

1) Seer el modelo de regresión

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

con  $\{t_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N$   $Q > P$   
 $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$

$$w \in \mathbb{R}^Q$$
$$\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$$

- mínimos cuadrados
- mínimos cuadrados regularizados
- máxima verosimilitud
- máximo a posteriori
- Bayesiano con modelo local gaussiano
- regresión rígida kernel
- predicciones gaussianas
- Salir
- Mínimos cuadrados

Asumimos que las observaciones  $t_n$  están relacionadas con  $x_n$  con el modelo

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

donde

$\phi(x_n)^T w$  es la predicción para la entrada  $x_n$

$w$  son los parámetros

$\eta_n$  es el ruido blanco gaussiano

Suponemos las observaciones  $t_n$  distribuidas como

$$P(t_n | X_n, \omega) = N(t_n | \phi(X_n)^T \omega, \sigma_n^2)$$

→ pera + la probabilidad conjunta (1)

$$p(+|\Phi, \omega) = \prod_{n=1}^N P(t_n | X_n, \omega)$$

→  $\Phi$  es la matriz de características

→ el error en cuanto a donde la predicción

$$e_n = t_n - \hat{t}_n \quad t_n \in \mathbb{R}^N$$

$$e_n = t_n - \phi(X_n)^T \omega$$

Ahora para plantear el problema

la distribución gaussiana de  $t_n$  y  $e_n$

$$P(+|\Phi, \omega) = \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma_n^2} \|t_n - \phi\omega\|^2 \right]$$

donde

$$\|t - \phi\omega\|^2 = (t - \phi\omega)^T (t - \phi\omega)$$

nos interesa minimizar la pérdida promedio del error

l2 norma

$$\text{Asumiendo i.i.d} = E\{X_n\} \approx \frac{1}{N} \sum_{n=1}^N X_n$$

debemos minimizar  $\omega$  de  $\|t - \phi\omega\|^2$

$$\omega^* = \underset{\omega}{\operatorname{argmin}} E\{e_n\}$$

$$\omega^* = \underset{\omega}{\operatorname{argmin}} E\{\|t - \phi\omega\|^2\}$$

$$\mathbb{E} \{ \| + - \Phi \omega \|^2 \} = \frac{1}{N} [ \| + - \Phi \omega \|^2 ]$$

$$\frac{1}{N} ( + - \Phi \omega ) ( + - \Phi \omega ) \left[ +^T + +^T +^T \Phi \omega - \Phi \omega^T + + (\Phi \omega)^T \Phi \omega \right] \frac{1}{N}$$

$$[ +^T + - 2 +^T \Phi \omega + \omega^T \Phi^T \Phi \omega ] \frac{1}{N}$$

luego

$$\frac{\partial}{\partial \omega} \mathbb{E} \{ \| + - \Phi \omega \|^2 \} = 0$$

$$\frac{\partial}{\partial \omega} \mathbb{E} \{ \| + - \Phi \omega \|^2 \} = - 2 ( +^T \Phi )^T + 2 \Phi^T \Phi \omega = 0 \cdot N = 0$$

$$2 \Phi^T \Phi \omega = 2 ( +^T \Phi )^T$$

$$\Phi^T \Phi \omega = \Phi^T +$$

Ahora la pseudo inversa

$$\boxed{\omega^* = (\Phi^T \Phi)^{-1} \Phi^T +}$$

### MINIMOS CUADRATICOS REGULARIZADOS

fuentes de ruido

$$t_n = \Phi(x_n)^T \omega + h_n$$

y el error cuadrático

$$e_n^2 = \| + - \Phi \omega \|_2^2$$

debido a que los datos son ruidosos o datos atípicos. es necesario regularizar los datos en términos de regularización  $L^2$  se define

$$\|\omega\|_2^2 = \sum_{j=1}^Q w_j^2$$

• Inga el termino da regularizacao (j):

$$\lambda \|\omega\|_2^2$$

• Inga termos

$$\|t - \Phi\omega\|_2^2 + \lambda \|\omega\|_2^2$$

→ Achar el problema de optimizacion

$$\omega = \underset{\omega}{\operatorname{Argmin}} \|t - \Phi\omega\|_2^2 + \lambda \|\omega\|_2^2$$

