Formula General del problema de optimización

$$\min_{W}\left(\underbrace{\frac{1}{2}\sum_{n=1}^{N}(t_n-\phi(X_n)W^T)^2}_{\text{Función de pérdida}}+\underbrace{\frac{\lambda}{2}\|W\|^2}_{\text{término de regularización } L_2}\right)$$

- La función de pérdida representa el error cuadrático medio entre las predicciones y los valores reales.
- El término de regularización el la norma L_2 de los pesos, que penaliza los valores grandes de W
- λ controla la fuerza de penalización.

1. Mínimos cuadrados ordinarios (OLS)

Se desea minimizar el error entre las predicciones y los valores reales.

$$\min_{W}\sum_{n=1}^{N}(t_n-\phi(X_n)W^T)^2$$

- Se define la matriz de diseño Φ $(N\times Q)$, donde cada fila es $\phi(X_n)$.
- El vector de observaciones es $t=[t_1,\dots,t_N]^T$.
- El error cuadrático se reescribe como:

$$J(w)=(t-\Phi W)^T(t-\Phi W)$$

Para encontrar el W que minimiza $J(W)$, necesitamos calcular la derivada de $J(W)$ con respecto a W e igualarla a cero.

$$\begin{aligned}J(W)&=(t^T-(\Phi W)^T)(t-\Phi W)\\&=(t^T-W^T\Phi^T)(t-\Phi W)\\&=t^Tt-t^T\Phi W-W^T\Phi^Tt+W^T\Phi^T\Phi W\end{aligned}$$

Simplificamos los términos. dado que $\mathbf{t}^T \Phi \mathbf{w}$ es un escalar, es igual a su traspuesta $(\mathbf{t}^T \Phi \mathbf{w})^T = \mathbf{w}^T \Phi^T \mathbf{t}$.

$$J(\mathbf{w}) = \mathbf{t}^T\mathbf{t} - 2\,\mathbf{w}^T \Phi^T \mathbf{t} + \mathbf{w}^T \Phi^T \Phi \mathbf{w}$$

Calculamos la derivada de $J(\mathbf{w})$ con respecto a $\mathbf{w}$

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = -2\Phi^T\mathbf{t} + 2\Phi^T\Phi\mathbf{w}$$

Igualamos a cero

$$-2\,\Phi^T\mathbf{t} + 2\,\Phi^T\Phi\,\mathbf{w} = 0$$
$$\Phi^T\Phi\,\mathbf{w} = \Phi^T\mathbf{t}$$

Si la matriz $\Phi^T\Phi$ es invertible (lo que significa que las columnas de Φ son linealmente independientes y $N \geq @$), podemos multiplicar ambos lados por $(\Phi^T\Phi)^{-1}$

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

2. Mínimos cuadrados regularizados

Para prevenir el sobre ajuste, se añade un término de penalización a la función de costo. La regularización L2 penaliza la suma de los cuadrados de los cuadrados de los coeficientes $\mathbf{w}$.

$$J_R(\mathbf{w}) = \sum_{n=1}^{N} (t_n - \phi(x_n)^T\mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2$$

λ es el parámetro de regularización, y $\|\mathbf{w}\|_2^2 = \mathbf{w}^T\mathbf{w}$.
en forma matricial

$$J_R(\mathbf{w}) = (\mathbf{t} - \Phi\mathbf{w}) + \lambda\,\mathbf{w}^T\mathbf{w}$$
$$= \mathbf{t}^T\mathbf{t} - 2\,\mathbf{w}^T\Phi^T\mathbf{t} + \mathbf{w}^T\Phi^T\Phi\mathbf{w} + \lambda\mathbf{w}^T\mathbf{w}$$

Para los primeros tres términos la derivada es la misma que en OLS. Solo necesitamos la derivada del término de regularización:

$$\frac{\partial(\lambda\mathbf{w}^T\mathbf{w})}{\partial(\mathbf{w})} = \lambda\,\frac{\partial(\mathbf{w}^T I\mathbf{w})}{\partial(\mathbf{w})} = \lambda(2I\mathbf{w}) = 2\lambda\mathbf{w}$$

