

PARCIAL 1: Teoría de Aprendizaje de Máquina.

Punto #1

Sea el modelo de regresión $t_n = \phi(X_n)w^\top + \eta_n$, con $\{t_n \in \mathbb{R}, X_n \in \mathbb{R}^P\}_{n=1}^N$, $w \in \mathbb{R}^Q$, $\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$, $Q \geq P$, y $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$. Presente el problema de optimización y la solución del mismo para los modelos mínimos cuadrados, mínimos cuadrados regularizados, máxima verosimilitud, máximo a posteriori, Bayesiano (con modelo lineal Gaussiano), regresión rígida Kernel y mediante procesos Gaussiano. Asuma datos i.i.d. Discuta las diferencias y similitudes entre los modelos estudiados.

Solución

• Planteamiento del modelo

Se considera el modelo de regresión: $t_n = \phi(X_n)w^\top + \eta_n$, $n = 1, \dots, N$.

donde cada $X_n \in \mathbb{R}^P$ representa el vector de entrada, $t_n \in \mathbb{R}$ es la salida observada, $w \in \mathbb{R}^Q$ son los parámetros del modelo. Se asume ruido η_n gaussiano e i.i.d $\rightarrow \eta_n \sim N(0, \sigma_n^2)$.

• Estimador por mínimos cuadrados

Para encontrar el vector de parámetros w que mejor se ajusta a los datos observados y bajo el supuesto de ruido gaussiano. Se debe minimizar el error cuadrático total, de tal forma:

$$w^* = \underset{w}{\operatorname{argmin}} \|t - \phi w^\top\|_2^2$$

Expandiendo el término cuadrático: $\langle t - \phi w^\top, t - \phi w^\top \rangle$

$$\langle t, t \rangle - \langle t, \phi w^\top \rangle - \langle \phi w^\top, t \rangle + \langle \phi w^\top, \phi w^\top \rangle$$

$$\text{obteniendo: } \|t - \phi w^\top\|_2^2 = t^\top t - t^\top \phi w^\top - (\phi w^\top)^\top t + w^\top \phi^\top \phi w^\top$$

$$\text{Teniendo en cuenta } \rightarrow \hat{t} = \phi w^\top$$

$$\text{Entonces: } w^* = t^\top t - 2t^\top \phi w^\top + w^\top \phi^\top \phi w^\top$$

$$\text{Derivando respecto a } w \Rightarrow \frac{\partial w^*}{\partial w} [t^\top t - 2t^\top \phi w^\top + w^\top \phi^\top \phi w^\top] = -2t^\top \phi + (2\phi^\top \phi w^\top)^\top$$

$$\text{Igualando a cero } \Rightarrow -2t^\top \phi + 2w^\top \phi^\top \phi = 0 \Rightarrow 2w^\top \phi^\top \phi = 2t^\top \phi$$

Despejando w , la solución para el problema de optimización es:

$$w = t^\top \phi (\phi^\top \phi)^{-1} \Rightarrow w^\top = (\phi^\top \phi)^{-1} \phi^\top t \quad \text{Si } \phi^\top \phi \text{ es invertible.}$$

- **Estimados por mínimos cuadrados regularizados**

Cuando las columnas de Φ son linealmente dependientes o el número de parámetros es muy grande, la matriz $\Phi^T \Phi$ puede no ser invertible o generar soluciones con gran varianza. Por lo tanto, se introduce una regularización que penaliza la magnitud de los parámetros w y estabiliza la solución.