→ se minimiza (usando termos)

$$\|t - \Phi\omega\|_2^2 = (t - \Phi\omega)^T(t - \Phi\omega) + \lambda \omega^T \omega$$

$$\|t - \Phi\omega\|_2^2 = t^T - 2t^T \Phi\omega + \omega^T \Phi^T \Phi\omega$$

$$\omega^* \text{ argmin} [t^T - 2t^T \Phi\omega + \omega^T \Phi^T \Phi\omega + \lambda \omega^T \omega] = 0$$

$$\frac{d}{d\omega} [t^T - 2t^T \Phi\omega + (\omega \Phi)^T \Phi\omega + \omega^T \omega \lambda] = 0$$

$$\frac{d}{d\omega} [t^T - 2t^T \Phi\omega + \omega^T [\Phi^T \Phi + \lambda I] \omega] = 0$$

$$[-2 \Phi^T t] + 2(\Phi^T \Phi + \lambda I) \omega = 0$$

$$-2 \Phi^T t + 2(\Phi^T \Phi + \lambda I) \omega = 0$$

$$2(\Phi^T \Phi + \lambda I) \omega = 2 \Phi^T t$$

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

Maxima verosimilitud

Se tiene el modelo de hipótesis

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

$\eta_n \sim N(\eta_n | 0, \sigma_n^2)$  ruido blanco gaussiano y datos i.i.d  
(suposición)

Entendemos que la verosimilitud

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

Ahora para una sola observación  $(x_n, t_n)$  tenemos la verosimilitud

$$p(t_n | \phi(x_n) w^T, \sigma_n^2) = N(t_n | \phi(x_n) w^T, \sigma_n^2)$$

Ahora basta el supuesto de I.I.D la verosimilitud conjunta

$$p(t | \bar{X}, w, \sigma^2) = \prod_{n=1}^N p(t_n | \phi(x_n) w^T, \sigma_n^2)$$

Ahora (una asunción) los verosimilitudos terminan  $\log(p(t_n | \phi(x_n) w^T, \sigma_n^2))$

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

Ahora la log-verosimilitud es la suma de los log-verosimilitudos individuales

$$\log(p(t)) = \log\left(\prod_{n=1}^N N(t_n | \mu, \sigma^2)\right)$$

$$= \log\left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|t_n - \mu\|^2}{2\sigma^2}\right)\right)$$

$$\log \left( \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \left( \exp \left( \frac{-||x_n - w||^2}{2\sigma^2} \right) \right)$$

Luego tenemos

$$\log \left( \frac{1}{(2\pi\sigma^2)^{N/2}} \right) + \log \left( \exp \left( -\left[ \frac{||x_1 - w||^2}{2\sigma^2} + \frac{||x_2 - w||^2}{2\sigma^2} + \dots + \right] \right) \right)$$

Pedimos escribir el argumento del exponente como la sumatoria de los tramos

$$\log \left( \frac{1}{(2\pi\sigma^2)^{N/2}} \right) + \log \left( \exp \left( \frac{1}{2\sigma^2} \sum_{n=1}^N ||x_n - w||^2 \right) \right)$$

Luego tenemos

$$\log(p(x)) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N ||x_n - w||^2$$

El objetivo es encontrar los parámetros que maximizan la verosimilitud y encajar los pesos y varianzas bgo las suposiciones anteriores, por lo tanto el problema de optimización es el siguiente:

$$w_{ML} = \operatorname{argmax}_{w, \sigma^2} \log \left( \prod_{n=1}^N N(t_n | \Phi(x_n)w^\top, \sigma_n^2) \right)$$

Resumiendo y usando log-verosimilitud

$$w_{ML}, \sigma^2_{ML} = \operatorname{argmax}_{w, \sigma^2} \left[ -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \|\Phi w\|^2 \right]$$

Here derivative respect to  $\sigma^2$

$$\frac{\partial}{\partial \sigma^2} \log p(t|x_i, \omega, \sigma_n^2) = \left( \frac{\partial}{\partial \sigma^2} \left[ -\frac{N}{2} \log(2\pi) \right] - \frac{\partial}{\partial \sigma^2} \left[ \frac{N}{2} \log(\sigma_n^2) \right] \right)$$

$$- \frac{\partial}{\partial \sigma^2} \left[ -\frac{1}{2\sigma^2} \sum_{n=1}^N \|t_n - \phi(x_n)\|^2 \right]$$

