed{\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}}$$

Modelo lineal multivariado

- El setting para el modelo multivariado es análogo salvo que

  Espacio de atributos: $\mathcal{X} \subset \mathbb{R}^D$

  pues disponemos de D atributos.

- Ahora el modelo es

$$\widehat{y} = h(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}$$

en donde los vectores $\boldsymbol{\theta}$ y $\boldsymbol{x}$ están dados por

$$\boldsymbol{\theta} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_D \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(D)} \end{bmatrix}$$

Modelo lineal multivariado

- El producto matricial representa

$$\widehat{y} = \begin{bmatrix} b & w_1 & \dots & w_D \end{bmatrix} \begin{bmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(D)} \end{bmatrix} = b + \sum_{j=1}^{D} w_j x^{(j)}$$

- Por lo que el sesgo inductivo es entonces

$$\mathcal{H} = \left\{ h_{\boldsymbol{\theta}} : \left[x^{(1)}, \dots, x^{(D)} \right] \mapsto b + \sum_{j=1}^{D} w_j x^{(j)} \right\}$$

Modelo lineal multivariado

- Considerando la matriz de diseño y el vector de etiquetas (o targets):

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(D)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(D)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N^{(1)} & \cdots & x_N^{(D)} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

- La MSE y el vector de parámetros óptimo quedan

$$J_T(\boldsymbol{\theta}) = \frac{1}{N} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \qquad \boxed{\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}}$$

- Los cálculos son idénticos y la solución es la misma.

Regresión lineal con funciones base

- Permite modelar relaciones no lineales $x \rightsquigarrow y$
- El sesgo inductivo es

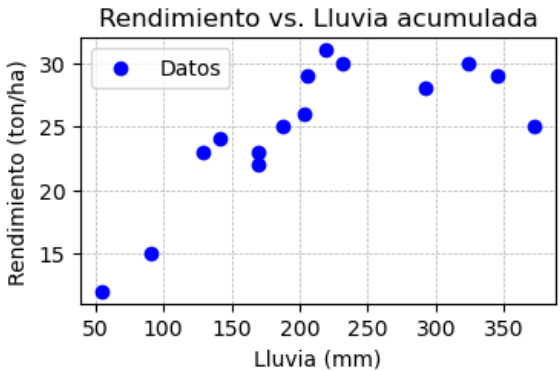
$$\mathcal{H} = \{h_{\boldsymbol{\theta}} : x \mapsto b + w_1 h_1(x) + w_2 h_2(x) + \cdots + w_D h_D(x)\}$$

- Casos particulares son:
  - Regresión polinomial: cuando $h_j(x) = x^j$
  - Regresión trigonométrica: cuando $h_j(x) = \cos(jx)$ o $h_j(x) = \sin(jx)$
- La relación entre y y los coeficientes $\boldsymbol{\theta} = (b, \boldsymbol{w})$ sigue siendo lineal
- Mismo procedimiento al caso multivariado considerando la matriz de diseño:

$$\boldsymbol{X} = \begin{bmatrix} 1 & h_1(x_1) & \cdots & h_D(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & h_1(x_N) & \cdots & h_D(x_N) \end{bmatrix}$$

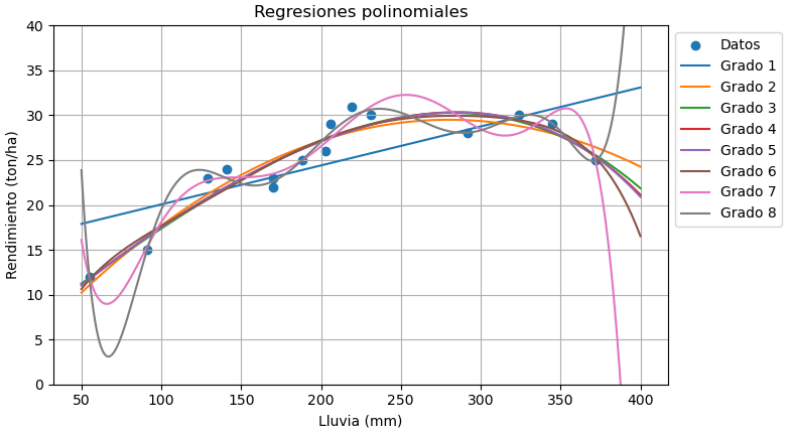
Ejemplo: regresión polinomial

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



Ejemplo: regresión polinomial

El siguiente gráfico muestra varios polinomios con grados que van desde 1 a 8:



