

Primer Parcial

1) Sea el modelo de Regresión $t_n = \phi(x_n)w^T + \eta_n$, con $\{t_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N$, $w \in \mathbb{R}^Q$, $\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$, $Q \geq P$, y $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$.

Donde:

- t_n : Variable objetivo (real)
- x_n : Vector de características (en \mathbb{R}^P)
- ϕ : Función de mapeo a un espacio de dimensión Q (con $Q \geq P$)
- w : Parámetros del modelo (en \mathbb{R}^Q)
- η_n : Ruido Gaussiano con distribución $N(0, \sigma_n^2)$

Desarrollo de cada modelo

↳ media
Cero

Desarrollo de cada modelo

1) Mínimos cuadrados (Ordinary Least Squares, OLS)

Se asume mapeo $\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$, con $Q \geq P$

Dado el conjunto de datos $\{\phi(x_n) \in \mathbb{R}^Q, t_n \in \mathbb{R}\}_{n=1}^N$, se puede definir el modelo lineal:

$$t_n = \phi(x_n)w^T + \eta_n \quad w \in \mathbb{R}^Q$$

NOTA: El modelo sería lineal en x_n si tuviera la forma $t_n = \alpha x_n + b$.
El modelo es lineal en w porque t_n depende linealmente de los parámetros w_j , sin importar cómo sea $\phi(x_n)$.

Cualquier valor de w mantiene la linealidad, porque la estructura del modelo es

la linealidad se refiere a que no hay términos como w_j^2 , $\sin(w_j)$ o productos $w_j w_k$ en el modelo.

→ Por tanto, el modelo es lineal en los parámetros w , ya que

$$t_n = \sum_{j=1}^Q w_j \phi_j(x_n) + \eta_n \quad \text{Esto permite usar técnicas de álgebra lineal para optimizar } w.$$

Objetivo: Encontrar w que minimice el error cuadrático medio (MSE)

$J(w) = \text{la función de costo} = \text{Error cuadrático Medio (MSE)}$

$$J(w) = \sum_{n=1}^N (t_n - \phi(x_n)w^T)^2 = \|t - \Phi w\|^2$$

Forma matricial (Agrupando todos los datos)

Donde: $t = [t_1, \dots, t_N]^T \in \mathbb{R}^N$, $\Phi = [\phi(x_1), \dots, \phi(x_N)] \in \mathbb{R}^{N \times Q}$

$\phi(x_n) \rightarrow$ Transformación
 $t \rightarrow t_n$ (Vector de targets)

• Se expande la función de pérdida

$$J(w) = (t - \Phi w)^T (t - \Phi w) \quad \rightarrow \text{La Transpuesta sale porque se debe de cambiar las filas } x \text{ columnas y las columnas } x \text{ filas.}$$

$$J(w) = t^T t - \underbrace{w^T \Phi^T t}_\text{①} - \underbrace{(w^T \Phi^T t)}_\text{②} + (\Phi w)^T (\Phi w)$$

Por tanto, reescribe la Mult ② de la misma forma que ①; esto pasa porque no se puede realizar la ②, como se ve abajo.

Donde:

$t^T t \rightarrow$ Constante (no depende de w)

$-2w^T \Phi^T t \rightarrow$ lineal en w

$w^T \Phi^T \Phi w \rightarrow$ cuadrática en w

Por propiedad $a \in \mathbb{R} \rightarrow a^T = a$

① $t^T \Phi w$
 $\begin{cases} 1 \times 1 & \text{1x1} \\ P \times P & P \times 1 \\ 1 \times P & \end{cases}$
 \downarrow
 se puede realizar

② $(\Phi w)^T t = \Phi^T w^T t$
 $\begin{cases} P \times N & P \times 1 \\ P \times P & 1 \times 1 \end{cases}$
 \downarrow
 $\text{No se puede realizar}$

