/ /

PARCIAL #1

## Preguntas

1. Sea el modelo de regresión $t_n = \phi(\mathbf{x}_n)\mathbf{w}^\top + \eta_n$, con $\{t_n \in \mathbb{R}, \mathbf{x}_n \in \mathbb{R}^P\}_{n=1}^N$, $\mathbf{w} \in \mathbb{R}^Q$, $\phi : \mathbb{R}^P \to \mathbb{R}^Q$, $Q \geq P$, 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 verosilimitud, 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) SEA EL MODELO DE REGRESIÓN

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

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

DONDE $\{t_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N$, $w \in \mathbb{R}^Q$ Y $\phi : \mathbb{R}^P \to \mathbb{R}^Q$ ES UN VECTOR DE FUNCIÓN BASE $(Q \geq P)$

EN DONDE DEFINIMOS:

$$t = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, \qquad \phi = \begin{pmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_N)^T \end{pmatrix} \in \mathbb{R}^{N \times Q}$$

DE MANERA QUE EL MODELO COMPLETO SE PUEDE ESCRIBIR COMO:

$$t = \Phi w + \eta, \qquad \therefore \eta \sim N(0, \sigma_\eta^2 I_N)$$

AHORA PARA LOS SIGUIENTES:

* MÍNIMOS CUADRADOS
* MÍNIMOS CUADRADO REGULARIZADOS
* MÁXIMA VEROSIMILITUD
* MÁXIMA A-POSTERIORI
* BAYESIANO CON MODELO LINEAL GAUSSIANO
* REGRESIÓN RIGIDA KERNEL
* PROCESOS GAUSSIANOS

→ MÍNIMOS CUADRADOS:

ASUMIMOS QUE LAS OBSERVACIONES $t_n$ ESTÁN RELACIONADAS CON $x_n$, CON EL MODELO:

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

EN DONDE:

* $\Phi(x_n)^T w$ → ES LA PREDICCIÓN PARA LA ENTRADA $X_n$
* $W$ SON LOS PARAMETROS
* $\eta_n$ ES EL RUIDO BLANCO GAUSSIANO

Suponemos las observaciones $t_n$ están distribuidas como:

$$P(t_n \mid x_n, w) = N(t_n \mid \phi(x_n)^T w, \sigma^2_n)$$

→ Para $t_n$ la probabilidad conjunta es:

$$P(t \mid \phi, w) = \prod_{n=1}^{N} P(t_n \mid x_n, w)$$

En donde $\phi$ es la matriz de características

→ El error $e_n$ (Cuanto x desvia la predicción)

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

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

Ahora para plantear el problema, la distribución gaussiana de $t_n$ y $e_n$

$$P(t \mid \phi, w) = \frac{1}{(2\pi\sigma^2_n)^{N/2}} \exp\left[-\frac{1}{2\sigma^2_n} \|t_n - \phi w\|^2\right]$$

Donde:

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

Nos interesa minimizar la perdida promedio del error entonces:

asumiendo i.i.d $= \mathcal{E}\{x_n\} \approx \frac{1}{N} \sum_{n=1}^{N} x_n$

Debemos minimizar w de $\|t - \phi w\|^2$:

$$w^* = \arg\min \mathcal{E}\{e_n\}$$

$$w^* = \arg\min \mathcal{E}\{\|t - \phi w\|^2\}$$

$$\mathcal{E}\{\|t - \phi w\|\} = \frac{1}{N}\left[\|t - \phi w\|^2\right]$$

$$= \frac{1}{N}(t - \phi w)(t - \phi w) = \left[t^T t - t^T \phi w - \phi w^T + (\phi w)^T Q w\right] \frac{1}{N}$$

$$= \left[t^T t - 2t^T \phi w + w^T \phi^T \phi w\right] \frac{1}{N}$$

Luego tenemos que:

$$\frac{\partial}{\partial w} \mathcal{E}\{\|t - \phi w\|\} = 0$$

$$\frac{\partial}{\partial w} \mathcal{E}\{\|t - \phi w\|^2\} = -2(t^T \phi)^T + 2\phi^T \phi w = 0 \cdot N = 0$$

$$2\phi^T \phi w = 2(t^T \phi)^T$$

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

↳ Ahora con la pseudoinversa: $\boxed{w^* = (\phi^T \phi w)^{-1} \phi t}$

/ /

→ Mínimos cuadrados regularizados

Teniendo el modelo:

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

Y el error cuadratico

$$e_n^2 = \| t - w\phi \|_2^2$$

Debido a que mínimos cuadrados es un modelo sensible en datos atipicos, es necesario un termino de regularización.