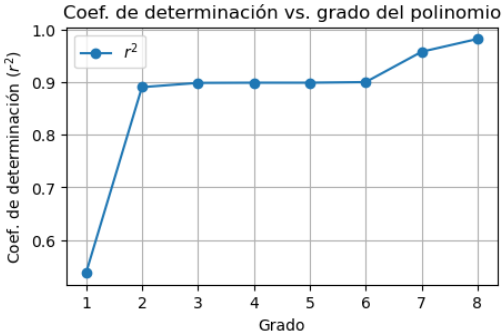
Ejemplo: regresión polinomial

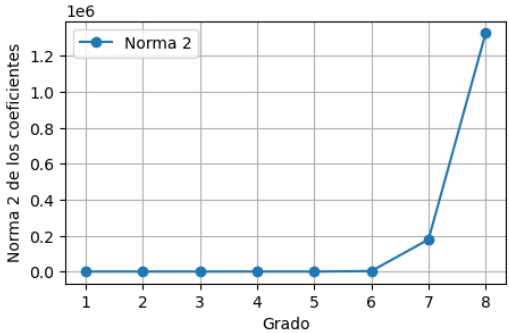
Coeficientes para los distintos grados tienden a crecer rápidamente:

Coeficiente	Grado 1	Grado 2	Grado 3	Grado 4	Grado 5	Grado 6	Grado 7	Grado 8
w_1	3.85	18.08	10.79	15.02	12.71	76.59	-1533.75	-5491.59
w_2		-14.56	3.56	-13.02	-0.23	-453.11	13365.95	52990.47
w_3			-11.13	10.75	-16.30	1318.39	-50622.38	-231805.88
w_4				-9.49	15.63	-1950.95	102868.06	571182.95
w_5					-8.56	1419.35	-116379.46	-841240.02
w_6						-406.93	69088.87	735492.68
w_7							-16787.07	-352547.47
w_8								71414.21

Ejemplo: regresión polinomial

Mejor ajuste en entrenamiento a costo de coeficientes gigantes





Regularización Ridge

Idea de regularización

Modificación a un algoritmo de aprendizaje que pretende reducir el error verdadero (en $\mathcal{D}$) pero no su error empírico (en T).

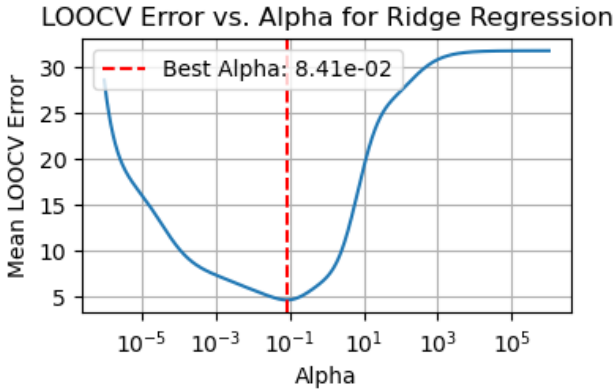
- Ridge consiste en penalizar con la norma-2 los coeficientes:

$$J_\alpha(\boldsymbol{w}) = J(\boldsymbol{w}) + \alpha\|\boldsymbol{w}\|^2 = J(\boldsymbol{w}) + \alpha\sum_{j=1}^{D} w_j^2$$

- En este caso, el objetivo es evitar los coeficientes gigantes.
- Notar que b no interviene en la penalización.

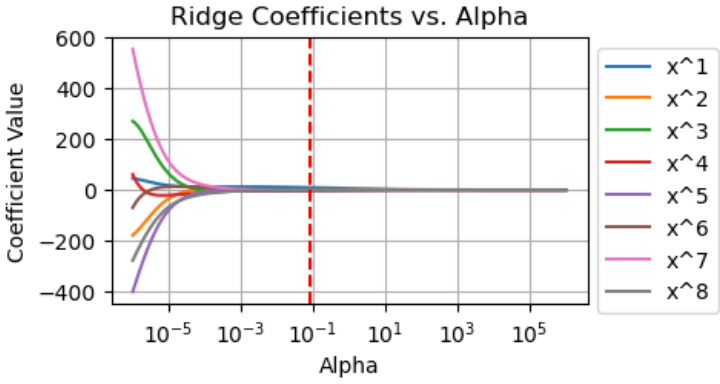
Ejemplo: regresión polinomial

Curva del riesgo (LOOCV) para varios valores de α:



Ejemplo: regresión polinomial

Evolución de los coeficientes en función de α:



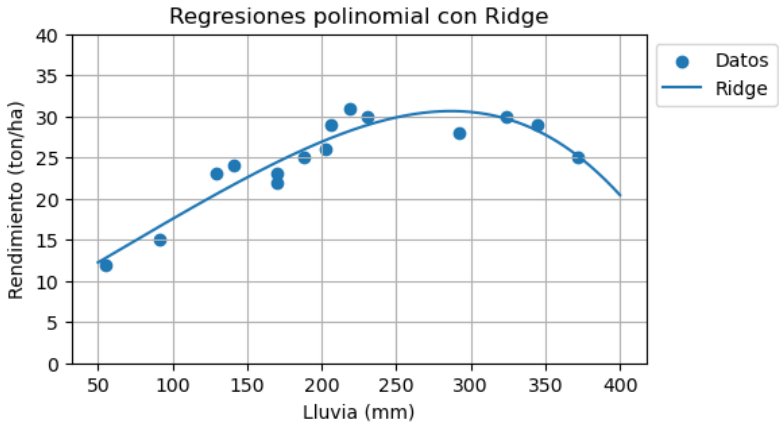
Ejemplo: regresión polinomial

Efecto de la regularización en los coeficientes:

Coef	Original	Ridge
x^1	-5491.591405	9.290364281
x^2	52990.4687	1.239116016
x^3	-231805.8798	-2.013422183
x^4	571182.9466	-2.683297203
x^5	-841240.0155	-2.179430752
x^6	735492.6833	-1.218635966
x^7	-352547.4682	-0.16433541
x^8	71414.20573	0.79483848

Ejemplo: regresión polinomial

Gráfico de la regresión regularizada:



Regularización Lasso

- Lasso (Least Absolute Shrinkage and Selection Operator)
- Consiste en penalizar con la norma-1 los coeficientes:

$$J_\alpha(\boldsymbol{w}) = J(\boldsymbol{w}) + \alpha\|\boldsymbol{w}\|_1 = J(\boldsymbol{w}) + \alpha\sum_{j=1}^{D}|w_j|$$

- También el objetivo es evitar los coeficientes gigantes.
- Tiene el efecto de seleccionar atributos

Ejemplo: regresión bivariada

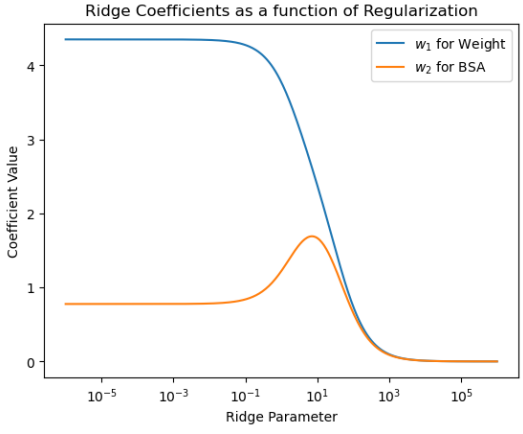
Datos sobre 20 personas con presión arterial alta:

- presión arterial (y = BP, en mm Hg)
- edad (x_1 = Age, en años)
- peso (x_2 = Weight, en kg)
- superficie corporal (x_3 = BSA, en m2)
- duración de la hipertensión (x_4 = Dur, en años)
- pulso basal (x_5 = Pulse, en latidos por minuto)
- índice de estrés (x_6 = Stress)

Queremos determinar la relación entre **BP** y (**Weight**,**BSA**) ($r = 0.875$).

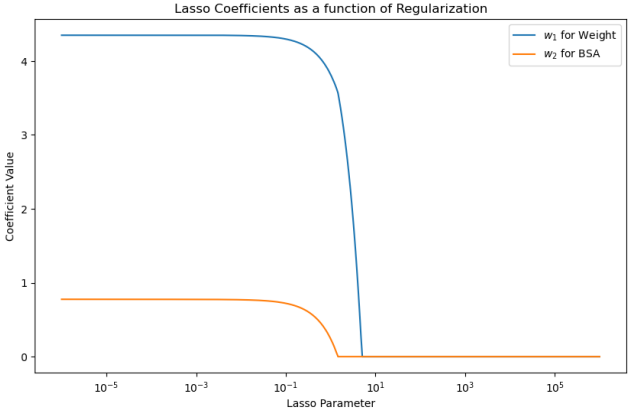
Ejemplo: regresión bivariada

Regularización Ridge



Ejemplo: regresión bivariada

Regularización Lasso



Bibliografía

- An introduction to statistical learning with applications in Python. Cap 3 y 6.2.
- Machine Learning - A First Course for Engineers and Scientists. Capítulo 5.3.