Maximizing:

$$\frac{\partial}{\partial \sigma^2} p(t|x_i, \omega, \sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^N \|t_n - \phi(x_n)\omega^\top\|^2 = 0$$

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^N \|t_n - \phi(x_n)\omega^\top\|^2 = 0$$

$$\boxed{\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N \|t_n - \phi(x_n)\omega^\top\|^2}$$

derived respects to  $\omega$ :

$$\frac{\partial}{\partial \omega} \left[ \log p(t|x_i, \omega, \sigma_n^2) \right] = -\frac{1}{2\sigma^2} \sum_{n=1}^N 2(t_n - \phi(x_n)\omega^\top)(-\phi(x_n))$$

$$\frac{\partial}{\partial \omega} \left[ \log p(t|x_i, \omega, \sigma_n^2) \right] = 0$$

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N 2(t_n - \phi(x_n)\omega^\top)(-\phi(x_n)) = 0$$

$$\frac{1}{\sigma^2} (t_n - \phi(x_n)\omega^\top) \phi(x_n) = 0$$

In form matricial

$$\frac{1}{\sigma^2} (t_n - \phi\omega^\top) \phi^\top = 0$$

$$\phi^T(t - \phi w) = 0$$

$$\phi^T t - \phi^T \phi w = 0$$

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

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

## • Mètode A-posteriori (MAP)

CST modifica laica estimar los parámetros probabilísticos  
en la incorporación de información previa antes de  
observar datos en estadística bayesiana, diciendo típicamente el  
teorema de Bayes

$$\text{Posterior} \rightarrow P(x|y) = \frac{f(y|x) f(x)}{f(y)} \xrightarrow{\substack{\text{verosimilitud} \\ \rightarrow \text{prior}}} \text{evidencia}$$

There is another notation

$$P(w|t) = \frac{P(t|w)p(w)}{P(t)}$$

en particular

$$P(t) = \int p(t|w)p(w)dw$$

- en la prior

$$p(w) = \prod_{n=1}^N N(w|0, \sigma_w^2)$$

- verosimilitud

$$p(t_n|\phi(x_n)w, \sigma_h^2) = N(t_n | \phi(x_n)w, \sigma_h^2)$$

$$p(t|w, \sigma_h^2) = \prod_{n=1}^N N(t_n | \phi(x_n)w, \sigma_h^2) \rightarrow \text{MAP}$$

Per lo que el modelo simplifica la proporcionalidad de Bayes  
mediante la proporcionalidad

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

Ahora para determinar la función de costo

$$\mathcal{L}(A, f(x)) = P(w|t, \Phi, \sigma_n^2)$$

$$= \prod_{n=1}^N N(t_n | \phi(x_n)w, \sigma_n^2) \prod_{j=1}^Q N(w_j | 0, \sigma_w^2)$$

$$\log(\mathcal{L}(A, f(x))) = \log\left(\prod_{n=1}^N N(t_n | \phi(x_n)w, \sigma_n^2)\right) + \log\left(\prod_{j=1}^Q N(w_j | 0, \sigma_w^2)\right)$$

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

$$\log\left(\prod_{j=1}^Q \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{\|w_j - 0\|^2}{2\sigma_w^2}\right)\right)$$

$$= \frac{N}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{n=1}^N \|t_n - \phi(x_n)w\|^2 + \log\left(\frac{1}{\sqrt{2\pi\sigma_w^2}}\right)^Q +$$

$$\log\left(\exp\left(-\frac{1}{2\sigma_w^2} \sum_{j=1}^Q \|w_j\|^2\right)\right)$$

$$= -\frac{N}{2} \log(2\pi\sigma_0^2) - \frac{Q}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_0^2} \|t - \phi(w)\|^2 - \frac{1}{2\sigma_w^2} \sum_{j=1}^Q \|w_j\|^2$$

Ahora para fines de optimización se ignora los términos que no dependen de  $w$

el problema queda

$$w_{MAP} = \underset{w}{\operatorname{argmax}} - \frac{1}{2G^2} \|t - \Phi w^\top\|_2^2 - \frac{1}{2G_w^2} \|w\|_2^2$$