- Minimización: Derivada e igualación a cero

$$n = \frac{d}{dw} (t^T t) - \frac{d}{dw} (2t^T \Phi w) + \frac{d}{dw} ((\Phi w)^T (\Phi w)) = -2 \frac{d}{dw} (t^T \Phi w) + \frac{d}{dw} (w^T \Phi^T \Phi w)$$

$\underbrace{-2t^T \Phi}_{L \times N} + \underbrace{2\Phi^T \Phi w}_{N \times P}$

$\frac{d}{dw} w^T \Phi^T \Phi w = 2 \Phi^T \Phi w$
cuadrático en w

• $\frac{d}{dw} w^T A w = 2 Aw$
Matriz simétrica

Aqui $A = \Phi^T \Phi \rightarrow$ simétrica
porque $(\Phi^T \Phi)^T = \Phi^T \Phi$

NOTA: Para que se pueda realizar la suma entre matrices, tiene que ser del mismo tamaño; por tanto, se transpone $(2t^T \Phi)^T = 2\Phi^T t$

Ahora, para encontrar el mínimo, se iguala a cero. Aquí $a = \Phi^T t$
y se despeja w para encontrar w^*

$$-2\Phi^T t + 2\Phi^T \Phi w = 0$$

$$\frac{1}{2} \cancel{\Phi^T \Phi} w = \frac{1}{2} \cancel{\Phi^T t} \rightarrow \text{dividiendo}$$

entre 2

• mult. ambos lados por $(\Phi^T \Phi)^{-1}$ para despejar w

$$(\Phi^T \Phi)^{-1} (\Phi^T \Phi) w = (\Phi^T \Phi)^{-1} \Phi^T t$$

$$w^* = (\Phi^T \Phi)^{-1} \Phi^T t$$

NOTAS: La inversa existe si $(\Phi^T \Phi)^{-1}$: debe ser invertible

• $\Phi^T \Phi$ es cuadrada, lo es porque Φ es $N \times Q$

y $\Phi^T \Phi$ es $Q \times Q$

• $\Phi^T \Phi$ tiene rango completo (i.e., sus columnas son linealmente independientes)

Si $\Phi^T \Phi$ NO es invertible (Por ejemplo, si $Q > N$)

Y en términos de Pseudoinversa, donde esta es $A^+ = (A^T A)^{-1} A^T$, entonces:

$$w^* = \Phi^+ t$$

• Se usa la pseudo-inversa (Moore-Penrose), denotada como $(\Phi^T \Phi)^+$
solución: $w = (\Phi^T \Phi)^+ \Phi^T t$

2) Mínimos Cuadrados regularizados

El estimador generalizado de mínimos cuadrados con regularización L2 También conocido como modelo lineal rígido-linear ridge regression, se prede plantear el modelo de optimización como:

$$W_{MC2} = \min_w \left[\sum_{n=1}^N (t_n - \Phi(x_n) w)^2 + \lambda \|w\|^2 \right]$$

Su forma matricial es:

$$W_{MC2} = \underset{w}{\operatorname{argmin}} \|t - \Phi w^T\|_2^2 + \lambda \|w\|_2^2$$

$\operatorname{argmin} \rightarrow$ Argumento del mínimo
Conjunto de valores que minimizan función objetivo.

Donde: $t = [t_1, t_2, \dots, t_n]^T \in \mathbb{R}^{N \times 1}$

$\Phi = [\Phi(x_1), \Phi(x_2), \dots, \Phi(x_n)]^T \in \mathbb{R}^{N \times Q}$

$$\lambda \in \mathbb{R}^+$$

$$w \in \mathbb{R}^{1 \times Q}$$

$$w^T \in \mathbb{R}^{Q \times 1}$$

- La función de costo es:

$$\begin{aligned} \|t - \Phi w^T\|_2^2 + \lambda \|w\|_2^2 &= (t - \Phi w^T)^T (t - \Phi w^T) + \lambda (w^T)^T (w) \\ &= t^T t - 2t^T \Phi w^T + (\Phi w^T)^T (\Phi w^T) + \lambda w^T w \end{aligned}$$

