

# Parcial #1: Teoría de Aprendizaje de Máquina

Julian Felipe Gutierrez Ramírez - 1193577231.

1. Sea el modelo de regresión  $t_n = \phi(\mathbf{x}_n)\mathbf{w}^\top + \eta_n$ , con  $\{t_n \in \mathbb{R}, \mathbf{x}_n \in \mathbb{R}^P\}_{n=1}^N$ ,  $\mathbf{w} \in \mathbb{R}^Q$ ,  $\phi : \mathbb{R}^P \rightarrow \mathbb{R}^Q$ ,  $Q \geq P$ , y  $\eta_n \sim \mathcal{N}(\eta_n | 0, \sigma_\eta^2)$ . Presente el problema de optimización y la solución del mismo para los modelos mínimos cuadrados, mínimos cuadrados regularizados, máxima verosimilitud, máximo a-posteriori, Bayesiano con modelo lineal Gaussiano, regresión rígida kernel y mediante procesos Gaussianos. Asuma datos i.i.d. Discuta las diferencias y similitudes entre los modelos estudiados.

## ① Modelo mínimos cuadrados:

Problema de Optimización:

$$W_{mc} = \underset{\mathbf{w}}{\operatorname{argmin}} \quad \frac{1}{N} \sum_{n=1}^N (t_n - \phi(\mathbf{x}_n)\mathbf{w}^\top)^2$$

En su forma matricial:

$$W_{mc} = \frac{1}{N} \|t - \Phi w\|^2$$

donde  $\Phi$  es la matriz de diseño Con  $\Phi_n = \phi(\mathbf{x}_n)$

Antes de derivar, se va a reescribir la función de costo  $\|t - \Phi w\|^2$  de la siguiente manera:

$$\|t - \Phi w\|^2 = \|P\|^2 = P^\top P$$

Regresando a la función:

$$\langle t - \Phi w, t - \Phi w \rangle = (t - \Phi w)^\top (t - \Phi w)$$

$$= [t^T - (\underline{\Phi}W)^T](t - \underline{\Phi}W)$$

Se realiza la distribución quedando de la siguiente manera:

$$= t^T t - \underbrace{t^T \underline{\Phi}W}_{\textcircled{1}} - \underbrace{(\underline{\Phi}W)^T t}_{\textcircled{2}} + (\underline{\Phi}W)^T (\underline{\Phi}W)$$

Analizando las multiplicaciones subrayada ① y ②:

$$\textcircled{1} \quad t^T \underline{\Phi} W$$

$1 \times N \quad N \times P \quad P \times 1$

$\begin{matrix} 1 \times P \\ \downarrow \\ 1 \times P \end{matrix}$

$1 \times 1$

$$\textcircled{2} \quad (\underline{\Phi}W)^T t = W^T \underline{\Phi}^T t$$

$1 \times P \quad P \times N \quad N \times 1$

$\begin{matrix} t^T \\ \downarrow \\ 1 \times N \end{matrix}$

$N \times 1$

Se puede realizar

Se puede realizar

$$\|t - \underline{\Phi}W\|^2 = t^T t - t^T \underline{\Phi}W - (\underline{\Phi}W)^T t + (\underline{\Phi}W)^T (\underline{\Phi}W)$$

$$= t^T t - 2t^T \underline{\Phi}W + (\underline{\Phi}W)^T (\underline{\Phi}W)$$

Ahora sí, se deriva respecto  $W$ :  $(ab)^T = b^T a^T$

$$\frac{\partial}{\partial W} \{ t^T t - 2t^T \underline{\Phi}W + (\underline{\Phi}W)^T (\underline{\Phi}W) \} = -2t^T \underline{\Phi} + \frac{\partial}{\partial W} \{ W^T \underline{\Phi}^T \underline{\Phi} W \}$$

cuadrático en  $W$

$$= -2t^T \underline{\Phi} + 2\underline{\Phi}^T \underline{\Phi} W$$

$1 \times N \quad N \times P \quad P \times 1$

$\begin{matrix} P \times N \quad N \times P \\ \downarrow \\ P \times P \end{matrix}$

$P \times 1$

Para que se pueda realizar la suma entre matrices, tiene que ser del mismo tamaño, por lo tanto, se transpone  $(\underline{\Phi}^T \underline{\Phi})^T$

$$\frac{\partial}{\partial w} \left\{ t^T t - 2t^T \underline{\Phi} w + (\underline{\Phi} w)^T (\underline{\Phi} w) \right\} = -(2t^T \underline{\Phi})^T + 2\underline{\Phi}^T \underline{\Phi} w \\ = -2\underline{\Phi}^T t + 2\underline{\Phi}^T \underline{\Phi} w$$