La regularización $L^2$ se define como:

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

Luego el termino de regularización es:

$$\lambda \|w\|_2^2$$

Por lo que tenemos:

$$\| t - \phi^T w \|_2^2 + \lambda \|w\|_2^2$$

Ahora el problema de optimización es:

$$w = \underset{w}{\mathrm{Arg\,min}} \; \| t - \phi w \|_2^2 + \|w\|_2^2 \lambda$$

De mínimos cuadrados tenemos

$$\| t - \phi w \|_2^2 = (t - \phi w)^T (t - \phi w)$$

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

$$w^* \mathrm{argmin} \left[ tt^T - 2t^T \phi w + w^T \phi^T \phi w + w^T w \lambda \right] = 0$$

$$\frac{d}{dw} \left[ tt^T - 2t^T \phi w + (w\phi)^T \phi w + w^T w \lambda \right] = 0$$

$$\frac{d}{dw} \left[ tt^T - 2t^T \phi w + w^T (\phi^T \phi + \lambda I) w \right] = 0$$

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

$$-2\phi^T t + 2(\phi^T \phi + \lambda I) w = 0$$

$$2(\phi^T \phi + \lambda I) w = \phi^T t$$

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

/ /

## → Máxima verosimilitud

Se tiene el modelo de regresión

$$t_n = \phi(x_n)w^T + \eta_n \qquad \therefore \eta_n \sim N(\eta_n \mid 0, \sigma_n)$$

→ Ruido blanco gaussiano y datos i.i.d (suponemos)

Entonces

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

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

$$P(t_n \mid \phi(x_n)w^T, \sigma_n^2) = N(t_n \mid \phi(x_n \mid w^T, \sigma_n^2)$$

Teniendo el supuesto de i.i.d la verosimilitud conjunta

$$p(t \mid \mathbb{X}, w, \sigma_n^2) = \prod N(t_n \mid \phi(x_n)w^T, \sigma_n^2)$$

Ahora como asumimos una normal tenemos $\log(N(t_n)\phi(x_n)w^T, \sigma_n)$

$$N(t \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(t_n - \phi(x_n)w^T)^2}{2\sigma_n^2}\right]$$

Entonces la log-verosimilitud es la suma de las log-verosimilitudes individuales

$$\log(P(x)) = \log\left(\prod_{n=1} N(x_n \mid \mu, \sigma^2)\right)$$

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

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

Luego vemos que:

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

Podemos escribir el argumento del exponente como la sumatoria de las muestras

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

Luego tenemos que:

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

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

$$w_{ML} = \underset{w, \sigma_L^2}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} N(t_n \mid \phi(x_n)w_T, \sigma_n^2)\right)$$

/ /

Resolviendo y usando log - verosimilitud

$$w_{ml} \mid \sigma^2_{ml} = \underset{w,\,\sigma^2_n}{\arg\max} \left[ -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\sigma^2_n) - \frac{1}{2\sigma^2} \|t - \phi w^T\| \right.$$

Ahora derivando respecto a $\sigma^2$

$$\frac{d}{d\sigma^2} \log P(t \mid x, u, \sigma^2_n) = \left( \frac{\partial}{\partial \sigma^2} \left[ -\frac{N}{2} \log(2\pi) \right] - \frac{\partial}{\partial \sigma^2} \left[ \frac{N}{2} \log(\sigma^2_n) \right] \right.$$

$$\frac{d}{d\sigma^2} \left[ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} \|t - \phi w^T\|^2 \right]$$

Resolviendo

$$\frac{d}{d\sigma^2} p(t \mid w, \sigma^2) = \frac{-N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^{N} \|t_n - \phi(x_n) w^T\|^2 = 0$$

$$-N\sigma_n^2 + \sum \|t_n - \phi(x_n) w^T\|^2 = 0$$

$$\sigma^2_{ml} = \frac{1}{N} \sum \|t_n - \phi(x_n) w^T\|^2$$

Derivando con respecto a $w$:

$$\frac{d}{dw} \left[ \log p(t \mid x, w, \sigma^2_n) \right] = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} 2 \left( t_n - \phi(x_n) w^T \right) \left( -\phi(x_n) \right)$$

