Parcial 1:

modelo : $t_n = \phi(x_n)\, w + n_n \;;\; n = 1, \dots, N.$

- $t_n \in \mathbb{R}$ → Respuesta
- $x_n \in \mathbb{R}^p$ → Entrada.
- $\phi : \mathbb{R}^p \to \mathbb{R}^Q$ → Características
- $w \in \mathbb{R}^Q$ → Parametros a Estimar
- n_n → ruido independiente $n_n \sim N(0, \sigma^2)$
- Matriz de diseño $\phi \in \mathbb{R}^{N \times Q}$ filas $\phi(x_n)^T$.
- Vector observaciones

1| Minimos Cuadrados:

Forma Vectorial ↓

minimizar Error / Perdida :

$$J(w) = \frac{1}{2} \| t - \phi w \|_2^2 = \frac{1}{2} (t - \phi w)^T (t - \phi w)$$

$$(t - \phi w)^T (t - \phi w) = t^T t - t^T \phi w - \phi^T w^T t + \phi^T w^T \phi w$$

$$= t^T t - 2\, w^T \phi^T t + w^T \phi^T \phi w$$

- $t^T \phi w$ → Escalar / su transpuesta es igual al escalar
- $t^T \phi w = \phi^T w^T t$

$$J(w) = \frac{1}{2} \left(t^T t - 2\, w^T \phi^T t + w^T \phi^T \phi w \right)$$

$$= \frac{1}{2} t^T t - w^T \phi^T t + \frac{1}{2} w^T \phi^T \phi w$$

↓ 0 ↑ A

• Derivamo $J(w)$ / w. $A = \Phi^T \Phi$.

• $\frac{\partial}{\partial w} \frac{1}{2}(w^T A w) = \frac{1}{2}(A + A^T) w = \frac{1}{2}(2Aw) = Aw$

termine Quadratico

termine Lineal:

$\frac{\partial}{\partial w}(\Phi^T t)\, w = \Phi^T t$

• $\frac{\partial J}{\partial w} = -\Phi^T t + Aw \rightarrow -\Phi^T t + \Phi \Phi^T w$

igualamos a Cero "condicion optimo gradiente"

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

$\Phi^T \Phi w = \Phi^T t$

$w_{ols} = (\Phi^T \Phi)^{-1} \Phi^T t$ //

↳ minimo global
Gradiente = 0.

2) Mínimos Cuadrados Regularizados:

$$J_\lambda(w) = \frac{1}{2}\|t - \Phi w\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \quad \lambda > 0.$$

T_1 T_2

Parámetro de regularización

$\|w\|_2^2 = w^T w$ $\quad \hookrightarrow \frac{\lambda}{2} w^T w$.

Derivadas:

T_1: $-\Phi^T t + \Phi\Phi^T w$ - Calculada anteriormente.

T_2: λw.

Gradiente: $-\Phi^T t + \Phi\Phi^T w + \lambda w$

- Condición de mínimo

$$\nabla_w J_\lambda(w) = 0$$

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

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

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

3.) Maxima Verosimilitud.

- Verosimilitud $p(t|w)$

$$p(t_n|w) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(t_n - \phi(x_n)^T w)^2}{2\sigma_n^2}\right)$$

- función de densidad Normal

$$\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- independientes, Ver. Conjunta

$$p(t|w) = \prod_{n=1}^{N} p(t_n|w) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(t_n - \phi(x_n)^T w)^2}{2\sigma_n^2}\right)$$

$$p(t|w) = (2\pi\sigma_n^2)^{-N/2} \exp\left(-\frac{1}{2\sigma_n^2} \sum_{n=1}^{N} (t_n - \phi(x_n)^T w)^2\right)$$

Nota: $\sum (t_n - \phi_n^T w)^2 = \|t - \Phi w\|_2^2$

$$P(t|w) = (2\pi\sigma_n^2)^{-N/2} \exp\left(\frac{-1}{2\sigma_n^2} \|t - \phi w\|_2^2\right)$$

- Log - Verosimilitud

$$L(w) = \log P(t|w) = -\frac{N}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\|t-\phi w\|_2^2$$

↑ No depende de w.

$$\arg\max L(w) = \arg\max\left(\frac{-1}{2\sigma_n^2}\|t-\phi w\|_2^2\right)$$

Nota: maximizar la Ver. similitud. = minimizar la Suma de cuadrados

$$= \arg\min \frac{1}{2\sigma_n^2}\|t-\phi w\|_2^2.$$

- Derivamos Log - Likelihood

$$g(w) = \frac{1}{2\sigma_n^2}\|t-\phi w\|_2^2$$

$$\text{Derivada} = \frac{\partial}{\partial w}\|t-\phi w\|_2^2 = (t-\phi w)(t-\phi w)^T$$

$$= -\frac{1}{2\sigma_n^2}(2\phi^T t - 2\phi\phi^T w) = \frac{1}{\sigma_n^2}(\phi^T t - \phi\phi$$

$$\nabla_w L(w) = \frac{1}{\sigma_n^2} \Phi^T (t - \Phi w)$$

* Máximo interior / mínimo de la suma de ^2 dos.