(sin los factores de escala en multo el punto maximo o minimo el problema se transforma)

$$w_{MAP} = \underset{w}{\operatorname{argmin}} \|t - \Phi w^\top\|_2^2 + \frac{G_w^2}{2} \|w\|_2^2$$

en un orden de ideas

el problema se reduce a una optimización por minimizar la función regularizada con  $\lambda = \frac{G_w^2}{2}$ , el cual se resuelve alternativamente y la solución es:  $G_w^\dagger$

$$\boxed{w_{MAP}^* = (\Phi^\top \Phi + \frac{G_w^2}{2} I)^{-1} \Phi^\top t}$$

• Bayesiano con modelo lineal gaussiano

considérese denas modelos SL regulares incorporar conocimiento previo en el prior

tendrá la forma

$$P(t|w) = \mathcal{N}(t|\Phi w, \sigma_0^2)$$

y el prior

$$P(w) = \mathcal{N}(w|0, G_w^{-2})$$

→ die Distribution der prior

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

→ Asumiert el prior

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

→ y ist posterior (umw)

$$p(w|t) = N(w|m_N, S_N)$$

→ S1  $p(x) = N(x|\mu, \Delta^{-1})$

$$p(y|x) = N(y|Ax+b|L^{-1})$$

$$\text{mit } y = f(x|A, b) = Ax + b$$

$$\text{Irgo } p(y|x) = N(x|\mu_{x|y}, \Sigma_{x|y})$$

$$\mu_{x|y} = (\Lambda + A^T L A)^{-1} [A^T L (y - b) + \Lambda \mu]$$

$$\Sigma_{x|y} = (\Lambda + A^T L A)^{-1}$$

$$\text{in 2. Fall } t_n = w^T \phi(x_n) + b = N(t_n | w^T \phi(x_n), \beta^{-1})$$

$$\hookrightarrow p(t|w) = N(t|dw, \tilde{\beta}^{-1} I)$$

$$\text{Irgo } w \sim p(w) = N(w|m_0, S_0)$$

$$\text{el posterior } p(w|t) = N(w|m_0, S_0)$$

$$\mu_N = \mu_{w|t} = (S_0^{-1} + \phi^T \beta I \phi)^{-1} [\phi^T \beta I (t - c) + S_0^{-1} m_0]$$

$$\Sigma_N = \Sigma_{w|t} = (S_0^{-1} + \beta \phi \phi^T)^{-1} [\beta \phi^T + S_0^{-1} m_0]$$

$$\mu_N = \mu_{w|t} = S_N (S_0^{-1} m_0 + \beta \phi^T t)$$

lugo

$$S_N = (S_0^{-1} + \beta \Phi^T \Phi)^{-1}$$

$$S_N^{-1} = S_0^{-1} = S_0^{-1} + \beta \Phi^T \Phi$$

$$\text{Ahiere } \beta = 1/\sigma_n^2$$

la media posterior

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

covarianza posterior

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

(una suposición)  $p(w|t) = N(w|\bar{w}_N, \tilde{S}_N)$   
→ Suposición  $m_0 = 0$

$$m_N = S_N \left( \frac{1}{\sigma_n^2} S_0 \Phi \Phi^T \right) \quad (*)$$

$$\rightarrow S_0 = \sigma_w^2 I_Q$$

$$S_N = \left( (16\omega^2 I_Q)^{-1} + \frac{1}{\sigma_n^2} \Phi \Phi^T \Phi \Phi \right)^{-1}$$

$$\tilde{S}_N = \left( \frac{1}{\sigma_n^2} I_Q + \frac{1}{\sigma_h^2} \Phi \Phi^T \Phi \Phi \right)^{-1}$$

$$\tilde{S}_N = \left[ \frac{1}{\sigma_n^2} \left( \frac{\sigma_h^2}{\sigma_0^2} I_Q + \Phi \Phi^T \Phi \Phi \right) \right]^{-1} = \sigma_n^2 \left( \frac{\sigma_h^2}{\sigma_0^2} I_Q + \Phi \Phi^T \Phi \Phi \right)^{-1} \quad (*)$$

