Mínimos Cuadrados

$$t_n = \varphi(x_n)^T w + \varepsilon_n$$

$$w = \arg\min J(w)$$

↳ Función de Costo

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

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

Nota: los factores $1/2$ son para simplificar la derivada

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

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

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

- Igualamos a Cero

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

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

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

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

Mínimos Cuadrados Regularizados

$$t_n = \varphi(x_n)^T w + \epsilon_n \quad \epsilon_n \sim N(0, \sigma_n^2)$$

$$\hat{w}_{ridge} = \text{argmin } J(w)$$

↳ Función de Costo

- Agregamos la norma L2 para penalizar pesos grandes

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

Nota: los factores $1/2$ son para simplificar la derivada

$$\hat{w}_{ridge} = \text{argmin}\left(\frac{1}{2}\|t - \phi w\|_2^2 + \frac{\lambda}{2}\|w\|_2^2\right)$$

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

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

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

$$\nabla_w J(w) = 0 - t^T\phi + \phi^T\phi w + \lambda w$$

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

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

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

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

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

Maxima Verosimilitud

$$t_n = \varphi(x_n)^T w + \eta_n \quad \eta_n \sim N(0, \sigma_n^2)$$

- Sigue una distribución normal.

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

- Asumiendo i.i.d. la verosimilitud conjunta, es el producto de todas las probabilidades individuales.

$$p(t | x, w, \sigma_n^2) = \prod_{n=1}^{N} N(t_n | \varphi(x_n)^T w, \sigma_n^2)$$

- Forma de una densidad gaussiana unidimensional.

$$N(t_n | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(t_n - \mu^2)\right)}$$

$$p(t | x, w, \sigma_n^2) = \prod_{n=1}^{N} p(t_n | x_n, w, \sigma_n^2)$$

$$p(t | x, w, \sigma_n^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\left(-\frac{1}{2\sigma_n^2}(t_n - \varphi(x)^T w)^2\right.}$$

- El producto de las constantes $(2\pi\sigma_n^2)^{-1/2}$ se convierte en:

$$\prod_{n=1}^{N} (2\pi\sigma_n^2)^{-1/2} = (2\pi\sigma_n^2)^{-N/2}$$

$$p(t | x, w, \sigma_n^2) = (2\pi\sigma_n^2)^{-N/2} \prod_{n=1}^{N} e^{\left(-\frac{1}{2\sigma_n^2}(t_n - \varphi(x)^T w\right.}$$

$$\prod_n e^{a_n} = e^{\sum_n a_n}$$

$$p(t|X, w, \sigma_2^2) = (2\pi\sigma_2^2)^{-N/2} e^{\left(\frac{-1}{2\sigma_2^2}\sum_{n=1}^{N}(t_n - \varphi(X_n)^T w)^2\right)}$$

$$\sum_{n=1}^{N}(t_n - \varphi(X_n)^T w)^2 = (t - \phi w)^T (t - \phi w)$$

$$p(t|X, w, \sigma_2^2) = (2\pi\sigma^2)^{-N/2} e^{\left(\frac{-1}{2\sigma_2^2}\|t - \phi w\|^2\right)}$$

- Es mas conveniente trabajar con el logaritmo de la verosimilitud

$$\ln p(t|X, w, \sigma_2^2) = \frac{-N}{2}\ln(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\|t - \phi w\|^2$$

$$\hat{w}_{ML} = \arg\max \ln p(t|X, w, \sigma_2^2)$$

↳ Maxima la log-verosimilitud

- Como el primer termino no depende de w, es equivalente a minimizar el termino cuadratico

$$\hat{w}_{ML} = \arg\min \|t - \phi w\|^2$$

$$\hat{w}_{ML} = (\phi^T \phi)^{-1} \phi^T t$$

$$\sigma_2^2 = \frac{1}{N}\|t - \phi \hat{w}_{ML}\|^2$$

Maximo a posteriori

$$t_n = \varphi(x_n)^T w + \eta_n, \quad \eta \sim N(0, \sigma_\eta^2)$$

- En basa al MLE (Maxima Valorsimilitud).

$$\hat{w}_{MLE} = \arg\max p(t|X, w)$$

$$p(t|X, w) = (2\pi\sigma_\eta^2)^{-N/2} e^{\left(\frac{-1}{2\sigma_\eta^2}\|t - \phi w\|^2\right)}$$

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

- Introducimos un prior sobre los pasos w, asumiendo que los pasos siguen una distribución normal centrada en 0.