$$\nabla_w L(w) = 0 \rightarrow \Phi^T (t - \Phi w) = 0.$$

$$\Phi \Phi^T w = \Phi^T t$$

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

* Segunda derivada:

$$H(w) = -\frac{1}{\sigma_n^2} \Phi^T \Phi$$

Nota:

1. Likelihood: $p(t|w) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \|t - \Phi w\|_2^2\right)$
2. Log. lik.: $l(w) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t - \Phi w\|_2^2$
3. Grad: $\nabla_w l(w) = \frac{1}{\sigma^2} \Phi^T (t - \Phi w)$.
4. $= 0$: $\Phi^T \Phi w = \Phi^T t$.
5. Soluc.: $(\Phi^T \Phi)^{-1} \Phi t$
6. Hessiano: $H = -\frac{1}{2\sigma^2} \Phi^T \Phi$ → máximo global

4. max. a- Posteriori:

Suponemos un Prior (a Priori) w también gaussiano

$$p(w) = \mathcal{N}(w | 0, \sigma_w^2 I)$$

a) Likelihood $\rightarrow$ $p(t|w, \sigma^2)$

$$= (2\pi\sigma^2)^{N/2} \exp\left[-\frac{1}{2\sigma^2}(t-\phi w)^T(t-\phi w)\right]$$

b) Prior. $p(w|\sigma_w^2) = (2\pi\sigma_w^2)^{-M/2} \exp\left[-\frac{1}{2\sigma_w^2} w^T w\right]$

- log Posterior.

$$\log p(w|t) = \log p(t|w) + \log p(w) + \text{Const}$$

$$= -\frac{1}{2\sigma^2}(t-\phi w)^T(t-\phi w) - \frac{1}{2\sigma_w^2} w^T w + \text{Const}$$

* Maximo Posterior: / minimizar su negativo

$$w_{MAP} = \arg\min\left[\frac{1}{2\sigma^2}(t-\phi w)^T(t-\phi w) + \frac{1}{2\sigma_w^2} w^T w\right]$$

Error Cuadrático — Penalización por pesos grandes (Regularización)

$\lambda = \frac{\sigma^2}{\sigma_w^2}$

$$J(w) = \frac{1}{2}(t-\phi w)^T (t-\phi w) + \frac{\sigma^2}{2\sigma_w^2} w^T w \quad \leftarrow \sigma^2$$

$$J(w) = \frac{1}{2}\|t-\phi w\|_2^2 + \frac{\lambda}{2}\|w\|^2$$

1) Derivamos el caso $/w$.

$$J(w) = \frac{1}{2}(t^T t - 2t^T \phi w + w^T \phi^T \phi w) + \frac{\lambda}{2} w^T w$$

- $\nabla \frac{1}{2} w^T \phi^T \phi w = \frac{1}{2}(2\phi^T \phi w) = \phi^T \phi w \rightarrow w^T = w \rightarrow$ Simetrica
- $\nabla(-t^T \phi w) = -\phi^T t$
- $\nabla \frac{\lambda}{2} w^T w = \frac{\lambda}{2}(2w) = \lambda w$

Nota: Reglas: ∇ matricial

$w^T A w$	$2Aw$
$C^T w$	C
$w^T w$	$2w$

∇ total: $-\phi^T (t - \phi w) + \lambda w$.

- $-\phi^T (t - \phi w)$: Grad. Error. → Error Real / w real
- λw: Grad. término regularizador. Evita sobreajuste.

$\nabla = 0$.

$$\phi^T \phi w + \lambda w = \phi^T t$$

$$(\phi^T \phi + \lambda I) w = \phi^T t$$

$$w_{map} = (\phi^T \phi + \lambda I)^{-1} \phi^T t \quad ; \quad \lambda = \frac{\sigma^2}{\sigma_w^2}$$

$\phi^T \phi$: Con. Correlac. — λI: Prior Gaussiano.

5. Bayesiano Con modelo lin G.

Usaremos:

Likelih: $p(t|w) = \mathcal{N}(t | \phi w, \sigma^2 I)$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-1}{2\sigma^2} \|t - \phi w\|_2^2\right)$$

Prior: $p(w) = \mathcal{N}(w | 0, \Sigma_p)$

$$= (2\pi)^{-Q/2} |\Sigma_p|^{-1/2} \exp\left(-\frac{1}{2} w^T \Sigma_p^{-1} w\right)$$