Incorporando un término adicional a la norma cuadrática. Entonces:

$$w^* = \underset{w}{\operatorname{argmin}} (||t - \Phi w^T||_2^2 + \lambda ||w||_2^2) \quad \text{donde } \lambda > 0 \text{ es el parámetro de regularización}$$

Se expande la función de costo agregando el nuevo término y derivando respecto a w se obtiene:

$$\frac{\partial w^*}{\partial w} = -2t^T \Phi + 2w\Phi^T \Phi + 2\lambda w$$

$$\text{Igualando a } 0 \Rightarrow -\frac{2t^T \Phi}{2} + \frac{2w\Phi^T \Phi}{2} + \frac{2\lambda w}{2} = 0$$

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

$$w\Phi^T \Phi + \lambda w = t^T \Phi \quad \text{Teniendo en cuenta } \Phi^T \Phi = I \text{ donde } I \text{ es la matriz}$$

$$w(\Phi^T \Phi + \lambda) = t^T \Phi \quad \text{Identidad de dimensión } Q \times Q$$

$$\Rightarrow w \underbrace{(\Phi^T \Phi + \lambda)}_{I} (\Phi^T \Phi + \lambda)^{-1} = t^T \Phi (\Phi^T \Phi + \lambda)^{-1}$$

Despejando w , la solución para este problema de optimización es:

$$w = t^T \Phi (\Phi^T \Phi + \lambda)^{-1} \Rightarrow w^T = (\Phi^T \Phi + \lambda)^{-1} \Phi^T t$$

- **Solución por máxima verosimilitud.**

$\eta_n \sim N(0, \sigma_n^2)$ es el término de ruido, asumido gaussiano e independiente. Por lo tanto la distribución gaussiana de las observaciones (datos) se puede expresar así:

$$P(t_n | \Phi(X_n)w^T, \sigma_n^2) \rightarrow \text{Probabilidad Gaussiana}$$

Dado un conjunto de datos i.i.d., la verosimilitud total se define como el producto de las probabilidades individuales

$$P(t | \Phi, w^T, \sigma_n^2) = \prod_{n=1}^N N(t_n | \Phi(X_n)w^T, \sigma_n^2)$$

Entonces, se tiene el siguiente planteamiento:

$$w^* = \underset{w, \sigma_n^2}{\operatorname{argmax}} \log \left(\prod_{n=1}^N N(t_n | \Phi(X_n)w^T, \sigma_n^2) \right)$$

Faltando el cálculo, se toma el logaritmo de la verosimilitud (log-verosimilitud):

$$w^* = w_{ML}, \hat{\sigma}_\eta^2 = \underset{w, \sigma_\eta^2}{\operatorname{argmax}} -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\sigma_\eta^2) - \frac{1}{2 \sigma_\eta^2} \|t - \phi w^\top\|_2^2$$

Así, la estimación por máxima verosimilitud es encontrar los parámetros w y σ^2 que maximizan la anterior expresión.

Derivando la log-verosimilitud con respecto a w , se obtiene:

$$\frac{\partial w^*}{\partial w} = \frac{1}{\hat{\sigma}_\eta^2} \phi^\top(t - \phi w)$$

Igualando a 0 se despeja w : $\frac{1}{\hat{\sigma}_\eta^2} \phi^\top(t - \phi w) = 0 \quad (\times \hat{\sigma}_\eta^2) \Rightarrow \phi^\top(t - \phi w) = 0$

$$\Rightarrow \phi^\top t - \phi^\top \phi^\top w = 0$$

$$\phi^\top \phi^\top w = \phi^\top t$$

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

De esta forma, $w_{ML} = (\phi^\top \phi)^{-1} \phi^\top t$ → Estimación de máxima verosimilitud para los pesos

Derivando la log-verosimilitud respecto a σ^2 , se obtiene:

$$\frac{\partial w^*}{\partial \sigma^2} = -\frac{N}{2} \frac{1}{\hat{\sigma}_\eta^2} + \frac{\|t - \phi w^\top\|_2^2}{2(\hat{\sigma}_\eta^2)^2} = 0$$

Se iguala a 0 y se despeja $\hat{\sigma}_\eta^2$:

$$\frac{\|t - \phi w^\top\|_2^2}{2(\hat{\sigma}_\eta^2)^2} - \frac{N}{2\hat{\sigma}_\eta^2} \Rightarrow \hat{\sigma}_\eta^2 = \frac{\|t - \phi w^\top\|_2^2}{N} \{ (t - \phi w)^\top (t - \phi w) \}$$

Obteniendo: $\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (t_n - \phi(x_n) w^\top)^2$ → Varianza estimada del ruido del modelo.