• Derivando respecto a W , teniendo en cuenta $(ab)^T = b^T a^T$

$$\frac{d}{dw} (t^T t) = 2 \frac{d}{dw} (t^T \Phi w) + \frac{d}{dw} (\Phi w)^T (\Phi w) + \frac{d}{dw} (\lambda w^T w)$$

$$= -2t^T \Phi + \frac{d}{dw} (w^T \Phi^T \Phi w) + 2\lambda w$$

$$= -2t^T \Phi + 2\Phi^T \Phi w + 2\lambda w$$

se iguala a cero y se divide por 2 a ambos lados de la igualdad

$$t^T \Phi = \Phi^T \Phi w + \lambda w \quad \text{factorizando}$$

$$t^T \Phi = (\Phi^T \Phi + \lambda) w$$

$$w = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t$$

3) Máxima Verosimilitud

Para el caso del Ruido Blanco Gaussiano, se tiene que:

$$r_n \sim p(r_n) = G(r_n | 0, \sigma_n^2) \quad \text{con} \quad t_n = \phi(x_n) W^T + r_n, \\ r_n = t_n - \phi(x_n) W^T$$

Por lo tanto

$$p(t_n | \phi(x_n) W^T, \sigma_n^2) = G(t_n | \phi(x_n) W^T, \sigma_n^2)$$

se pueden encontrar los pesos maximizando el log-verosimilitud

$$W_{ML} = \underset{W}{\operatorname{argmax}} \log \left(\prod_{n=1}^N G(t_n | \phi(x_n) W^T, \sigma_n^2) \right)$$

Por lo tanto:

$$\begin{aligned} \log(p(x)) &= \log \left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left(\frac{-|t_n - \phi|^2}{2\sigma_n^2} \right) \right) \\ &= \log \left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma_n^2}} \right) + \log \left(\prod_{n=1}^N \exp \left(\frac{-|t_n - \phi|^2}{2\sigma_n^2} \right) \right) \\ &= \log \left(\frac{1}{(2\pi\sigma_n^2)^{N/2}} \right) + \log \left(\exp \left(-\sum_n \frac{|t_n - \phi|^2}{2\sigma_n^2} \right) \right) \\ &= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2} \sum_{n=1}^N |t_n - \phi|^2 \end{aligned}$$

$$= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{N}{2} \log(\sigma_n^2) - \frac{1}{2\sigma_n^2} \sum_{n=1}^N |t_n - \phi|^2$$

Como la varianza del ruido (σ_n^2) es constante respecto a W , esto equivale a:

$$W_{MV} = \underset{w}{\operatorname{argmax}} \left(-\frac{1}{2\sigma_n^2} \sum_{n=1}^N (t_n - \phi(x_n) w^\top)^2 \right)$$

Finalmente se convierte el problema de optimización a minimización

$$W_{MV} = \underset{w}{\operatorname{argmin}} \left[-\frac{1}{2} \left(-\frac{1}{2\sigma_n^2} \sum_{n=1}^N (t_n - \phi(x_n) w^\top)^2 \right) \right]$$

$$W_{MV} = \underset{w}{\operatorname{argmin}} \left(\frac{1}{2\sigma_n^2} \sum_{n=1}^N (t_n - \phi(x_n) w^\top)^2 \right)$$

Maximizar $\log p(t_n | \phi(x_n) w^\top, \sigma_n^2)$ equivale a minimizar el error cuadrático

$$W_{MV} = (\Phi^\top \Phi)^{-1} \Phi^\top t$$

4) Máximo a-posteriori

Es un modelo bayesiano porque incorpora la incertidumbre en los parámetros de modelo, como los pesos, mediante el uso de distribuciones de probabilidad.

El enfoque bayesiano combina la información previa (a priori) de los pesos con la verosimilitud de los datos mediante el teorema de Bayes.