Je iguala a cero para encontrar el mínimo:

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

$$2\underline{\Phi}^T t = 2\underline{\Phi}^T \underline{\Phi} w$$

Dividiendo en ambos lados por 2:

$$\underline{\Phi}^T t = \underline{\Phi}^T \underline{\Phi} w$$

Multiplicando ambos lados por  $(\underline{\Phi}^T \underline{\Phi})^{-1}$ , para despejar w:

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

II

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

## ② Mínimos cuadrados Regularizados:

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

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

En su forma matricial queda de la siguiente manera:

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

donde:  $t = [t_1, t_2, \dots, t_N]^T \in \mathbb{R}^N$

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

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

La función de costo queda de la siguiente manera:

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

Ahora se deriva respecto  $w$ :  $(ab)^T = b^T a^T$

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

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

$$\bar{\Phi}^T t = \bar{\Phi}^T \bar{\Phi} w + \lambda w, \text{ Se factoriza:}$$

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

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

### ③ Máxima Verosimilitud:

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

$$\eta_n \sim p(\eta_n) = G(\eta_n | 0, \sigma_n^2)$$

Con:

$$t_n = \phi(x_n) w^T + \eta_n$$

$$\eta_n = t_n - \phi(x_n) w^T$$

Por lo tanto

$$p(t_n | \underbrace{\phi(x_n) w^T}_{\times}, \sigma_n^2) = G(t_n | \phi(x_n) w^T, \sigma_n^2)$$

Se pueden encontrar los pesos y la varianza maximizando el log-Verosimilitud

$$W_{ML} = \underset{w, \sigma^2}{\operatorname{arg \max}} \log \left( \prod_{n=1}^N G(t_n | \underbrace{\phi(x_n) w^T}_{\Phi}, \sigma_n^2) \right)$$

Por lo tanto:

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

Como la varianza del ruido ( $\sigma_n^2$ ) es constante respecto a  $w$ , esto equivale a:

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

Finalmente se convierte el problema de optimización a minimización:

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

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

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

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

#### ④ Máximo a-posteriori:

El modelo Máximo a-posteriori es un modelo bayesiano porque incorpora la incertidumbre en los parámetros del 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 máximos 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 idénticamente distribuidos; el máximo a-posteriori dado un conjunto de datos, busca encontrar el vector de parámetros  $w$  que maximiza la probabilidad posterior.

$$W_{MAP} = \arg \max_w \log \left( \prod_{n=1}^N N(t_n | \Phi(x_n)w^\top, \Sigma_n^2) \right) \frac{Q}{\prod_{q=1}^Q N(W_q | 0, \Sigma_w^2)} \right)$$

$$\log \left( \prod_{n=1}^N N(t_n | \Phi(x_n)w^\top, \Sigma_n^2) \right) = -\frac{1}{2\Sigma_n^2} \|t - \Phi w^\top\|_2^2 - \frac{1}{2\Sigma_w^2} \|w\|_2^2$$

Quedando el problema de Optimización de la siguiente manera:

$$W_{MAP} = \arg \max_w \left( -\frac{1}{2\Sigma_n^2} \|t - \Phi w^\top\|_2^2 - \frac{1}{2\Sigma_w^2} \|w\|_2^2 \right)$$

Ahora, se convierte el problema de Optimización (Maximizar  $\rightarrow$  Minimizar):

$$W_{MAP} = \arg \min_w \left[ (-1) \left( -\frac{1}{2\Sigma_n^2} \|t - \Phi w^\top\|_2^2 - \frac{1}{2\Sigma_w^2} \|w\|_2^2 \right) \right]$$

$$W_{MAP} = \arg \min_w \left( \frac{1}{2\Sigma_n^2} \|t - \Phi w^\top\|_2^2 + \frac{1}{2\Sigma_w^2} \|w\|_2^2 \right)$$

Para facilitar la minimización Se multiplican en ambos lados por  $2\Sigma^{-2}$ :

$$W_{MAP}^* = \arg \min_w \left( \|t - \Phi w^\top\|_2^2 + \frac{\Sigma_n^2}{\Sigma_w^2} \|w\|_2^2 \right)$$