Con ello, la solución de predicción para un nuevo dato está dada por:

$$x^* = N(t_n | \phi^\top X_n w_{ML}^\top, \hat{\sigma}_{ML}^2)$$

• Solución pos. máximo a posteriori.

La probabilidad de observar t_n dado x_n y w está dada por:

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

Donde el valor observado t_n se distribuye normalmente alrededor de la predicción del modelo con una desviación σ_n .

Asumiendo datos i.i.d., la probabilidad conjunta de todas las observaciones es el producto de las probabilidades individuales:

$$p(t|w) = \prod_{n=1}^N N(t_n | \phi(x_n)^T w, \sigma_n^2) \rightarrow \text{Verosimilitud.}$$

Este método se nombra el teorema de Bayes de tal manera que busca calcular la distribución posterior de los parámetros, es decir, su probabilidad después de observar los datos:

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

Teorema de Bayes

- donde:
- $p(t|w) \rightarrow$ Verosimilitud: mide qué tan probables son los datos dados los parámetros
 - $p(w) \rightarrow$ Prior: conocimiento previo sobre los parámetros
 - $p(t) \rightarrow$ Evidencia / Normalizadora: Asegura que la probabilidad total sea 1 (Margen de probabilidad)

$p(t)$ no depende de w , por lo tanto se puede ignorar para la optimización.

Teniendo el modelo: $t_n = \phi(x_n)w^T + \eta_n$, entonces: $\eta_n = t_n - \phi(x_n)w^T$. La salida t_n se comporta como una variable aleatoria normal con media en la predicción del modelo $\phi(x_n)w^T$ y varianza igual a la del inicio.

Siguiendo la probabilidad definida al inicio y suponiendo i.i.d.:

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

Donde I , representa que los errores de distintas observaciones son independientes.

Tomando en cuenta que el prior (distribución previa de los pesos) se define como:

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

Se combina todo mediante el teorema de Bayes para construir el posterior y así:

$$p(w|t) \propto p(t|w)p(w) \Rightarrow p(w|t) \propto N(t | \phi w, \sigma_n^2 I) N(w | 0, \sigma_w^2 I)$$

Resolviendo a través de logaritmos, la función de verosimilitud se expresa como:

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

Por independencia de los datos:

$$P(t|w) = \prod_{n=1}^N P(t_n|w) = (2\pi\sigma_n^2)^{-N/2} \exp\left[-\frac{1}{2\sigma_n^2} \sum_{n=1}^N (t_n - \phi(x_n)w)^2\right]$$

Entonces:

$$P(t|w) = (2\pi\sigma_w^2)^{-N/2} \exp\left[-\frac{1}{2\sigma_w^2} \|t - \phi w\|_2^2\right]$$

Para el PRIOR: $P(w) = N(0, \sigma_w^2 I)$, por definición de la densidad gaussiana multivariada:

$$P(w) = (2\pi\sigma_w^2)^{-Q/2} \exp\left[-\frac{1}{2\sigma_w^2} \|w\|_2^2\right]$$

De forma proporcional: $P(w) \propto \exp\left(-\frac{1}{2\sigma_w^2} \|w\|_2^2\right)$ y para el modelo condicional:

$$P(w|t) \propto \left[(2\pi\sigma_w^2)^{-N/2} \exp\left(-\frac{1}{2\sigma_w^2} \|t - \phi w\|_2^2\right) \right]$$