Se asume un prior Gaussiano sobre w :

$$p(w) = N(w | 0, \sigma_w^2 I_Q)$$

El modelo por máximo a-posteriori, simplifica la relación de Bayes mediante la proporcionalidad:

Es decir, la distribución posterior es proporcional al producto de la verosimilitud y el prior:

$$P(w|t) \propto p(t|w) P(w)$$

Por consiguiente, asumiendo datos independientes e identicamente distribuidos; el máximo a-posteriori dado un conjunto de datos busca encontrar el vector de parámetros w que maximiza la probabilidad posterior

$$W_{MAP} = \underset{w}{\operatorname{argmax}} \log \left(\prod_{n=1}^N G(t_n | \phi(x_n) w^\top, \sigma_n^2) \prod_{q=1}^Q G(w_q | 0, \sigma_w^2) \right)$$

prior

$$W_{MAP} = \underset{w}{\operatorname{argmax}} -\frac{1}{2\sigma_n^2} \|t - \Phi w^\top\|_2^2 - \frac{1}{2\sigma_w^2} \|w\|_2^2$$

Ahora se convierte el problema de optimización (maximizar \rightarrow minimizar) = (multiplicar por -1)

$$W_{MAP} = \underset{w}{\operatorname{argmin}} \left[\frac{1}{2\sigma_n^2} \|t - \Phi w^\top\|_2^2 + \frac{1}{2\sigma_w^2} \|w\|_2^2 \right]$$

Para facilitar la minimización se multiplica en ambos lados por σ_n^2 :

$$W^{MAP} = \underset{W}{\operatorname{argmin}} \left(\|t - \Phi W\|_2^2 + \frac{\sigma_n^2}{\sigma_w^2} \|W\|_2^2 \right)$$

Bajo estas suposiciones, el problema de optimización de MAP asumiendo ruido y prior Gaussiana, es equivalente a la optimización de mínimos cuadrados regularizados con $\lambda = \sigma_n^2 / \sigma_w^2$:

$$W^{MAP} = \left(\Phi^\top \Phi + \frac{\sigma_n^2}{\sigma_w^2} \mathbb{I} \right)^{-1} \Phi^\top t$$

5) bayesiano con modelo lineal Gaussiano:

Se quiere estimar una distribución posterior para los pesos W en el modelo lineal con ruido gaussiano, y usarla para hacer predicciones con incertidumbre.

El enfoque bayesiano introduce incertidumbre sobre los parámetros W mismos. Se asume un prior gaussiano:

$$P(W) = N(W | m_0, S_0)$$

Donde:

- $m_0 \in \mathbb{R}^M$: media previa de W
- $S_0 \in \mathbb{R}^{MM}$: Matriz de covarianza previa, que refleja la confianza sobre W antes de ver los datos. Por ejemplo, si se asume que todos los pesos son independientes con varianza σ_w^2 , $S_0 = \sigma_w^2 \mathbb{I}$ y $m_0 = 0$

Como el prior y la verosimilitud son Gaussiana, el posterior también es Gaussiana (Propiedades de conjugación).

$$\text{Para combinar: } P(t|W) \propto \exp \left(-\frac{1}{2\sigma_n^2} \|t - \Phi W\|^2 \right)$$

$$P(W) \propto \exp \left(-\frac{1}{2} (W - m_0)^\top S_0^{-1} (W - m_0) \right)$$

Multiplicando y completando cuadrados, el posterior es:

$$S_N = \left(S_0^{-1} + \frac{1}{\sigma_n^2} \Phi^\top \Phi \right)^{-1} \quad m_N = S_N \left(S_0^{-1} m_0 + \frac{1}{\sigma_n^2} \Phi^\top t \right)$$

Interpretación:

- S_N es la covarianza de la posterior, combina la incertidumbre previa y la información de los datos.
- m_N es la media del posterior, que representa la estimación Bayesiana actualizada de los parámetros.