Entonces la derivada completa es:

$$\frac{\partial J_R(\mathbf{w})}{\partial(\mathbf{w})} = -2\Phi^T \mathbf{t} + 2\Phi^T\Phi\mathbf{w} + 2\lambda\mathbf{w}$$

Igualamos a cero y despejamos $\mathbf{w}$

$$-2\Phi^T\mathbf{t} + 2\Phi^T\Phi\mathbf{w} + 2\lambda\mathbf{w} = 0$$
$$\Phi^T\Phi\mathbf{w} + \lambda\mathbf{w} = \Phi^T\mathbf{t}$$
$$(\Phi^T\Phi + \lambda\mathbb{I})\mathbf{w} = \Phi^T\mathbf{t}$$

La matriz $(\Phi^T\Phi + \lambda\mathbb{I})$ es invertible para $\lambda > 0$

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

3. Máxima verosimilitud (MLE)

Problema de optimización:

El modelo de regresión es $t_n = \phi(\mathbf{x}_n)^T\mathbf{w} + \eta$ con $\eta_n \sim \mathcal{N}(0, \sigma_\eta^2)$.
esto implica que t_n dado $\mathbf{x}_n, \mathbf{w}, \sigma_\eta^2$ sigue una distribución gausiana:

$$p(t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma_\eta^2) = \mathcal{N}(t_n \mid \phi(\mathbf{x}_n)^T\mathbf{w}, \sigma_\eta^2)$$
$$p(t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma_\eta^2) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{(t_n - \phi(\mathbf{x}_n)^T\mathbf{w})^2}{2\sigma_\eta^2}\right)$$

Dado que los términos son iid., la función de verosimilitud para todos los datos $\mathbf{t}$ (dado $\mathbf{x}$, $\mathbf{w}$ y σ_η^2) es el producto de las probabilidades individuales:

$$L(\mathbf{w}, \sigma_\eta^2) = p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \sigma_\eta^2) = \prod_{n=1}^{N} p(t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma_\eta^2)$$

$$L(\mathbf{w}, \sigma_\eta^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{(t_n - \phi(\mathbf{x}_n)^T\mathbf{w})^2}{2\sigma_\eta^2}\right)$$

$$L(\mathbf{w}, \sigma_\eta^2) = \left(\frac{1}{2\pi\sigma_\eta^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma_\eta^2}\sum_{n=1}^{N}(t_n - \phi(\mathbf{x}_n)^T\mathbf{w})^2\right)$$

Es computacional más conveniente maximizar la log-verosimilitud (log-likelihood):

$$\mathcal{L}(W, \sigma_n^2) = \ln \mathcal{L}(W, \sigma_n^2)$$
$$\mathcal{L}(W, \sigma_n^2) = \frac{N}{2} \ln(2\pi \sigma_n^2) - \frac{1}{2\sigma_n^2} \sum_{n=1}^{N} (t_n - \phi(x_n)^T W)^2$$

En forma matricial, la suma de cuadrados es $(t - \Phi W)^T (t - \Phi W)$.

$$\mathcal{L}(W, \sigma_n^2) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma_n^2) - \frac{1}{2\sigma_n^2} (t - \Phi W)^T (t - \Phi W)$$

El problema de optimización es encontrar (W y σ_n^2) que maximicen $\mathcal{L}(W, \sigma_n^2)$.

Calculamos la derivada de $\mathcal{L}(W, \sigma_n^2)$ con respecto a W.

Los términos $-\frac{N}{2} \ln(2\pi)$ y $-\frac{N}{2} \ln(\sigma_n^2)$ no dependen de W, por lo que su derivada es cero.