2) Posterior = Prior × likelihood

$$\log p(w|t) = \frac{-1}{2\sigma^2}(t-\Phi w)^T(t-\Phi w) - \frac{1}{2} w^T \Sigma_p^{-1} w + \text{Const}$$

$$= \frac{-1}{2\sigma^2}(t^T t - 2w^T \Phi^T t + w^T \Phi^T \Phi w) - \frac{1}{2} w^T \Sigma_p^{-1} w + \text{Const}$$

$$= \frac{1}{2}\left[w^T \left(\Sigma_p^{-1} + \frac{1}{\sigma^2} \Phi^T \Phi \right) w - 2\left(\frac{1}{\sigma^2} \Phi^T t \right)^T w \right] + C$$

$A = \Sigma_p^{-1} + \frac{1}{\sigma^2} \Phi^T \Phi \qquad b = \frac{1}{\sigma^2} \Phi^T t$

$$= \frac{1}{2} w^T A w - 2 b^T w.$$

Completa Cuadrado

$$w^T A w - 2 b^T w = (w - A^{-1} b)^T A (w - A^{-1} b) - b^T A^{-1} b$$

$$\log p(w|t) = \frac{-1}{2}(w - A^{-1} b)^T A (w - A^{-1} b)$$

$$S_N = A^{-1} = \left(\Sigma_p^{-1} + \frac{1}{\sigma^2} \Phi^T \Phi \right)^{-1}$$

$$m_N = A^{-1} b = \frac{1}{\sigma^2} S_N \Phi^T b.$$

Regresion Rigida por Kernel

Queremos permitir no linealidad sin trabajar explicitamente con w en $\mathbb{R}^Q$, usando un kernel
$k(x, x') = \phi(x)^T \phi(x')$

Ridge en espacio de caracteristicas

$$\min_w \|t - \Phi w\|^2 + \lambda \|w\|^2$$

Solucion dual (teorema de representacion):

La solucion w pertenece al subespacio generado por las filas de Φ, es decir, existe $\alpha \in \mathbb{R}^N$ tal que $w = \Phi^T \alpha$.

$$t - \Phi w = t - \Phi \Phi^T \alpha = t - K\alpha$$

$K = \Phi \Phi^T \in \mathbb{R}^{N \times N}$ es la matriz de Kernel con $K_{ij} = k(x_i, x_j)$

El problema en α es:

$$\min_\alpha \|t - K\alpha\|^2 + \lambda \alpha^T \Phi \Phi^T \alpha = \|t - K\alpha\|^2 + \lambda \alpha^T K \alpha$$

Derivando e igualando a cero:

$$(K + \lambda I)\alpha = t \Rightarrow \alpha = (K + \lambda I)^{-1} t$$

La prediccion para un nuevo x_* usa:

$$\hat{t}_* = \phi(x_*)^T w = \phi(x_*)^T \Phi^T \alpha = k_*^T \alpha$$

Donde $k_* = (k(x_*, x_1), \ldots, k(x_*, x_N))^T$

Por lo tanto: $\hat{t}_* = k_*^T (K + \lambda I)^{-1} t$

Interpretacion: kernel ridge permite operar con k directamente sin calcular Φ. Es equivalente a Ridge en el espacio de caracteristicas.

Regresion Por proceso Gaussiano

Un Proceso Gaussiano (GP) pone un prior directamente sobre funciones $f(\cdot)$

$$f(\cdot) \sim GP(0, k(\cdot,\cdot))$$

$t_n = f(x_n) + \eta_n$ con $\eta_n \sim \mathcal{N}(0, \sigma_\eta^2)$

Prior Conjunto sobre valores observados y no observados

Sea $f = [f(x_1), \ldots, f(x_N)]^T$ Entonces

$$p(f) = \mathcal{N}(0, K), \quad K_{ij} = k(x_i, x_j)$$

$t = f + \eta$ con $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I)$ por lo tanto.

$$p(t) = \mathcal{N}(0, k + \sigma_\eta^2 I)$$

Predictiva para x_*

La distribucion Conjunta de $\begin{pmatrix} t \\ f_* \end{pmatrix}$ es gaussiana, Condicionando se obtiene la predictiva normal.

$$p(f_* \mid x_*, t) = \mathcal{N}(\mu_*, \sigma_*^2)$$

$$\mu_* = k_*(k + \sigma_\eta^2 I)^{-1} t$$

$$\sigma_*^2 = k(x_*, x_*) - k_*(k + \sigma_\eta^2 I)^{-1} k_*$$

Si deseamos la predictiva sobre $t_* = f_* + \eta_*$ añadimos σ_η^2 a la varianza.

Relacion Con Kernel ridge y modelo Bayesiano lineal

La media predictiva de GPR: $\mu_* = k_*^T (k + \sigma_n^2 I)^{-1} t$

Es identica a la prediccion de kernel ridge con $\lambda = \sigma_\eta^2$

GPR y kernel ridge comparten la misma media predictiva cuando se toma hiperparametros coincidentes. la diferencia principal es que GPR entrega una varianza predictiva natural.

El GP puede verse como el limite no parametrico del modelo lineal bayessiano cuando el numero de caracteristicas Q tiende a infinito apropiadamente y el prior induce el kernel k.