Se impone un prior de la forma:

$$P(W) = N(W | 0, \sigma_w^2 \mathbb{I})$$

Entonces:

$$P(w|t) = N(w | \tilde{m}_N, \tilde{S}_N)$$

$$\tilde{m}_N = \frac{1}{\sigma_w^2} \left(\frac{1}{\sigma_n^2} \right)^{-1} \left(\frac{\sigma_n^2}{\sigma_w^2} \mathbb{I}Q + \Phi^\top \Phi \right)^{-1} \quad \tilde{S}_N = \frac{1}{\sigma_w^2} \mathbb{I}Q$$

Reemplazando en la media condicional:

$$\tilde{m}_N = \frac{1}{\sigma_n^2} \left(\frac{1}{\sigma_n^2} \right)^{-1} \left(\frac{\sigma_n^2}{\sigma_w^2} \mathbb{I}Q + \Phi^\top \Phi \right)^{-1} \Phi^\top t$$

$$\tilde{m}_N = \left(\frac{\sigma_n^2}{\sigma_w^2} \mathbb{I}Q + \Phi^\top \Phi \right)^{-1} \Phi^\top t$$

NOTA: La solución del modelo lineal Gaussiano para el prior $p(w) = N(w | 0, \sigma_w^2)$ y ante ruido blanco Gaussiano $p(y_n | w) = N(y_n | \phi(x_n)^\top w, \sigma_n^2)$, es equivalente en la media \tilde{m}_N a la solución de mínimos cuadrados regularizados.

Predictiva

Para un nuevo dato x_* , la distribución predictiva referente a la salida t_* se puede calcular como:

$$P(t_* | x_*, t, w) = \int P(t_* | x_*, w) P(w | t) dw$$

$$P(t_* | t) = \int N(t_* | \phi(x_*)^\top w, \sigma_n^2) N(w | \tilde{m}_N, \tilde{S}_N) dw$$

$$P(t_* | x_*, t, w) = N(t_* | \phi(x_*)^\top \tilde{m}_N, \sigma_n^2 + \phi(x_*)^\top \tilde{S}_N \phi(x_*)^\top)$$

6) Regresión Rígida Kernel

Se quiere predecir lo siguiente: $f(x) = \phi(x)w$

Dónde:

- $\phi(w) \in \mathbb{R}^Q$: Vector de características no lineales de la entrada x .

• $w \in \mathbb{R}^Q$: Vector de pesos en el espacio transformado

Problema de regresión de Ridge:

Dado un conjunto de entrenamiento $\{(x_n, y_n)\}_{n=1}^N$, se minimiza:

$$W_{RK} = \frac{1}{N} \sum_{n=1}^N (y_n - \phi(x_n)^\top w)^2 + \lambda \|w\|^2$$

En forma matricial:

- Se define la matriz $\Phi \in \mathbb{R}^{N \times Q}$, donde la fila n es $\phi(x_n)$
- El vector $y \in \mathbb{R}^N$ contiene las etiquetas y_n

$$\text{Entonces: } W_{RK} = \frac{1}{N} \|y - \Phi w\|^2 + \lambda \|w\|^2$$

$$W_{RK} = \left(\frac{1}{N} (y^\top y - 2y^\top \Phi w + (\Phi w)^\top (\Phi w)) + \lambda w^\top w \right)$$

Derivando e igualando a cero:

$$\frac{\partial}{\partial w} \left(\frac{1}{N} (y^T y - 2y^T \Phi w + (\Phi^T \Phi w)^T (\Phi^T \Phi w)) + \lambda w^T w \right) = 0$$

$$-\frac{2}{N} \Phi^T y + \frac{2}{N} \Phi^T \Phi w + 2\lambda w = 0$$

$$\text{Dividiendo entre } 2: \frac{1}{N} \Phi^T y = \frac{1}{N} \Phi^T \Phi w + \lambda w$$

$$\text{Se multiplica por } N: \Phi^T y = \Phi^T \Phi w + N\lambda w$$

$$\text{Se agrupan términos: } \Phi^T y = (\Phi^T \Phi + N\lambda I)w$$

Recordar: w es un vector y I es la matriz identidad del mismo tamaño de $\Phi^T \Phi$.