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial}{\partial W}\left(-\frac{1}{2\sigma_n^2} (t - \Phi W)^T (t - \Phi W)\right)$$
$$= -\frac{1}{2\sigma_n^2} \frac{\partial}{\partial W} (t^T t - 2 W^T \Phi^T t + W^T \Phi^T \Phi W)$$
$$= -\frac{1}{2\sigma_n^2} (-2\Phi^T t + 2\Phi^T \Phi W)$$
$$= \frac{1}{\sigma_n^2} (\Phi^T t - \Phi^T \Phi W)$$

Igualamos la derivada a cero y resolver para W

$$\frac{1}{\sigma_n^2} (\Phi^T t - \Phi^T \Phi W) = 0$$

Asumiendo $\sigma_n^2 \neq 0$:

$$\Phi^T t - \Phi^T \Phi W = 0$$
$$\Phi^T \Phi W = \Phi^T t$$

$$W_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T t$$

Notamos que $W_{MLE} = W_{OLS}$

Solución para σ_η^2, MLE

Calculamos la derivada de $\mathcal{L}(\mathbf{W}, \sigma_\eta^2)$ con respecto a σ_η^2

$$\mathcal{L}(\mathbf{W}, \sigma_\eta^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_\eta^2) - \frac{1}{2\sigma_\eta^2} SSE(\mathbf{W})$$

donde $SSE(\mathbf{W}) = (\mathbf{t} - \mathbf{\Phi}\mathbf{W})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{W})$.

$$\frac{\partial \mathcal{L}}{\partial \sigma_\eta^2} = -\frac{N}{2}\frac{1}{\sigma_\eta^2} - \left(-\frac{1}{2(\sigma_\eta^2)^2}\right) SSE(\mathbf{W})$$

$$= -\frac{N}{2\sigma_\eta^2} + \frac{SSE(\mathbf{W})}{2(\sigma_\eta^2)^2}$$

Igualamos a cero y resolvemos para σ_η^2

$$-\frac{N}{2\sigma_\eta^2} + \frac{SSE(\mathbf{W})}{2(\sigma_\eta^2)^2} = 0$$

Multiplicando por $2(\sigma_\eta^2)^2$

$$-N\sigma_\eta^2 + SSE(\mathbf{W}) = 0$$
$$N\sigma_\eta^2 = SSE(\mathbf{W})$$

Solucionamos para σ_η^2, MLE

$$\sigma_\eta^2, MLE = \frac{SSE(\mathbf{W}_{MLE})}{N}$$

$$= \frac{1}{N}(\mathbf{t} - \mathbf{\Phi}\mathbf{W}_{MLE})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{W}_{MLE})$$

$$= \frac{1}{N}\sum_{n=1}^{N}(t_n - \phi(\mathbf{x}_n)^T \mathbf{W}_{MLE})^2$$

4. Máximo a-posteriori (Maximun a-posteriori - MAP)

Problema de optimización:

En la estimación MAP, incorporamos conocimiento previo sobre los parámetros $\mathbb{W}$ a través de una distribución a priori $p(\mathbb{W})$.

El objetivo es maximizar la probabilidad a posteriori $p(\mathbb{W} \mid \mathbf{t}, \mathbb{X})$, que por el teorema de Bayes, es proporcional al producto de la verosimilitud y la prior:

$$p(\mathbb{W} \mid \mathbf{t}, \mathbb{X}) \underbrace{\propto}_{\text{"es proporcional a"}} p(\mathbf{t} \mid \mathbb{X}, \mathbb{W})\, p(\mathbb{W})$$

Maximizar la posterior es equivalente a maximizar su logaritmo:

$$\ln p(\mathbb{W} \mid \mathbf{t}, \mathbb{X}) = \ln p(\mathbf{t} \mid \mathbb{X}, \mathbb{W}) + \ln p(\mathbb{W}) + \text{constante}$$

Asumimos una prior Gaussiana para $\mathbb{W}$ con media cero y covarianza isotrópica $\alpha^{-1}\mathbb{I}_Q$:

$$p(\mathbb{W} \mid \alpha) = \mathcal{N}(\mathbb{W} \mid 0, \alpha^{-1}\mathbb{I}_Q)$$
$$= \left(\frac{\alpha}{2\pi}\right)^{Q/2} \exp\left(-\frac{\alpha}{2}\mathbb{W}^T\mathbb{W}\right)$$

La log-prior es:

$$\ln p(\mathbb{W} \mid \alpha) = \frac{Q}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{W}^T\mathbb{W}$$