$$\frac{d}{dw} \left[ \log p(t \mid x, w, \sigma^2_n) \right] = 0$$

$$\frac{-1}{\cancel{2}\sigma^2} \sum_{n=1}^{N} \cancel{2} \left( t_n - \phi(x_n) w^T \right) \left( -\phi(x_n) \right) = 0$$

$$\boxed{\frac{1}{\sigma^2} \left( t_n - \phi(x_n) W^T \right) \phi(x_n) = 0}$$

De forma matricial

$$\frac{1}{\sigma^2} \left( t_n - \phi w^T \right) \phi^T = 0$$

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

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

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

$$\boxed{w^* = \left( \phi^T \phi \right)^{-1} \phi^T w}$$

/ /

→ Máximo a-posteriori (MAP)

Este modelo busca estimar los parametros probabilisticos con la interpretación de información previa antes de observar datos en estadistica bayesiana, de aquí tenemos el teorema de Bayes

$$\underbrace{P(x,y)}_{\text{Posterior}} = \frac{\overbrace{f(y|x)}^{\text{Verosimilitud}}\ \overbrace{f(x)}^{\text{Prior}}}{\underbrace{f(y)}_{\text{Evidencia}}}$$

Entonces

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

* Con evidencia

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

* Con prior

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

* Verosimilitud

$$P(t_n|\phi(x_n)w, \sigma_n^2) = N(t_n|m(x_n)w, \sigma_n^2)$$

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

Por lo que el modelo simplifica la aplicación de Bayes mediante la proporcionalidad

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

Ahora para determinar la función de costo

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

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

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

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

$$= \log\left(\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left(-\frac{\|t_n - \phi(x_n)w\|^2}{2\sigma_n^2}\right)\right) + \log\left(\prod_{j=1}^{Q}\frac{1}{\sqrt{2\pi\sigma_n^2}}\right)\exp\left(-\frac{\|w_j - 0\|_2^2}{2\sigma_w^2}\right)$$

$$= \frac{N}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\sum_{n=1}^{N}\|t_n - \phi(x)w\|^2 + \log\left(\frac{1}{\sqrt{2\pi\sigma_w^2}}\phi\right) + \log\left(\exp\left(\frac{-1}{2\sigma_w^2}\sum_{j=1}^{Q}\|w_j\|^2\right)\right)$$

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

/ /

Ahora para fines de optimización se ignoran los terminos que no dependen de $w$, el problema queda

$$w_{MAP} = \underset{w}{\arg\max} \ \frac{-1}{2\sigma_n^2} \|t - \phi w^T\|_2^2 - \frac{1}{2\sigma_w^2} \|w\|_2^2$$

Como los factores de escala no modifican el punto máximo o minimo, el problema se transforma

$$w_{MAP} = \arg\min \|t - \phi w^T\|_2^2 + \frac{\sigma_n^2}{\sigma_w^2} \|w\|_2^2$$

En otras palabras, el problema se reduce a una optimización por minimos cuadrados con $\lambda = \frac{\sigma_n^2}{\sigma_w^2}$, el cual se resolvió anteriormente y queda:

$$w^*_{map} = \left( \phi^T\phi + \frac{\sigma_n^2}{\sigma_w^2} I \right)^{-1} \phi^T t$$

→ Bayesiano con modelo lineal Gaussiano

Como los demas modelos se requiere incorporar conocimiento previo en el prior

Tenemos la verosimilitud

$$p(t|w) = \mathcal{N}(t \mid \phi w \ \sigma_n^2)$$

y el prior

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

La distribución del prior

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

Asumimos el prior

$$p(w) = \mathcal{N}(w_0 | S_0)$$

y el posterior como

$$p(w|t) = \mathcal{N}(w \mid m_N, S_N)$$

Si $P(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$

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

Con $y = f(x|A,b) = Ax + b$

Luego $p(y|x) = \mathcal{N}(x \mid \mu x|y, \Sigma_{xy})$

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

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

/ /

En el caso $t_n = w^T\phi(x_n) t_n = N(t_n | w^T\phi(x_n), \beta^{-1})$

$$L \quad p(t|w) = N(t|\phi w, \beta^{-1} I)$$

Luego $w \sim p(w) = N(w|m_0, S_0)$

El posterior $P(w|t) = N(w|m_0, S_0)$

$$M_n = M_w|t = (S_0^{-1} + \phi^T \beta I \phi)^{-1} (\phi^T \beta I (t - 0) + S_0^{-1} m_0]$$