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

$$p(w) = (2\pi\sigma_w^2)^{-N/2} e^{\left(\frac{-1}{2\sigma_w^2}\|w\|^2\right)}$$

- Se aplica al teorema de bayes.

$$p(w|t, x) = \frac{p(t|X, w)p(w)}{p(t|X)}$$

$$W_{map} = \arg\max \log p(t|X, w) + \log p(w)$$

$$\log p(t|X, w) = \frac{-N}{2}\log(2\pi\sigma_\eta^2) - \frac{1}{2\sigma_\eta^2}\|t - \phi w\|^2$$

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

$$\log p(w|t, x) = C - \frac{1}{2\sigma_\eta^2}\|t - \phi w\|^2 - \frac{1}{2\sigma_w^2}\|w\|^2$$

$$J(w) = \frac{1}{2\sigma_2^2}\|t - \phi w\|^2 + \frac{1}{2\sigma_w^2}\|w\|^2$$

$$\lambda = \frac{\sigma_2^2}{\sigma_w^2}$$

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

$$\nabla_w J(w) = -\phi^T t + \phi^T \phi w + \lambda w = 0$$

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

↳ Mismo resultado que con mínimos cuadrados regularizados

Día___ Mes___ Año___ Lun Mar Mié Jue Vie Sáb

Primer parcial TAM

1 $t_n = \phi(x_n)w^T + \eta_n$

Bayesiano con modelo lineal gaussiano:

En lugar de buscar un solo numero w, el modelo devuelve una distribucion completa sobre w, que refleja incertidumbre.

Expresión del posterior

$p(w|t, X) = N(w|w_{MAP}, \Sigma_{MAP})$ → la distribución de w despues de ver los datos es una normal multivariada con media y covarianza

media covarianza

Media:

$$w_{MAP} = \Sigma_N\left(\Sigma_0^{-1}m_0 + \frac{1}{\sigma_n^2}\Phi^T t\right)$$

Covarianza:

$$\Sigma_{MAP} = \left(\Sigma_0^{-1} + \frac{1}{\sigma_n^2}\Phi^T\Phi\right)^{-1}$$

Al tener más datos, más pequeño se vuelve Σ_N, o sea, el modelo se vuelve más seguro sobre los pesos w.

Si asumimos $m_0 = 0$ → $w_{MAP} = \frac{1}{\sigma_n^2}\Sigma_N\Phi^T t$

Prior sobre pesos:

$$p(w) = N(w|m_0, \Sigma_0)$$

Verosimilitud:

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

$$p(w|t_n, x_n) = p(t|w)\cdot p(w)$$

$$p(w|t_n, x_n) = -\frac{1}{2\sigma_n^2}\left(\|t_n\|_2^2 - 2t_n\phi(x_n)^T w + \phi(x_n)^T w\phi(x_n)w^T\right)$$

$$= -\frac{1}{2}\Sigma_0^{-1}(w-m_0)^T(w-m_0)$$

$$= -\frac{1}{2}\left(w^T\Sigma_0^{-1}w - (w^T)^T\Sigma_0^{-1}(m_0)^T - m_0^T\Sigma_0^{-1}w + \Sigma_0^{-1}m_0^T m_0\right)$$

$$= -\frac{1}{2}\left(\underbrace{\Sigma_0^{-1}w^T w}_{\text{cuadratico}} - \underbrace{2\Sigma_0^{-1}m_0^T w}_{\text{Lineal}} + \underbrace{\Sigma_0^{-1}m_0^T m_0}_{\text{constante}}\right)$$

