Teoría de Aprendizaje de máquina / Juan Jerónimo Castaño Rivera

1) Sea el modelo

$t_n = \phi(x_n)\mathbb{W}^T + \eta_n$, $t_n \in \mathbb{R}$, $x_n \in \mathbb{R}^P$

$\mathbb{W} \in \mathbb{R}^Q$; $\phi \in \mathbb{R}^P \to \mathbb{R}^Q$

$\eta_n \sim N(\eta_n | 0, \sigma_n^2)$

De forma matricial $\quad t = \Phi \mathbb{W} + \eta$

Donde

$\Phi \in \mathbb{R}^{N \times Q}$ = Datos

$t \in \mathbb{R}^N$ = Vector de salida

$\mathbb{W} \in \mathbb{R}^Q$ = Modelo

- Minimos Cuadrados

Optimización $\quad \mathbb{W}^* = \underset{\mathbb{W}}{\arg\min} \frac{1}{N} \| t - \phi \mathbb{W} \|_2^2$

El valor minimo se puede hallar derivando e igualando a cero

$\| t - \phi \mathbb{W} \|_2^2 = \langle t - \phi \mathbb{W}, t - \phi \mathbb{W} \rangle = (t - \phi \mathbb{W})^T (t - \phi \mathbb{W})$

$= t^T t - t^T \phi \mathbb{W} - (\phi \mathbb{W})^T t + (\phi \mathbb{W})^T (\phi \mathbb{W})$

$\frac{\partial}{\partial \mathbb{W}} \frac{1}{N} \{ t^T t - 2 t^T \phi \mathbb{W} + (\phi \mathbb{W})^T (\phi \mathbb{W}) \} = 0$

$= 0 - (2 t^T \phi)^T + 2 \phi^T \phi \mathbb{W} = 0$

$-2 \phi^T t + 2 \phi^T \phi \mathbb{W} = 0 \quad \Rightarrow \quad \cancel{2} \phi^T t = \cancel{2} \phi^T \phi \mathbb{W}$

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

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

- Mínimos Cuadrados regularizados

$$\mathbb{W} = \underset{\mathbb{W}}{\operatorname{argmin}} \|\mathbb{t} - \Phi \mathbb{W}\|_2^2 + \lambda \|\mathbb{W}\|_2^2$$

$$\mathbb{W}^* = \frac{\partial}{\partial \mathbb{W}} \left\{ \|\mathbb{t} - \Phi \mathbb{W}\|_2^2 + \lambda \|\mathbb{W}\|_2^2 \right\} = 0$$

$$\Rightarrow -2\Phi^T \mathbb{t} + 2\Phi^T \Phi \mathbb{W} + 2\lambda \mathbb{W} = 0$$

$$(2\Phi^T\Phi + 2\lambda \mathbb{I})\mathbb{W} = 2\Phi^T \mathbb{t}$$

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

- Máxima verosimilitud

$$t_n = \phi(x_n)^T \mathbb{W} + \eta_n \; ; \quad \eta \sim \mathcal{N}(0, \sigma_n^2)$$

Se puede ver que $p(t_n | x_n, \mathbb{W}) = \mathcal{N}(t_n | \phi(x_n)^T \mathbb{W}, \sigma_n^2)$

Dado que son iid $\prod_{n=1}^{N} p(t_n | \phi(x_n)\mathbb{W}, \sigma_n^2)$

$$\log\left(\prod_{n=1}^{N} p(t_n | \phi_n \mathbb{W}, \sigma_n^2)\right) = \log\left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{|t_n - \phi_n \mathbb{W}|^2}{2\sigma_n^2}\right)\right)$$

$$= N \log\left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right) - \frac{1}{2\sigma_n^2} \sum_{n=1}^{N} |t_n - \phi_n \mathbb{W}|^2$$

$$= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} \sum_{n=1}^{N} |t_n - \phi_n \mathbb{W}|^2$$