El log-posterior (ignorando constante que no dependen de $\mathbb{W}$), es:

$$\ln p(\mathbb{W} \mid \mathbf{t}, \mathbb{X}, \sigma_n^2, \alpha) \propto -\frac{1}{2\sigma_n^2}(\mathbf{t} - \Phi\mathbb{W})^T(\mathbf{t} - \Phi\mathbb{W}) - \frac{\alpha}{2}\mathbb{W}^T\mathbb{W}$$

Maximizar esta cantidad es equivalente a minimizar su negación

$$J_{MAP}(\mathbb{W}) = \frac{1}{2\sigma_n^2}(\mathbf{t} - \Phi\mathbb{W})^T(\mathbf{t} - \Phi\mathbb{W}) + \frac{\alpha}{2}\mathbb{W}^T\mathbb{W}$$

Si Multiplicamos por $2\sigma_n^2$ (asumiendo σ_n^2 es una constante conocida), minimizar esto:

$$(\mathbf{t} - \Phi\mathbb{W})^T(\mathbf{t} - \Phi\mathbb{W}) + \alpha\sigma_n^2\mathbb{W}^T\mathbb{W}$$

Esto es idéntico a la función de costo de la Regresión Ridge si definimos $\lambda = \alpha \sigma_\eta^2$.

Tomamos la derivada de la función objetivo (negativo de log-posterior sin constantes) con respecto a $\mathbf{w}$.

Sea $f(\mathbf{w}) = \frac{1}{2\sigma_\eta^2} (\mathbf{t} - \Phi \mathbf{w})^T (\mathbf{t} - \Phi \mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$

$$\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{2\sigma_\eta^2} (-2 \Phi^T \mathbf{t} + 2 \Phi^T \Phi \mathbf{w}) + \frac{\alpha}{2} (2\mathbf{w})$$

$$= -\frac{1}{\sigma_\eta^2} \Phi^T \mathbf{t} + \frac{1}{\sigma_\eta^2} \Phi^T \Phi \mathbf{w} + \alpha \mathbf{w}$$

Igualamos a cero y resolvemos para $\mathbf{w}$

$$-\frac{1}{\sigma_\eta^2} \Phi^T \mathbf{t} + \frac{1}{\sigma_\eta^2} \Phi^T \Phi \mathbf{w} + \alpha \mathbf{w} = 0$$

Multiplicamos por σ_η^2

$$-\Phi^T \mathbf{t} + \Phi^T \Phi \mathbf{w} + \alpha \sigma_\eta^2 \mathbf{w} = 0$$

$$\Phi^T \Phi \mathbf{w} + \alpha \sigma_\eta^2 \mathbf{w} = \Phi^T \mathbf{t}$$

$$(\Phi^T \Phi + \alpha \sigma_\eta^2 \mathbb{I}) \mathbf{w} = \Phi^T \mathbf{t}$$

Obtenemos la solución para $\mathbf{w}_{MAP}$

$$\mathbf{w}_{MAP} = (\Phi^T \Phi + \alpha \sigma_\eta^2 \mathbb{I})^{-1} \Phi^T \mathbf{t}$$

Nótese que es idéntico a $\mathbf{w}_{Ridge}$ con $\lambda = \alpha \sigma_\eta^2$

Scribe

5. Modelo Bayesiano con Modelo Lineal Gaussiano

En un enfoque Bayessiano completo, no se busca una estimación puntual de $\mathbb{W}$, si no que calcular la posterior completa

$$p(\mathbb{W} \mid \mathbb{t}, \mathbb{X}, \sigma_\eta^2, \alpha).$$

La verosimilitud es $p(\mathbb{t} \mid \mathbb{X}, \mathbb{W}, \sigma_\eta^2) = \mathcal{N}(\mathbb{t} \mid \Phi \mathbb{W}, \sigma_\eta^2 \mathbb{I}_N)$.
la prior es $P(\mathbb{W} \mid \alpha) = \mathcal{N}(\mathbb{W} \mid 0, \alpha^{-1} \mathbb{I}_Q)$.