Usando la propiedad de logaritmos: $\log(ab) = \log a + \log b$ y $\log(ex) = x$. Se tiene:

$$\log P(w|t) = -\frac{N}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} \|t - \phi w\|_2^2 - \frac{Q}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} \|w\|_2^2$$

$$\Rightarrow \log P(w|t) = -\frac{1}{2\sigma_w^2} \|t - \phi w\|_2^2 - \frac{1}{2\sigma_w^2} \|w\|_2^2$$

Para el problema de optimización:

$$w_{MAP} = \arg \max_w \frac{-1}{2\sigma_w^2} \|t - \phi w\|_2^2 - \frac{1}{2\sigma_w^2} \|w\|_2^2$$

Esto es equivalente a minimizar el negativo del logaritmo posterior:

$$w_{MAP} = \arg \min_w \|t - \phi w\|_2^2 + \lambda \|w\|_2^2 \quad \text{donde } \lambda = \frac{\sigma_w^2}{\sigma_n^2}$$

$$\text{Expandiendo: } w_{MAP} = \arg \min_w (t - \phi w)^T (t - \phi w) + \lambda w^T w$$

$$\text{Desarrollando: } w_{MAP} = t^T t - 2t^T \phi w + w^T \phi^T \phi w + \lambda w^T w$$

$$\text{Derivando con respecto a } w: \frac{\partial w_{MAP}}{\partial w} = -2\phi^T t + 2\phi^T \phi w + 2\lambda w$$

$$\text{Igualando a cero: } -2\phi^T t + 2\phi^T \phi w + 2\lambda w = 0 \Rightarrow (\phi^T \phi + \lambda I) w = \phi^T t$$

Se despeja w y se obtiene que la solución al problema de optimización es:

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

Solución por Bayesiano con modelo lineal Gaussiano

El prior para el modelo lineal gaussiano está definido por: $p(w) = N(w | M_0, S_0)$

Donde $M_0 = 0$, $S_0 = \sigma_w^2 I$. Entonces $p(w) = N(w | 0, \sigma_w^2 I)$ y $p(y | x, w) = N(y | Ax, \sigma_y^2 I)$ (con: $\sigma_y^2 = \sigma_w^2$, $\sigma_w^2 = \alpha^{-1}$)

Antes de ver los datos, se tiene la priori; después de ver los datos se obtiene la posteriori.

Resolviendo con la fórmula de la normal:

$$p(x) = N(x | \mu, \Lambda^{-1}) \Rightarrow p(x) \propto \exp\left(-\frac{1}{2}(x - \mu)^T \Lambda (x - \mu)\right) \rightarrow \text{PRIOR}$$

$$\text{Verosimilitud: } p(y|x) = N(y | Ax + b, L^{-1}) \propto \exp\left(-\frac{1}{2}(y - Ax - b)^T L (y - Ax - b)\right)$$

y ya que $p(x|y) \propto p(y|x)p(x)$ se simplifica con algoritmos para encontrar el problema de optimización.

$$-2 \log p(x|y) = (y - Ax - b)^T L (y - Ax - b) + (x - \mu)^T \Lambda (x - \mu) + \text{cte} \rightarrow \text{POSTERIOR}$$

Expandiendo y agrupando:

$$\rightarrow (y - b)^T L (y - b) - 2x^T A^T L (y - b) + x^T A^T L A x + x^T \Lambda x - 2\mu^T \Lambda x - y^T \Lambda \mu + \text{cte}$$

Agrupando por potencias de x :

- Término cuadrático: $\underbrace{x^T x (\Lambda^T L A + \Lambda)}_{\Sigma^{-1} x^T y}$

La matriz Σ^{-1} (predicción posterior) representa la suma de las influencias de la prior y los datos observados. Es decir, la combinación de ambas fuentes de información.

- Término lineal: $\underbrace{x^T (A^T L (y - b) + \Lambda \mu)}_{\Sigma^{-1} x^T y}$

Mejora la evidencia de los datos y con el conocimiento previo μ . Donde $\Sigma^{-1} x^T y | y$ es el punto donde el exponente cuadrático alcanza su mínimo.