$w \sim \mathcal{N}$

(\*) (\*) (\*)

$$\tilde{m}_N = \frac{1}{\omega^2} \cancel{\frac{6^2}{2}} \left( \frac{6^2}{6^2} I_N + \Phi \Phi^T \Phi \Phi \right)^{-1} \Phi \Phi^T +$$

$$\boxed{\tilde{m}_N = \left( \frac{6^2}{6^2} I_N + \Phi \Phi^T \right)^{-1} \Phi \Phi^T +}$$

Per lo que la solución del modelo es equivalente a  $\tilde{m}_N$  a la solución de máximos cuadrados regulares teniendo

$$\boxed{\lambda = \frac{6^2}{6^2}}$$

## Regresion Ridge Kernel

A diferencia de la regresión lineal clásica, PRK incluye en su marco la regularización que penaliza la magnitud de los coeficientes para evitar el sobreajuste.

Definimos

$$y \in \mathbb{R}^N; X \in \mathbb{R}^{N \times P} \quad H \subseteq \mathbb{R}^Q \quad Q \rightarrow \infty$$

$$\hat{y} = f(\phi(X)|\omega) = \phi(x)\omega$$

$$f: \mathbb{R}^Q \rightarrow \mathbb{R} \quad \omega \in H \subseteq \mathbb{R}^Q$$

el problema se reduce a:

$$\omega^* = \underset{\omega}{\operatorname{argmin}} J(y, f(\phi(x)|\omega)) + R(f|\lambda)$$

consistiendo minimizar el error cuadrático promedio de regularización, similar al problema de mínimos cuadrados regulados  $J = \|y - \phi(x)\omega\|_2^2 \quad R(f|\lambda) = \lambda \|\omega\|^2$

entonces

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \frac{1}{N} \|y - \phi(x)\omega\|_2^2 + \lambda \|\omega\|^2 \quad (1)$$

dnde  $\lambda \in \mathbb{R}^+$  (fijo)

$$\|y - \phi(x)\omega\|^2 = (y - \phi(x)\omega)(y - \phi(x)\omega)^T$$

$$yy^T - 2y^T \phi(x)\omega + ((\phi(x)\omega)^T (\phi(x)\omega))$$

$$yy^T - 2y^T \phi(x)\omega + \omega^T \phi^T \phi \omega \quad (2)$$

(2) en (1)

desarrollando la función de costo

$$\frac{\partial}{\partial w} \{ L(y, f) + R(f | \lambda) \} =$$

$$\frac{1}{N} \frac{\partial}{\partial w} \left[ y^T y - 2 \phi(x)^T y + w^T \phi^T(x) \phi(x) w \right] + \frac{\partial}{\partial w} [\lambda w^T w]$$

$$\frac{1}{N} \left[ -2 \phi^T(x) y + 2 w^T \phi^T(x) \phi(x) + 2 \lambda w \right] = 0$$

$$\frac{2 w^T \phi(x) \phi(x)}{N} + 2 \lambda w = \frac{2}{N} \phi^T(x) y$$

$$w^T \phi^T(x) \phi(x) + N \lambda w = \frac{2}{N} \phi^T(x) y$$

$$w [ \phi^T(x) \phi(x) + N \lambda I ] = \frac{2}{N} \phi^T(x) y$$

$$w^* = \boxed{[\phi^T(x) \phi(x) + N \lambda I]^{-1} \phi^T(x) y}$$

Ahora la dimensión del espacio transformado es muy grande ( $\infty$ )  
lo que hace que calcular la inversa sea un poco  
completo y costoso o sea imposible: Por lo que debemos  
pensar en  $[\phi^T(x) \phi(x) + N \lambda I]^{-1}$  usando la identidad

$$(I + AB)^{-1} A = A (I + BA)^{-1}$$

$$\text{en la que simplificamos } (\phi^T(x) \phi(x) + N \lambda I)^{-1} \phi^T(x)$$