El logaritmo de la posterior es proporcional a:

$$\ln p(\mathbb{W} \mid \mathbb{t}, \mathbb{X}, \sigma_\eta^2, \alpha) \propto -\frac{1}{2\sigma_\eta^2} (\mathbb{t} - \Phi \mathbb{W})^T (\mathbb{t} - \Phi \mathbb{W}) - \frac{\alpha}{2} \mathbb{W}^T \mathbb{W}$$

Expandiendo los términos cuadráticos en $\mathbb{W}$

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

Así, el exponente de la posterion (Mutiplicado por $-\frac{1}{2}$) es:

$$\frac{1}{\sigma_\eta^2} (\mathbb{t}^T \mathbb{t} - 2\mathbb{t}^T \Phi \mathbb{W} + \mathbb{W}^T \Phi^T \Phi \mathbb{W}) + \alpha \mathbb{W}^T \mathbb{W}$$

$$\frac{1}{\sigma_\eta^2} \mathbb{W}^T \Phi^T \Phi \mathbb{W} + \alpha \mathbb{W}^T \mathbb{W} - \frac{2}{\sigma_\eta^2} \mathbb{t}^T \Phi \mathbb{W} + \frac{1}{\sigma_\eta^2} \mathbb{t}^T \mathbb{t}$$

$$\mathbb{W}^T \left(\frac{1}{\sigma_\eta^2} \Phi^T \Phi + \alpha \mathbb{I}_Q \right) \mathbb{W} - 2\mathbb{W}^T \left(\frac{1}{\sigma_\eta^2} \Phi^T \mathbb{t} \right) + \text{costantes.}$$

Esta es la forma del exponente de una Gaussiana multivariada

$$\mathcal{N}(\mathbb{W} \mid \mu_N, \Sigma_N),$$

cuyo logaritmo (ignorando constantes) es

$$-\frac{1}{2} (\mathbb{W} - \mu_N)^T \Sigma_N^{-1} (\mathbb{W} - \mu_N) = -\frac{1}{2} (\mathbb{W}^T \Sigma_N^{-1} \mathbb{W} - 2\mathbb{W}^T \Sigma_N^{-1} \mu_N + \mu_N^T \Sigma_N^{-1} \mu_N)$$

Comparando los términos cuadráticos en $\mathbb{W}$ (el término $\mathbb{W}^T A \mathbb{W}$), identificamos la precisión (inversa de la covarianza) de la posterior:

$$\Sigma_N^{-1} = \frac{1}{\sigma_\eta^2} \Phi^T \Phi + \alpha \mathbb{I}_Q$$

la covarianza de la posterior es:

$$\Sigma_N = (\alpha \mathbb{I}_Q + \frac{1}{\sigma_\eta^2} \Phi^T \Phi)^{-1}$$

Comparando los términos lineales en $\mathbb{W}$ (el término $-2\mathbb{W}^T b$), identificamos:

$$\Sigma_N^{-1} \mu_N = \frac{1}{\sigma_\eta^2} \Phi^T \mathbb{t}$$

la media de la posterior es:

$$\mu_N = \Sigma_N \left(\frac{1}{\sigma_\eta^2} \Phi^T \mathbb{t} \right)$$

$$\mu_N = \left(\alpha \mathbb{I}_Q + \frac{1}{\sigma_\eta^2} \Phi^T \Phi \right)^{-1} \frac{1}{\sigma_\eta^2} \Phi^T \mathbb{t}$$

La distribución posterior para $\mathbb{W}$ es Gaussiana:

$$p(\mathbb{W} \mid \mathbb{t}, \mathbb{X}, \sigma_\eta^2, \alpha) = \mathcal{N}(\mathbb{W} \mid \mu_N, \Sigma_N)$$

con:

- Media: $\mu_N = (\Phi^T \Phi + \alpha \sigma_\eta^2 \mathbb{I}_Q)^{-1} \Phi^T \mathbb{t}$ (que es igual a W_{MAP})
- Covarianza: $\Sigma_N = \left(\alpha \mathbb{I}_Q + \frac{1}{\sigma_\eta^2} \Phi^T \Phi \right)^{-1}$

Prediccion Bayesiana:

Para una nueva entrada $\mathbb{X}_*$, la distribución predictiva para $t_* = \phi(\mathbb{X}_*)^T \mathbb{W} + \eta_*$ se obtiene integrando sobre la posterior de $\mathbb{W}$:

$$p(t_* \mid \mathbb{X}_*, \mathbb{t}, \mathbb{X}, \sigma_\eta^2, \alpha) = \int p(t_* \mid \mathbb{X}_*, \mathbb{W}, \sigma_\eta^2) \, p(\mathbb{W} \mid \mathbb{t}, \mathbb{X}, \sigma_\eta^2, \ldots) \, d\mathbb{W}$$

Esto resulta en una Gaussiana $\mathcal{N}(t_* \mid \mu_*, \sigma_*^2)$ con:

- Media predictiva: $\mu_* = \phi(\mathbb{X}_*)^T \mu_N$
- Varianza predictiva: $\sigma_*^2 = \sigma_\eta^2 + \phi(\mathbb{X}_*)^T \Sigma_N \phi(\mathbb{X}_*)$

6. Regresión Rigida Kernel (Kernel Ridge Regression - KRR)

Partiendo de la regresión ridge $\mathbb{W} = (\Phi^T \Phi + \lambda \mathbb{I}_Q)^{-1} \Phi^T \mathbb{t}$

Se puede demostrar que $\mathbb{W}$ se puede expresar como una combinación lineal de las transformaciones de las características de entrada: $\mathbb{W} = \Phi^T \mathbb{a}$ partir de algún vector $\mathbb{a} \in \mathbb{R}^N$.
Usando la identidad matricial:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} y = P B^T (R + B P B^T)^{-1} y.$$

Aplicandola a $\mathbb{W}$

$$\mathbb{W} = (\lambda \mathbb{I}_Q + \Phi^T \mathbb{I}_N^{-1} \Phi)^{-1} \Phi^T \mathbb{I}_N^{-1} \mathbb{t} \quad (\text{con } P = \lambda^{-1} \mathbb{I}_Q, B = \Phi, R = \mathbb{I}_N, y = \mathbb{t});$$

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

Definimos la matriz Kernel (o Gram) $\mathbb{K} = \Phi \Phi^T$. Los elementos de $\mathbb{K}$ son $K_{ij} = \phi(\mathbb{x}_i)^T \phi(\mathbb{x}_j) = k(\mathbb{x}_i, \mathbb{x}_j)$, donde k es la función Kernel, Entonces:

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

Si definimos:

$$\mathbb{a} = (\mathbb{K} + \lambda \mathbb{I}_N)^{-1} \mathbb{t}$$

Entonces,

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

Si definimos $\mathbb{a} = (\mathbb{K} + \lambda \mathbb{I}_N)^{-1} \mathbb{t}$, entonces $\mathbb{W} = \Phi^T \mathbb{a}$

Problema de optimización (en términos de $\mathbb{a}$):

La función de costo original es:

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

Sustituyendo $\mathbb{W} = \Phi^T \mathbb{a}$:

$$\begin{aligned} J_{KRR} &= (\mathbb{t} - \Phi\Phi^T \mathbb{a})^T (\mathbb{t} - \Phi\Phi^T \mathbb{a}) + \lambda (\Phi^T \mathbb{a})^T (\Phi^T \mathbb{a}) \\ &= (\mathbb{t} - \mathbb{K}\mathbb{a})^T (\mathbb{t} - \mathbb{K}\mathbb{a}) + \lambda \mathbb{a}^T \Phi \Phi^T \mathbb{a} \\ &= (\mathbb{t} - \mathbb{K}\mathbb{a})^T (\mathbb{t} - \mathbb{K}\mathbb{a}) + \lambda \mathbb{a}^T \mathbb{K} \mathbb{a} \\ &= \mathbb{t}^T \mathbb{t} - 2 \mathbb{t}^T \mathbb{K} \mathbb{a} + \mathbb{a}^T \mathbb{K}^T \mathbb{K} \mathbb{a} + \lambda \mathbb{a}^T \mathbb{K} \mathbb{a} \end{aligned}$$