Se deriva e iguala a cero, el problema corresponde a mínimos cuadrados

$$\mathbb{W}^* = \operatorname{argmin} -\frac{1}{2\sigma_n^2} \|\mathbb{t} - \Phi \mathbb{W}\|_2^2 - \frac{N}{2} \log(2\pi\sigma_n^2)$$

$$\mathbb{W}^* = (\Phi^T \Phi)^{-1} \Phi^T \mathbb{t}$$

- Maximo a-posteriori

En lugar de encontrar el valor de $\mathbb{w}$ que maximiza la verosimilitud, se asume un prior sobre $\mathbb{w}$ y buscamos el valor que maximiza la posterior

$$\mathbb{w}^* = \operatorname{argmax}\ p(\mathbb{w} \mid \mathbb{t})$$

$$p(t, w) = p(w, t) \Rightarrow p(t|w)p(w) = p(w|t)p(t)$$

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

Para simplificar $p(t) = 1$

Al igual que en máxima verosimilitud

$$p(t|w) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \phi(x_n) w, \sigma_n^2) \qquad p(w) = \mathcal{N}(w \mid 0, \sigma_w^2)$$

$$= \prod_{n=1}^{Q} \mathcal{N}(w_q \mid 0, \sigma_w^2)$$

$$= \log\left(\prod_{n=1}^{N} p(t_n \mid \phi_n w, \sigma_n^2) \prod_{n=1}^{Q} p(w_q \mid 0, \sigma_w^2)\right)$$

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

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

$$+ \log\left(\prod_{q=1}^{Q} \exp\left(-\frac{|w_q|^2}{2\sigma_w^2}\right)\right)$$

$$= \underbrace{-\frac{N}{2}\log(2\pi\sigma_n^2)}_{\text{No aporta}} - \underbrace{\frac{Q}{2}\log(2\pi\sigma_q^2)}_{\text{no aporta}} - \frac{1}{2\sigma_n^2}\sum_{n=1}^{N} |t_n - \phi_n w|^2 - \frac{1}{2\sigma_w^2}\sum_{q=1}^{Q} |w_q|^2$$

$$w^* = \min_{\mathbb{w}} \left[\| \mathbb{t} - \Phi \mathbb{w} \|^2 + \frac{2\sigma_n^2}{2\sigma_w^2} \| \mathbb{w} \| \right] \qquad \text{Min. cuad. regularizados}$$

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

- Bayesiano con modelo lineal Gaussiano

Este modelo no solo busca un único vector w, sino la distribución completa posterior sobre w dados los datos observados

$$p(w|t) = N(w|m_N, \Sigma_N)$$

Σ_N: Matriz covarianza $\Sigma_N = \left(\frac{1}{\sigma_w^2}\mathbb{I} + \frac{1}{\sigma_n^2}\Phi^T\Phi\right)$

$$m_N = \frac{1}{\sigma_n^2}\Sigma_N \Phi^T t$$

Optimización: Maximizar la evidencia

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

Demostración:

Sea un vector $x \in \mathbb{R}^q$ con prior Gaussian

$$p(x) = N(x|\mu, \Lambda^{-1})$$

Sea el modelo $y = Ax + b$; $p(y|x) = N(y|Ax+b, L^{-1})$

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

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

Para nuestro caso

$t_n = \Phi(x_n)w^T + \eta_n$ $\quad t = \Phi w + \eta$; $p(t|w) = N(t|\Phi w, \sigma_n^2)$

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

Sabemos que $\log p(t|w) = -\frac{1}{2\sigma_n^2}\|t - \Phi w\|_2^2 + cte.$

$$\log(N(x|\mu,\Sigma)) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi)$$

Por tanto $\log(p(w)) = -\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0) + cte$

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

$$= -\frac{1}{2}\left[\frac{1}{\sigma_n^2}(\mathbf{t}-\Phi\mathbb{W})^T(\mathbf{t}-\Phi\mathbb{W}) + (\mathbb{W}-m_0)^T S_0^{-1}(\mathbb{W}-m_0)\right] + ctc$$