Al reescribir:  $(I + A^{-1}B)^{-1} A = A (I + B A^{-1})^{-1}$

$$(\phi^T(x) \phi(x) + N \lambda I)^{-1} \phi^T(x) = \left[ N \lambda \left( \frac{1}{N \lambda} (\phi^T \phi + I) \right) \right]^{-1} \phi^T$$

$$= \frac{1}{N \lambda} \left( I + \frac{\phi^T \phi}{N \lambda} \right)^{-1} \phi^T = \phi^T \frac{1}{N \lambda} \left( I + \frac{\phi}{N} \phi^T \right)^{-1}$$

$$= \phi^T \left[ N\lambda (I + \frac{1}{N\lambda} \phi \phi^T) \right]^{-1} = \phi^T [N\lambda I + \phi \phi^T]$$

$\boxed{\omega^* = \phi^T(x) [N\lambda + \phi(x) \phi^T(x)]^{-1} y}$

Ahora la predicción de un nuevo punto

$$\hat{y} = f(\phi(x_{new}) | \omega) = \phi^T(x_{new}) \omega$$

$$\phi^T(x_{new}) \phi^T(x) [N\lambda I + \phi(x) \phi^T(x)]^{-1} y$$

$$\phi(x_{new}) \in \mathbb{R}^Q$$

Ahora por kernel Trick

$$K = \phi(x) \phi(x)^T \quad K_{ij} = K(x_i, x_j)$$

y el de la nueva predicción

$$k^T_{new} = [k(x_{new}, x_n)]_{n=1}^N$$

$$k^T_{new} = \phi(x) \phi(x)^T$$

$\hat{y} = f(x_{new}) = k^T_{new} (N\lambda I + K)^{-1} y$

• Pruebas gaussianas

este modelo basa las predicciones en los parametros, el medidor (que es los puntajes)

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

→ Asumir que el prior es de forma gaussiana

$$p(w) = N(w|w_0, \sigma_w^2)$$

→ considerar  $\bar{m} = 0$   $S_0 = \sigma_w^2 I_d$  interc.  $p(w)$

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

→ Ahora la distribución del posterior

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

→ verosimilitud

$$p(t|w) = N(t|\Phi w, \sigma_h^2)$$

y la kernel

$$\log p(t|w) = -\frac{N}{2} \log(2\pi\sigma_h^2) - \frac{1}{2\sigma_h^2} (t - \Phi w)^T (t - \Phi w)$$

per lo que el prior

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

$$\log p(w) = -\frac{N}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} w^T w$$

Ahora sumando los exponentes

$$\log p(w|t) = \log p(t|w) + \log p(w)$$

$$\log p(w|t) \propto -\frac{1}{2\sigma_n^2} (t - \Phi w)^T (t - \Phi w) - \frac{1}{2\sigma_w^2} w^T w$$

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

$$\log p(w|t) \propto -\frac{1}{2\sigma_n^2} (t^T t - 2t^T \Phi w + w^T \Phi^T \Phi w) - \frac{1}{2\sigma_w^2} w^T w$$

$$\log p(w|t) \propto -\frac{1}{2\sigma_n^2} t^T t + \frac{1}{\sigma_n^2} t^T \Phi w + \frac{1}{2\sigma_n^2} w^T \Phi^T \Phi w - \frac{1}{\sigma_w^2} w^T w$$

$$\log p(w|t) \propto -\frac{1}{2} w^T \left( \frac{1}{\sigma_n^2} \Phi^T \Phi + \frac{1}{\sigma_w^2} I_Q \right) w + \frac{1}{\sigma_n^2} w^T \Phi^T t$$

$$\text{hence } \tilde{S}_N^{-1} = \frac{1}{\sigma_n^2} \Phi^T \Phi + \frac{1}{\sigma_w^2} I_Q$$

$$S_N^{-1} \tilde{m}_N = \frac{1}{\sigma_n^2} \Phi^T t$$

$$\tilde{m}_N = S_N \left( \frac{1}{\sigma_n^2} \Phi^T t \right) = \left( \frac{1}{\sigma_n^2} \Phi^T \Phi + \frac{1}{\sigma_w^2} I_Q \right)^{-1} \frac{1}{\sigma_n^2} \Phi^T t$$