Dado que $\mathbb{K}$ es simétrica ($\mathbb{K}^T = \mathbb{K}$):

$$J_{KRR}(\mathbb{a}) = \mathbb{t}^T \mathbb{t} - 2 \mathbb{t}^T \mathbb{K} \mathbb{a} + \mathbb{a}^T \mathbb{K}\mathbb{K} \mathbb{a} + \lambda \mathbb{a}^T \mathbb{K} \mathbb{a}$$

Calculamos la derivada de $J_{KRR}(\mathbb{a})$ con respecto a $\mathbb{a}$

$$\frac{\partial J_{KRR}(\mathbb{a})}{\partial \mathbb{a}} = -2 \mathbb{K}^T \mathbb{t} + 2 \mathbb{K}^T \mathbb{K} \mathbb{a} + 2\lambda \mathbb{K}^T \mathbb{a}$$ (usando $\mathbb{K}^T$ por formalidad, aunque $\mathbb{K} = \mathbb{K}^T$)

$$\frac{\partial J_{KRR}(\mathbb{a})}{\partial \mathbb{a}} = -2 \mathbb{K} \mathbb{t} + 2 \mathbb{K}\mathbb{K} \mathbb{a} + 2\lambda \mathbb{K} \mathbb{a}$$

Igualamos la derivada a cero y resolvemos para $\mathbb{a}$

$$-2 \mathbb{K} \mathbb{t} + 2 \mathbb{K}\mathbb{K} \mathbb{a} + 2\lambda \mathbb{K} \mathbb{a} = 0$$

$$\begin{aligned} \mathbb{K}\mathbb{K}\mathbb{a} + \lambda \mathbb{K}\mathbb{a} &= \mathbb{K}\mathbb{t} \\ \mathbb{K}(\mathbb{K} + \lambda \mathbb{I}_N)\mathbb{a} &= \mathbb{K}\mathbb{t} \end{aligned}$$

Obtenemos la solución para $\mathbb{a}$, si $\mathbb{K}$ es invertible (no siempre es el caso, pero $(\mathbb{K} + \lambda \mathbb{I}_N)$ si lo es para $\lambda > 0$):

$$(\mathbb{K} + \lambda \mathbb{I}_N)\mathbb{a} = \mathbb{t}$$

$$\mathbb{a}_{KRR} = (\mathbb{K} + \lambda \mathbb{I}_N)^{-1} \mathbb{t}$$

La predicción para una nueva entrada $\mathbf{x}_*$ es:

$$t_* = \phi(\mathbf{x}_*)^T \mathbf{w} = \phi(\mathbf{x}_*)^T \Phi^T \mathbf{a} = \mathbb{k}(\mathbf{x}_*)^T \mathbf{a}$$

donde $\mathbb{k}(\mathbf{x}_*)$ es un vector de $N \times 1$ con elementos:

$$k_n(\mathbf{x}_*) = k(\mathbf{x}_n, \mathbf{x}_*).$$

7. Procesos Gaussianos (Gaussian Processes - GP)

Un Proceso Gaussiano es una colección de variables aleatorias, cualquier subconjunto finito de las cuales tiene una distribución Gaussiana conjunta. Un GP define una distribución sobre funciones $f(x)$.

Asumimos que:

$$t_n = f(\mathbf{x}_n) + \eta_n, \quad \text{con } \eta_n \sim \mathcal{N}(0, \sigma_\eta^2).$$

Un GP se especifica por una función de media $m(\mathbf{x})$ (a menudo $m(\mathbf{x})=0$) y una función de covarianza (kernel) $k(\mathbf{x}, \mathbf{x}')$.

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Las salidas observadas $\mathbf{t} = [t_1, \ldots, t_N]^T$, dado que $f(\mathbf{x}) + \eta$ sigue una distribución Gaussiana:

$$\mathbf{t} \mid \mathbf{X} \sim \mathcal{N}(0, \mathbb{K} + \sigma_\eta^2 \mathbb{I}_N)$$

donde $\mathbb{K}$ es la matriz kernel $\mathbb{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Sea $\mathbf{C} = \mathbb{K} + \sigma_\eta^2 \mathbb{I}_N$.