Despejando w :

$$w_{PRK}^* = (\Phi^T \Phi + N\lambda I)^{-1} \Phi^T y$$

Problema: $\phi(w)$ no se puede calcular directamente:

Si $\phi(w) \in \mathbb{R}^Q$ y $Q \rightarrow \infty$, no se puede construir ni almacenar $\phi(w)$ ni la matriz Φ . solo aparecen productos escalares:

• Φw = Vector con entradas $\phi(x_n)w$

• $\Phi^T \Phi$ = Suma de productos escalares $\phi(x_n) \cdot \phi(x_m)$

Una función Kernel $K(x, x')$, da directamente el producto escalar $\phi(x) \cdot \phi(x')$ en un espacio de características (posiblemente de dimensión infinita), sin necesidad de calcular $\phi(w)$ explícitamente. Así, se evita construir o manejar el espacio transformado.

función kernel: $K(x, x') = \phi(w) \cdot \phi(x')$

Ahora se expresa la solución en términos de combinaciones de los datos mediante α , usando solo productos escalares:

$$w_{PRK}^* = \sum_{n=1}^N \alpha_n \phi(x_n)$$

Es decir: $w_{PRK}^* = \Phi^T \alpha$

convirtiéndose el modelo en:

$$f(x) = \phi(x) w = \phi(x) (\Phi^T \alpha) = \sum_{n=1}^N \alpha_n \phi(x) \cdot \phi(x_n)$$

y usando el Kernel:

$$f(x) = \sum_{n=1}^N \alpha_n K(x, x_n)$$

$\alpha_n \in \mathbb{R}$ y se encuentra mediante mínimos cuadrados regularizados en RKHS.

7) Procesos Gaussianos:

Dado un conjunto de datos de entrenamiento:

$$D = \{(x_i, y_i)\}_{i=1}^n$$

Donde: $t_i = f(x_i) + \eta$, $\eta \sim N(0, \sigma_n^2)$

Se quiere predecir el valor de t_* , en un nuevo punto x_*

Ahora, se define el proceso Gaussiano para la regresión; asumiendo que $f(x)$ es una función sacada de un proceso gaussiano con $m(x) = 0$ y Kernel $K(x, x')$ o Kernel gaussiano:

$$K(x, x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right)$$

Donde: σ_f^2 , Varianza de la función

l , longitud de la escala

se define la matriz de Covarianza:

$$\text{Sea } X = [x_1, x_2, x_3, \dots, x_n]^T$$

Se construyen las matrices:

- $K \in \mathbb{R}^{n \times n}$ con entradas $K_{ij} = K(x_i, x_j)$ (Covarianza entre puntos de entrenamiento)
- $K_* \in \mathbb{R}^n$ vector con entradas $K(x_i, x_*)$ (Covarianza entre los datos y el punto a predecir)
- $K_{**} = K(x_*, x_*)$ (Varianza en el punto nuevo)

La distribución conjunta de los valores observados y el valor en el nuevo punto, dado que el proceso es Gaussiano, la distribución conjunta es:

$$\begin{bmatrix} t \\ t_* \end{bmatrix} \sim N \left(\begin{bmatrix} t \\ t_* \end{bmatrix}, \begin{bmatrix} 0 & K + \sigma_n^2 & K_* \\ 0 & K^T & K(x_*, x_*) + \sigma_n^2 \end{bmatrix} \right)$$

con $K_* = [K(x_*, x)]$

La probabilidad condicional $p(t_* | f(x_*), f(x))$ se puede determinar como:

$$p(t_* | f(x_*), f(x)) = N(t_* | m(x_*), \text{Cov}(f(x_*), f(x)))$$

$$\text{con: } m(x_*) = K_*^T (K + \sigma_n^2 I)^{-1} t$$

$$\text{Cov}(f(x_*), f(x)) = K(x_*, x_*) + \sigma_n^2 - K_*^T (K + \sigma_n^2 I)^{-1} K_*$$