Above is a criteria MAP problem

$$w_{MAP} = \arg \max_w p(w|t) = \arg \max_w (\log p(t|w) + \log p(w))$$

$$w^* = \tilde{m}_N$$

$$f(x) \sim N(\mu_f, \sigma_f^2)$$

quando calcolerò la distribuzione predittiva di  $f^*$

$$p(f^* | x^*, D) = \int p(f^* | x^*, w) p(w | D) dw$$

per tanto da: integral optimus

$$p(f^* | x^*, D) = N(f^* | \phi(x^*)^T \tilde{m}_N, \phi(x^*)^T \tilde{S}_N^{-1} \phi(x^*) + \sigma_n^2)$$

$$\mu_f = \phi(x^*)^T \tilde{m}_N$$

$$\sigma_f^2 = \phi(x^*)^T \tilde{S}_N^{-1} \phi(x^*) + \sigma_n^2$$

Allora spesso la prior sarà la funzione  $f(x)$

$$f(x) \sim N(0, k(x, x))$$

per cui si considera

$$p(y | D) \sim N(0, k + \sigma_n^2 I)$$

$$\begin{pmatrix} y \\ f^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k + \sigma_n^2 I & k^* \\ k^T & k(x^*, x^*) \end{pmatrix}\right)$$

Allora usiamo le proprietà di gaussiane multivariate

$$\begin{pmatrix} v \\ u \end{pmatrix} \sim N\left(\begin{pmatrix} m_u \\ m_v \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AC} \\ \Sigma_{CA} & \Sigma_{CC} \end{pmatrix}\right)$$

$$p(u | v) \sim N(m_u + \Sigma_{AC} \Sigma_{CC}^{-1} (v - m_v), \Sigma_{AA} - \Sigma_{AC} \Sigma_{CC}^{-1} \Sigma_{CA})$$

Allora se mostro le SO

$$p(f^* | x^*, D) \sim N(\mu_f + k^T (K + \sigma_n^2 I)^{-1} (x^* - x), k(x^*, x^*) - k^T (K + \sigma_n^2 I)^{-1} k)$$

Simplifying

$$p(f^* | y, x^*) \sim N(k_x^T (k + \sigma_n^2 I)^{-1} +, k(x^*, x^*) - k_x^T (k + \sigma_n^2 I)^{-1} k_x)$$

After taking express in terms

$$\mu_f = k_x^T (k + \sigma_n^2 I)^{-1} y$$

$$\sigma_f^2 = k(x^*, x^*) - k_x^T (k + \sigma_n^2 I)^{-1} k_x$$

## DISCUSIÓN

### Disección

- Mínimos cuadrados: los modelos estadísticos se basan en este para veran significativamente la curva de menor  $w$  y la incorporación de priors y su regularización

Solución:

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

Si  $(\phi^T \phi)$  es invertible

- Mínimos cuadrados regularizados

Este modelo incluye un término de regularización  $\lambda \|L^2\|$  con el objetivo de prevenir sobreajuste ( $\lambda > 0$ ) y mejorar la generalización del mismo

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

- Máxima Verosimilitud

Es coincidente con el modelo de regresión lineal en su solución

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

Permite integrar la estimación de parámetros de regresión lineal dentro de los procedimientos que maximizan la verosimilitud bayesiana de observar los parámetros dadas las observaciones

- MAP

Es un caso de mínimos cuadrados regularizados donde  $\lambda = \frac{\sigma_0^2}{\sigma_w^2}$  en la que se incorpora información previa o a priori

$$w^* = (\phi^T \phi + \lambda I)^{-1} \phi^T t \quad \lambda = \frac{\sigma_0^2}{\sigma_w^2}$$

## Capítulo 4

→ Bayesiane con modelo lineal gaussiano

el objetivo es inferir la distribución completa de los parámetros y el sistema nos da la  $\pi(\theta)$  (mitad posterior) y la varianza posterior  $\Sigma_{\theta\theta}$  y los ~~entendemos~~ coinciden con el modelo MAP ya que se solucionan con la diferencia de que el infijo bayesiano proporciona una distribución posterior mientras que MAP, ofrece una estimación puntual.

→ kernel ridge regression

el objetivo es extender la regresión Ridge a un espacio de características definido por un kernel, Al igual que similar a procesos gaussianos, kernel Ridge proporciona una estimación puntual de  $w$ .

→ procesos gaussianos

el objetivo es definir una distribución de funciones (funciones  $f(x)$ ), aunque similar a Kernel Ridge, este proporciona una distribución completa de  $f^*$  junto con su varianza.