$$M_n = M_w|t = (S_0^{-1} + \beta \phi^T \phi)^{-1} (\beta \phi^T t + S_0^{-1} m_0)$$

$$M_N = M_n|t = S_N (S_0^{-1} m_0 + \beta \phi^T t)$$

Luego se tiene

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

$$S_N^{-1} = S_{0N}^{-1} = S_0^{-1} + \beta \phi^T \phi$$

Ahora $\beta = \frac{1}{\sigma_n^2}$

La media posterior

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

La covarianza posterior

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

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

Suponemos $m_0 = 0$

$$m_n = S_n \left( \frac{1}{\sigma_n^2} S_N \phi \phi^T t \right) \quad (*)(*)$$

$S_0 = \sigma_w^2 I_Q$

$$S_N = \left( (\sigma_w^2 I_Q)^{-1} + \frac{1}{\sigma_n^2} \phi \phi^T \phi \phi \right)^{-1}$$

$$\tilde{S}_n = \left( \frac{1}{\sigma_w^2} I_Q + \frac{1}{\sigma_n^2} \phi \phi^T \phi \phi \right)^{-1}$$

$$\tilde{S}_n = \left[ \frac{1}{\sigma_n^2} \left( \frac{\sigma_n^2}{\sigma_w^2} I_Q + \phi \phi^T \phi \phi \right) \right]^{-1} = \sigma_n^2 \left( \frac{\sigma_n^2}{\sigma_w^2} I_Q + \phi \phi^T \phi \phi \right)^{-1} \quad (*)$$

Ahora $(*)$ en $(*)(*)$

$$\bar{m}_n = \frac{1}{\sigma_n^2} \sigma_n^2 \left( \frac{\sigma_n^2}{\sigma_w^2} I\phi + \phi\phi^T\phi\phi \right)^{-1} \phi\phi^T t$$

$$\boxed{\bar{m}_n = \left( \frac{\sigma_n^2}{\sigma_w^2} I\phi + \phi\phi^T \right)^{-1} \phi\phi^T t}$$

Por lo que la solución del modelo es equivalente de $\bar{m}_n$ a la solución de mínimos cuadrados regularizados, teniendo

$$\boxed{\lambda = \frac{\sigma_n^2}{\sigma_w^2}}$$

→ Regresión Rígida Kernel

A diferencia de la regresión lineal clásica, RRK incluye un término de regularización que penaliza la magnitud de los coeficientes para evitar el sobreajuste

Tenemos

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

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

$$f : \mathbb{R}^Q \to \mathbb{R} \qquad w \in H \subseteq \mathbb{R}^Q$$

El problema se resume en:

$$w^* = \arg\min_w \frac{1}{N} (y, f(\phi(x)|w)) + R(t|\lambda)$$

Consiste en minimizar el error más un parámetro de regularización, similar al problema de mínimos cuadrados regularizado)

$$e = \|y - \phi(x) w\|^2 \qquad R(t|\lambda) = \lambda \|w\|^2$$

Entonces

$$w^* = \arg\min_w \frac{1}{N} \|y - \phi(x) w\|_2^2 + \lambda \|w\|^2 \qquad (i)$$

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

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

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

$$yy^T - 2y\,\phi(x) w + w^T\phi^T\phi w \qquad (ii)$$

/ /

Ahora haciendo (i) en (ii), derivando la función de costo

$$\frac{d}{dw}\left\{ L(y,f) + R(f\mid\lambda)\right\} = \frac{1}{N}\frac{d}{dw}\left[ y^T y - 2\phi(x) y + w^T\phi^T(x)\phi(x) w\right] + \frac{d}{dw}\left[\lambda w^T w\right]$$

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

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

$$w^T\phi^T(x)\phi(x) + N\lambda w = \phi^T(x) y$$

$$w\left[\phi^T(x)\phi(x) + N\lambda I\right] = \phi^T(x) y$$

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

Ahora la dimensión del espacio transformado es demasiado grande $(\infty)$, lo que hace que calcular la inversa cueste mucho computacionalmente o sea imposible, por lo que se debe reescribir $[\phi^T\phi(x) + N\lambda I]$ usando la identidad:

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

Con lo que simplificamos $(\phi(x)^T\phi(x) + N\lambda I)^{-1}\phi^T(x)$, ahora usando la identidad:

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

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

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

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

$$\phi(x_n e w) \in \mathbb{R}^Q$$

Ahora por Kernel trick

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

y el de la nueva predicción

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

$$K^T_{new} = \phi(x)\phi(x_n)$$

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

/ /

## → Procesos Gaussianos

Este modelo busca incorporar a los parametros del modelo como en las predicciones

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

Asumimos el prior de forma Gaussiana

$$p(w) = N(w \mid m_0, S_0)$$

Consideramos $m_0 = 0$, $S_0 = \sigma_w^2 I_Q$ entonces $p(w)$

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

Ahora la distribución del posterior

$$p(w \mid t) \propto p(t \mid w)\, p(w)$$

La verosimilitud

$$p(t \mid w) = N(t \mid \phi w, \sigma_w^2)$$

y la log-verosimilitud

$$\log p(t \mid w) = \frac{-N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} (t - \phi w)^T (t - \phi w)$$

Por lo que el prior

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

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

Ahora sumando las expresiones

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

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

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

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

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

$$\log p(w \mid t) \propto -\frac{1}{2} w^T \left( \frac{1}{\sigma_n^2} \phi^T \phi + \frac{1}{\sigma_w^2} I \phi \right) + \frac{1}{\sigma_n^2} w^T \phi^T t$$

/ /

Luego $\bar{S}_n^{-1} = \frac{1}{\sigma_n^2} \phi^T \phi - \frac{1}{\sigma_w^2} I \phi$

$$S_n^{-1} m_N = \frac{1}{\sigma_n^2} \phi^T t$$

Ahora, con el criterio MAP, el problema es:

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

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

Queremos calcular la distribución predictiva de $f^*$

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

Por tanto, de la integral obtenemos:

$$p(t^*/x^*, D) = N(t^* \mid \phi(x^*)^T m_N, \ \phi(x^*)^T S_N \phi(x^*) + \sigma_n^2)$$

$$\mu t^* = \phi(x^*)^T m_N$$

$$\sigma_{t^*}^2 = \phi(x^*)^T S_N \phi(x^*) + \sigma_n^2$$

Ahora expresamos el prior sobre la función $f(x)$

$$f(x) \sim N(0, K(x, x'))$$

Podemos considerar

$$p(y|0) \sim N(0, K + \sigma_n^2)$$

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

Ahora usamos las propiedades gaussianas multivariadas

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

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

Aplicando esto a nuestro caso

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

Simplificando

$$p(t^*|y^*, x^*) \sim N\left(K_*^T (K + \sigma_n^2 I)^{-1} y, \ K(x^*, x^*) - K_*^T (K + \sigma_n^2 I)^{-1} K_*\right)$$

Donde las expresiones son:

$$\boxed{\begin{aligned} \mu_* &= k_*^T (K + \sigma_n^2 I)^{-1} y \quad \rightarrow \text{Predicción} \\ \sigma_*^2 &= k(x^*, x^*) - k_*^T (K + \sigma_n^2 I)^{-1} k_* \quad \rightarrow \text{Incertidumbre} \end{aligned}}$$