Problema de optimización (Hiperparámetros)

La "optimización" en GPs se refiere a encontrar los hiperparámetros θ del kernel k y la varianza del ruido σ_η^2. Esto se hace maximizando la log-verosimilitud marginal (o evidencia):

$$\ln p(\mathbf{t} \mid \mathbf{X}, \theta) = -\frac{1}{2} \mathbf{t}^T \mathbf{C}(\theta)^{-1} \mathbf{t} - \frac{1}{2} \ln |\mathbf{C}(\theta)| - \frac{N}{2} \ln(2\pi)$$

Scribe

Se usan métodos de optimización basados en gradiente para encontrar θ que maximice esta expresión.

La derivada con respecto a un hiperparámetro θ_j es:

$$\frac{\partial \ln p(\mathbf{t}|\theta)}{\partial \theta_j} = \frac{1}{2} \mathbf{t}^T \mathbb{C}^{-1} \frac{\partial \mathbb{C}}{\partial \theta_j} \mathbb{C}^{-1} \mathbf{t} - \frac{1}{2} \operatorname{tr}\left(\mathbb{C}^{-1} \frac{\partial \mathbb{C}}{\partial \theta_j}\right)$$

Si definimos $\alpha_{GP} = \mathbb{C}^{-1}\mathbf{t}$:

$$\frac{\partial \ln p(\mathbf{t}|\theta)}{\partial \theta_j} = \frac{1}{2} \operatorname{tr}\left(\left(\alpha_{GP} \alpha_{GP}^T - \mathbb{C}^{-1}\right) \frac{\partial \mathbb{C}}{\partial \theta_j}\right)$$

Predicción con GPs:

Para predecir el valor de la función latente $f_* = f(\mathbb{X}_*)$ en un nuevo punto $\mathbb{X}_*$, consideramos la distribución conjunta $\mathbf{t}$ y f_*:

$$\begin{pmatrix} \mathbf{t} \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \mathbb{K} + \sigma_\eta^2 \mathbb{I}_N & \mathbb{K}_* \\ \mathbb{K}_*^T & K_{**} \end{pmatrix}\right) = \mathcal{N}\left(0, \begin{pmatrix} \mathbb{C} & \mathbb{K}_* \\ \mathbb{K}_*^T & K_{**} \end{pmatrix}\right)$$

donde $\mathbb{K}_*$ es un vector de $N \times 1$ con elementos $k(\mathbb{X}_n, \mathbb{X}_*)$, y $K_{**} = k(\mathbb{X}_*, \mathbb{X}_*)$. La distribución predictiva condicionada $p(f_*|t, \mathbb{X}, \mathbb{X}_*)$ es gaussiana con:

- Media: $\bar{f}_* = E[f_*|\mathbf{t}] = \mathbb{K}_*^T \left(\mathbb{K} + \sigma_\eta^2 \mathbb{I}_N\right)^{-1} \mathbf{t} = \mathbb{K}_*^T \mathbb{C}^{-1} \mathbf{t}$
- Varianza: $\operatorname{Var}(f_*) = K_{**} - \mathbb{K}_*^T \left(\mathbb{K} + \sigma_\eta^2 \mathbb{I}_N\right)^{-1} \mathbb{K}_* = K_{**} - \mathbb{K}_*^T \mathbb{C}^{-1} \mathbb{K}_*$

se se predice

$t_* = f_* + \eta_*$, la media es la misma, $\bar{t}_* = \bar{f}_*$

y la varianza es

$$\operatorname{Var}(t_*) = \operatorname{Var}(f_*) + \sigma_\eta^2$$

Discuciones

- OLS, MLE, Ridge, MAP: Producen una estimación puntual para los pesos $\mathbb{W}$.

- El modelo Bayesiano Lineal, GP: Producen una distribución de probabilidad completa sobre los pesos o directamente sobre las funciones. Permitiendo cuantificar la incertidumbre

- OLS, MLE: No incorporan regularización explicita

- Ridge, MAP: introducen regularización L2 lo una prior Gaussiana, lo que ayuda a prevenir el sobreajuste y estabilizar la solución.

- Modelo Bayesiano: en este modelo la prior actúa como forma de regularización

- GP: la elección del Kernel y sus hiperparámetros, junto con el término de ruido, inherentemente regularizan el modelo y controlan su suavidad y complejidad