Agrupando términos

$$\mathbb{W}^T\left(\frac{1}{\sigma_n^2}\Phi^T\Phi + S_0^{-1}\right)\mathbb{W} - 2\mathbb{W}^T\left(\frac{1}{\sigma_n^2}\Phi^T\mathbf{t} + S_0^{-1}m_0\right)$$

Se sabe que $\log(p(w)) = -\frac{1}{2}(W-m_0)^T S_0^{-1}(W-m_0)$

$$= -\frac{1}{2}\left[W^T S_0^{-1} W - 2W^T S_0^{-1} m_0 + \cancel{m_0^T S_0^{-1} m_0}^{ctc}\right]$$

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

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

como $p(w) = N(W|m_0, S_0) = N(W|0, \sigma_w^2)$

$$S_N = \left(\frac{1}{\sigma_w^2}\mathbb{I} + \frac{1}{\sigma_n^2}\Phi^T\Phi\right)^{-1} \qquad m_N = \frac{1}{\sigma_n^2} S_N \Phi^T\mathbf{t}$$

$$m_N = \left(\frac{1}{\sigma_n^2}\right)\left(\frac{1}{\sigma_n^2}\right)^{-1}\left(\frac{\sigma_n^2}{\sigma_w^2}\mathbb{I} + \Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t} = \left(\frac{\sigma_n^2}{\sigma_w^2}\mathbb{I} + \Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t}$$

- Regresión rígida Kernel

Se extiende a espacios de funciones no lineales

$$\hat{\mathbb{Y}} = \Phi\mathbb{W} \quad ; \quad \Phi \in \mathbb{R}^{N \times Q} \quad ; \quad Q \to \infty \qquad (RKHS)$$

$$\hat{\mathbb{Y}} : \mathbb{R}^Q \to \mathbb{R}$$

$$\mathbb{W}^* = \underset{\mathbb{W}}{\arg\min}\ \frac{1}{N}\|\mathbf{t} - \Phi\mathbb{W}\|_2^2 + \lambda\|\mathbb{W}\|_H^2$$

Por mínimos cuadrados $\mathbb{W}^* = (\Phi^T\Phi + \lambda\mathbb{I})^{-1}\Phi^T\mathbf{t}$

Sin embargo $\Phi^T\Phi \in \mathbb{R}^{Q\times Q}$; $Q \to \infty$

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

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

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

Se hace predicción para nuevos puntos $f(x_*) = \phi(x_*)^T \mathbb{W}^*$

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

$$K = \Phi\Phi^T \quad ; \quad K_* = \phi(x_*)^T \Phi^T$$

$$f(x_*) = [K_*(\cdot)^T](K + \lambda I) t$$

- Gaussian Process

Extienden el método paramétrico para definir la incertidumbre de los parámetros del regresor al imponer un prior sobre funciones directamente en RKHS

El GP se define por su media y covarianza

$f(x) \in \mathbb{R}$; $f(x) = \phi(x)^T w$; $p(w) = N(w|0, \Sigma_w)$; $\Sigma_w \in \mathbb{R}^{Q \times Q}$

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

$$\text{cov}(f(x), f(x')) = \kappa(x, x') = \phi(x)^T \Sigma_w \phi(x')$$

$$f \sim GP(f|0, K) \quad ; \quad K = [\kappa(x, x')] \in \mathbb{R}^{N \times N}$$

Este modelo busca maximizar la verosimilitud $p(t|\theta)$, siendo θ un hiperparámetro de $\kappa(\cdot, \cdot|\theta)$

Se puede obtener que

$$m(x_*) = \mathbb{K}_*^T(K + \sigma_\epsilon^2 I)^{-1} t \quad , \quad \mathbb{K}_* = [\kappa(x_*, x_1), \kappa(x_*, x_2), \dots, \kappa(x_*, x_N)]^T$$

$$\text{cov}(f(x), f(x')) = \kappa(x_*, x_*) + \sigma_\epsilon^2 - \mathbb{K}_*^T(K + \sigma_\epsilon^2 I)^{-1}\mathbb{K}_*$$