| NOMBRE | Problema Optimización | Solución analítica | Explicación |
|---|---|---|---|
| LS - Least Squares | $W_{LS} = \underset{w}{\text{argmin}}\ RSS(w) = \underset{w}{\text{argmin}} \sum_{n=1}^{N} (t_n - \phi(X_n).W^T)^2$ | Invertible<br>$(\phi^T \phi)^{-1} \phi^T . t = W$ | Busca minimizar la distancia entre los datos observados y el dato predicho |
| RLS- Regularized Least Squares | $\underset{w}{\text{argmin}} \sum_{n=1}^{N} (t_n - \phi(X_n).W^T)^2 + \lambda \lVert w \rVert_2^2$ | Pseudoinversa<br>$W^* = (\phi^T \phi + \lambda I)^{-1} \phi^T t$ | Se penaliza la norma de los pesos de w, con el objetivo de que mi modelo no sea tan sensible a cambios pequeños en x. $\lambda$ es un hiperparámetro y controla la "fuerza" de la penalización y garantiza la invertibilidad. Esto evita overfiting → Para evitar que mi modelo se 'aprenda' los datos y compense con grandes pesos las variaciones |
| MLE - Maximun Likelihood Estimation | $\underset{w,\sigma^2}{\text{argmax}}\ \text{Log}\ p(t \mid \Phi, w, \sigma^2)$<br>$= \underset{w,\sigma^2}{\text{argmax}}\ \text{Log}\left(\prod_{n=1}^{N} N(t_n \mid \phi(x_n) w^T, \sigma_n^2)\right)$ | $(\phi^T \phi)^{-1} \phi^T . t = W$ | La verosimilitud es la probabilidad de haber observado exactamente esa secuencia de datos, dado cierta hipótesis. El principio de máxima verosimilitud establece que el mejor modelo es el que hace que los datos observados sean lo más probable posible.<br>La solución coincide con la de mínimos cuadrados, pero ya no es 'quiero minimizar el error' ahora quiero que mis datos sean lo mas probables posibles. |
| MAP- Maximum a posteriori. | $= \underset{w}{\text{argmin}} \left(\frac{1}{2\sigma^2} \lVert t - \phi W \rVert^2 + \frac{\sigma_n^2}{\sigma_w^2} \lVert w \rVert^2\right)$ | $W^* = (\phi^T \phi + \lambda I)^{-1} \phi^T t$<br>$\lambda = \sigma_n^2 / \sigma_w^2$ | Aparece el concepto de Prior, el concepto de incorporar conocimiento, creencias o experiencia que ya se tiene sobre el fenómeno antes de ver cualquier dato. Luego se lleva a cabo el experimento y se actualiza la creencia. Ese prior es una distribución de probabilidad sobre w. |
| Bayesiano Lineal Gaussiano | $= \underset{w}{\text{argmax}}\ p(w \mid t) = \underset{w}{\text{argmax}}\ p(t \mid w) p(w)$ | $p(w \mid t) = N(w \mid m_N, S_N)$<br>Parámetros actualizados<br>$S_N^{-1} = \alpha I + \frac{1}{\sigma^2} \phi^T . \phi$<br>$m_N = \frac{1}{\sigma^2} . S_N . \phi^T . t$ | Aquí aparece la "confianza de cada predicción" porque aparece la varianza (Incertidumbre). Quiero llegar a tener una distribución. Mi posterior tiene: Media, varianza. No busca minimizar el costo, busca actualizar una distribución de probabilidad sobre los parámetros (Media y varianza |
| Regresión Rigida Kernel | $\underset{\alpha}{\text{argmin}} \left(\lVert t - K\alpha \rVert^2 + \frac{\lambda}{2} \alpha^T K \alpha\right)$<br>$\alpha \rightarrow$ Son los pesos (w) pero en otro espacio de representación<br>$w = a_1 \phi(x_1) + a_2 \phi(x_2) \dots a_N \phi(x_N)$<br>a = Qué tanto $X_1$ aporta al modelo<br>$K_{ij} = \phi(x_i)^T . \phi(x_i)$ | $\hat{a} = (K + \lambda I_N)^{-1} . t$<br>Predicción<br>$t(X_*) = K(X_*)^T . \hat{a}$ | El Kernel es una función que mide qué tan parecidas son dos cosas, pero en un espacio transformado no en un espacio original $K = \Phi \Phi^T$. $K_{ij} = k(x,x)$ mide la similitud entre dos puntos de entrenamiento<br>↑ Parecidos ↑ Kernel Su principio parte de mínimos cuadrados regularizados. Es un modelo que extiende la regresión lineal a contextos no lineales. En su espacio original no se ajustan linealmente, pero se pueden volver lineales a un espacio de características mayor.<br>Tiene un enfoque determinista. |
| Procesos Gaussianos | $= \underset{w}{\text{argmax}}\ p(w \mid t) = \underset{w}{\text{argmax}}\ p(t \mid w) p(w)$<br>$= \underset{w}{\min} \left(\frac{1}{2\sigma_n^2} \lVert t - \phi w \rVert_2^2 + \frac{1}{2} w^T \Sigma_p^{-1} w\right)$ | Solución Optima<br>$\hat{a} = (K + \sigma_n^2 I)^{-1} . t$<br>Media posterior<br>$\mu_*(X_*) = k(x_*)^T . (K + \sigma_n^2 I)^{-1} . t$<br>Varianza asociada<br>$\sigma_*^2(X_*) = k(x_*, x_*) - k(x_*)^T (K + \sigma_n^2 I)^{-1} k(x_*)$ | Un proceso Gaussiano es una distribución sobre funciones. Me da la varianza que se asocia a la incertidumbre del modelo. GP asume que los valores observados como los que quiero predecir siguen una distribución normal multivariada.<br>$\begin{bmatrix} t \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K + \sigma_n^2 I & k(x_*) \\ k(x_*) & k(x_*, x_*) \end{bmatrix}\right)$ |