Derivamos $p(w|t_n, x_n)$ respecto a w:

$$\frac{dp(w|t_n, x_n)}{dw} = -\frac{1}{2\sigma_n^2}\left(0 - 2t_n\phi(x_n)^T + 2\phi(x_n)^T\phi(x_n)w\right) - \frac{1}{2}\left(2\Sigma_0^{-1}w - 2\Sigma_0^{-1}m_0^T + 0\right)$$

$$-\frac{1}{2}\left(\frac{-2}{\sigma_n^2}t_n\phi(x_n)^T + \frac{2\phi(x_n)^T\phi(x_n)w}{\sigma_n^2} - 2\Sigma_0^{-1}w - 2\Sigma_0^{-1}m_0^T\right) = 0$$

$$\underbrace{\frac{\phi(x_n)^T\phi(x_n)w}{\sigma_n^2}}_{\text{cuadratico}} + \Sigma_0^{-1}w = \underbrace{\frac{t_n\phi(x_n)^T}{\sigma^2 n}}_{\text{lineal}} + \Sigma_0^{-1}m_0^T$$

Marden

Dia ___ Mes ___ Año ___ Lun Mar Mié Jue Vie Sab

$$w\left(\Sigma_0^{-1} I + \frac{\phi(x_n)^T \phi(x_n)}{\theta^2 n}\right) = \frac{t_n \phi(x_n)^T}{\theta^2 n} + \Sigma_0^{-1} m_0^T$$

Pesos optimos: S_N M_N

$$w = \left(\Sigma_0^{-1} I + \frac{\phi(x_n)^T \phi(x_n)}{\theta^2 n}\right)^{-1} \left(\Sigma_0^{-1} m_0^T + \frac{t_n \phi(x_n)^T}{\theta^2 n}\right)$$

$$p(w) = N(w|0, \theta^2 w) \rightarrow \frac{p(w|t_n, x_n)}{p(w|t_n)} = N(w|\bar{m}_N, \bar{S}_N)$$

$$w = \left(\frac{1}{\theta^2 n} I + \frac{\phi(x_n)^T \phi(x_n)}{\theta^2 n}\right)^{-1} \frac{\phi(x_n)^T t_n}{\theta^2 n} = \bar{M}_N$$

$\bar{S}_N$

$$\bar{S}_N = \left(\frac{1}{\theta^2 n}\right)^{-1} \left(\frac{\theta^2 n}{\theta_\omega^2} I + \frac{\phi(x_n)^T \phi(x_n)}{\theta^2 n}\right)^{-1} \phi(x_n)^T t_n$$

$$w = \left(\frac{\phi(x_n)^T \phi(x_n)}{\theta^2 n} + \lambda I\right)^{-1} \phi(x_n)^T t_n$$

Regresion rigida Kernel:

Este modelo mapea x_n a un espacio de dimension mayor con $\phi(x_n)$

$k(x, x') = \phi(x)^T \phi(x')$ mejora

Arrancamos de Bayesiano

$$p(w|t_n) = \|t_n - \phi(x_n) w^T\|_2^2 + \lambda \|w\|_2^2$$

$$w^* = \underset{w}{\operatorname{argmin}} \|t_n - \phi(x_n) w^T\|_2^2 + \lambda \|w\|_2^2 = (\phi^T(x_n)\phi(x_n) + \lambda I)^{-1} \phi^T(x_n) t_n$$

$$w^* = \overbrace{(\phi^T(x_n)\phi(x_n) + \lambda I)^{-1}}^{\alpha_i} t_n \underbrace{\phi^T(x_n)}_{\phi(x_i)^T}$$

$k(x, x')$

$$w^* = \sum_i \alpha_i \phi(x_i)^T$$

Se necesita predicción para un nuevo punto proyectando sobre w^*

$$t_* = \phi(x_*)^T w^*$$

$$= \underbrace{\phi(x_*)^T \phi^T(x_n)}_{k(*)^T} \overbrace{(\phi^T(x_n)\phi(x_n) + \lambda I)^{-1}}^{\alpha_i} t_n$$

$k(x, x')$

$$t_* = (k + \lambda I)^{-1} t_n k(*)^T \qquad k(*)^T = [k(x_*, x_1), \ldots, k(x_*, x_N)]$$