Bajo estas suposiciones, el problema de Optimización de MAP asumiendo ruido y prior Gaussianos, 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 un 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:

$$p(w) = N(w | m_0, S_0)$$

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

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

Para combinar:

$$p(t | w) \propto \exp \left( -\frac{1}{2\sigma_n^2} \|t - \Phi w\|^2 \right)$$

$$p(w) \propto \exp \left( -\frac{1}{2} (w - m_0)^T S_0^{-1} (w - m_0) \right)$$

Multiplicando y completando cuadrados, el posterior es:

$$S_N = \left( S_0^{-1} + \frac{1}{\sigma_n^2} \bar{\Phi}^T \bar{\Phi} \right)^{-1}$$

$$\mathbf{m}_N = S_N \left( S_0^{-1} \mathbf{m}_0 + \frac{1}{\sigma_n^2} \bar{\Phi}^T \mathbf{t} \right)$$

Interpretación:

- $S_N$  es la covarianza de la posterior, combina la incertidumbre previa y la información de los datos.
- $\mathbf{m}_N$  es la media del posterior, que representa la estimación bayesiana actualizada de los parámetros.

Se impone un prior de la forma:

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

Entonces:

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

$$\tilde{\mathbf{m}}_N = \frac{1}{\sigma_n^2} \left( \frac{1}{\sigma_w^2} \right)^{-1} \left( \frac{\sigma_n^2}{\sigma_w^2} \mathbb{I}_Q + \bar{\Phi}^T \bar{\Phi} \right)^{-1}$$

$$\tilde{S}_N = \left( \frac{1}{\sigma_w^2} \mathbb{I}_Q \right)$$

Reemplazando en la media condicional:

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

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

**NOTA:** La solución del modelo lineal Gaussiano para el prior  $p(w) = N(w|0, \Omega_w^{-2})$  y ante ruido blanco Gaussiano  $\eta_n \sim p(\eta_n) = N(\eta_n|0, \Omega_n^{-2})$ , 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 variable  $t_*$  se puede calcular como:

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

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

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

⑥ Regresión Rígida Kernel:

Se quiere predecir lo siguiente:

$$f(x) = \phi(x)w$$

Donde: •  $\phi(x) \in \mathbb{R}^Q$ : 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:

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

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  $y_n$ .

Entonces:

$$W_{Ridge} = \frac{1}{N} \|y - \Phi w\|^2 + \lambda \|w\|^2$$

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

Derivando e igualando a cero:

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

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

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

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

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

Recorda:  $w$  es un vector e  $\mathbb{I}$  es la matriz identidad del mismo tamaño de  $\bar{\Phi}^T \bar{\Phi}$ .

Despejando  $w$ :

$$w_{\text{Ran}}^* = (\bar{\Phi}^T \bar{\Phi} + N\lambda \mathbb{I})^{-1} \bar{\Phi}^T y$$

Problema:  $\Phi(x)$  No se puede calcular directamente:

Si  $\Phi(x) \in \mathbb{R}^Q$  y  $Q \rightarrow \infty$ , no se puede construir ni almacenar  $\Phi(x)$  ni la matriz  $\bar{\Phi}$ . Solo aparecen productos escalares:

- $\bar{\Phi}w$  = Vector con entradas  $\Phi(x_n)w$
- $\bar{\Phi}^T \bar{\Phi}$  = 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  $\Phi(x)$  explícitamente. Así, se evita construir o manejar el espacio transformado.

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

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

$$w_{\text{Ran}}^* = \sum_{n=1}^N \alpha_n \Phi(x_n)$$

Eso decir:  $W_{\text{RBF}}^* = \Phi^T \alpha$

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 regulados en RKHS.

## ⑦ Procesos Gaussiano:

Dado un conjunto de datos de entrenamiento:

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

Donde:

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

Se quiere predecir el valor  $t_*$ , 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:  $\bar{f}_f^2$ , Variante de la función

$l$ , longitud de la escala.

Se define la matriz de Covarianza:

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:

$$\begin{bmatrix} t \\ t_* \end{bmatrix} \sim N \left( \begin{bmatrix} t \\ t_* \end{bmatrix} \mid \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K + \bar{f}_f^2 \mathbb{I} & K_* \\ K^T & K(x_*, x_*) + \bar{f}_f^2 \end{bmatrix} \right)$$

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

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

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

$$\text{Con: } m(x_*) = K_*^\top (K + \sigma_\eta^2 I)^{-1} t$$

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