Marden

Dia___ Mes___ Año___ Lun Mar Mié Jue Vie Sab

Procesos gaussianos

Este modelo generaliza la regresion de kernel modelando distribuciones sobre funciones y se define por la funcion media y funcion covarianza

$\mathcal{GP}: F(x) \sim \mathcal{GP}(m(x), k(x,x'))$

$f(x) = \phi(x_n)^T w \quad : w \sim N(0, \Sigma_{MAP}) \rightarrow$ sin ruido
$\Sigma_{MAP} \downarrow SN$

media:

$$m(x) = \mathcal{E}\{f(x)\} = \mathcal{E}\{\phi(x_n)^T w\} = \phi(x_n)^T \mathcal{E}\{w\} = 0$$

Covarianza:

$$k(x,x') = \text{cov}(f(x), f(x'))$$

$$k(x,x') = \mathcal{E}\{f(x), f(x')\} = \mathcal{E}\{\phi(x)^T w \phi^T(x') w^T\}$$

$$k(x,x') = \phi(x)^T \mathcal{E}\{w w^T\} \phi(x')$$

$$k(x,x') = \phi(x)^T \Sigma_w \phi(x')$$

$f(x) \sim \mathcal{GP}(f(x) | 0, K): \quad K \in \mathbb{R}^{N \times N} \sim K_{ij} = k(x_i, x_j)$

$t_n = f(x) + \eta_n \longrightarrow$ Ruido $N(0, \theta_n^2)$

Prior sobre $f(x)$:

$$p(f(x)) = N(f(x) | 0, K)$$

Verosimilitud:

$$p(t_n | f(x)) = N(t_n | f(x), \theta_n^2 I_N)$$

Distribucion marginal de t_n:

$$p(t_n) = \int p(t | f(x))\, p(f(x))\, df(x) = N(t | 0, K_t + \theta_n^2 I_N)$$

$$= \int N(t_n | f(x), \theta_n^2 I_N) (\theta_n^2 I_N)^{-1} (t_n - f(x)) - \frac{1}{2} f(x)^T K^{-1} f(x)\, df(x)$$

$$= \int e^{\left(-\frac{1}{2}\left((\theta_n^2 I_N)^{-1}(\|t_n\|^2 - 2 t_n^T f(x) + f(x)^T f(x)) + f(x)^T f(x) K^{-1}\right)\right)} df(x)$$

$$= \int e^{\left(-\frac{1}{2}\underbrace{\left((\theta_n^2 I_N)^{-1} + K^{-1}\right)}_{a} f(x)^T f(x) + \underbrace{(\theta_n^2 I_N)^{-1} t_n^T f(x)}_{b} \underbrace{- \frac{(\theta_n^2 I_N)^{-1} t_n^T t_n}{2}}_{c}\right)} df(x)$$

$$= \int e^{\left(-\frac{1}{2} a f(x)^T f(x) + b f(x) + c\right)} df(x) \rightarrow \mu_f = a^{-1} b$$

$$= \int e^{\left(-\frac{1}{2}(f(x) - \mu_f)^T a (f(x) - \mu_f) + \frac{1}{2} b^T a^{-1} b + c\right)} df(x)$$

$$= e^{\left(\frac{1}{2} b^T a^{-1} b + c\right)}$$

$$= e^{\left(\frac{1}{2}(\theta_n^4 I_N)^{-1} t_n^T a^{-1} t_n - \frac{(\theta_n^2 I_N)}{2} t_n^T t_n\right)}$$

Marden

Dia____ Mes____ Año____ Lun Mar Mie Jue Vie Sab

$$p(t_n) = e^{\left(-\frac{1}{2} t_n^T \left(-(\theta_n^2 I_N)^{-1} (k + \theta_n^2 I_N)^{-1} k + (\theta_n^2 I_N)^{-1}\right) t_n\right)}$$

$$= e^{\left(-\frac{1}{2} t_n^T \underbrace{(k + \theta_n^2 I_N)^{-1}}_{\text{varianza}} t_n\right)}$$

$$p(t_n) = N(t_n | 0, k + \theta_n^2 I_N)$$