$\Sigma^{-1} y$ → Matriz de covarianza del posterior dado y : mide la confianza en la estimación

Σ^{-1}_{xy} → Inversa llamada matriz de precisión. Refleja el grado de certeza (+ precisión, - incertidumbre)

μ_{xy} → Vector de media de la distribución posterior. Solución óptima

$$\hookrightarrow \Sigma_{xy}^{-1} (A^T L (y - b) + \Lambda \mu) \quad \text{Ec. 1.} \quad y \quad \Sigma_{xy}^{-1} = (\Lambda + A^T L A)^{-1} \quad \text{Ec. 2}$$

Aplicando el modelo de regresión lineal bayesiana $\rightarrow t_n = \phi(x_n)w^T + \eta_n$, $p(\eta_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\eta^2}{2\sigma^2}}$

La verosimilitud se puede escribir como:

$P(t_n | w) = N(t_n | \phi(x_n)w^T, \sigma^2) \rightarrow$ la salida t_n se distribuye de forma gaussiana alrededor de la predicción lineal $\phi(x_n)^T w$ con varianza σ^2 .

El prior es: $p(w) = N(w | m_0, S_0) \rightarrow m_0 = 0, S_0 \rightarrow \alpha^{-1} I = \sigma_w^2 I$.

El posterior es: $p(w | t) \propto p(t|w)p(w) = N(w | m_N, S_N)$ Donde $m_N \rightarrow$ Media
 $S_N \rightarrow$ varianza

Y con las ecuaciones (Ec. 1. y Ec. 2) se obtiene:

- Media posterior: $m_N = S_N \left(S_0^{-1} m_0 + \frac{1}{\sigma^2} \phi^T t \right)$ Donde: $x \rightarrow w$ $I \rightarrow \frac{1}{\sigma^2} I$
 $y \rightarrow t$
 $A \rightarrow \phi$
 $b \rightarrow 0$
 $\Lambda \rightarrow S_0^{-1}$
- Covarianza posterior: $S_N = \left(S_0^{-1} + \frac{1}{\sigma^2} \phi^T \phi \right)^{-1}$

Donde la media posterior corresponde al punto que maximiza la posterior

Con ello, el problema de optimización aplicado al modelo es.

$$\arg \min_w \frac{1}{2\sigma^2} \|t - \phi w\|_2^2 + \frac{1}{2} (w - m_0)^T S_0^{-1} (w - m_0)$$

Sustituyendo $m_0 = 0$ y $S_0 = \sigma_w^2 I$, entonces:

$$w^* = \frac{1}{2\sigma^2} \|t - \phi w\|_2^2 + \frac{1}{2} w^T (\sigma_w^2 I)^{-1} w = \frac{1}{2\sigma^2} \|t - \phi w\|_2^2 + \frac{1}{2\sigma^2} w^T w$$

$$\text{Expandiendo } \|t - \phi w\|_2^2 \Rightarrow (t - \phi w)^T (t - \phi w) = t^T t - 2w^T \phi^T t + w^T \phi \phi w$$

$$\text{Se reescribe: } w^* = \frac{1}{2\sigma^2} (t^T t - 2w^T \phi^T t + w^T \phi \phi w) + \frac{1}{2\sigma^2} w^T w$$

$$w^* = \frac{1}{2\sigma^2} t^T t - \frac{1}{2\sigma^2} w^T \phi^T t + \frac{1}{2\sigma^2} w^T \phi \phi w + \frac{1}{2\sigma^2} w^T w$$

$$\text{Derivando con respecto a } w: \frac{\partial w^*}{\partial w} = \frac{1}{\sigma^2} \phi^T \phi w + \frac{1}{\sigma^2} w - \frac{1}{\sigma^2} \phi^T t$$

$$\Rightarrow \frac{\partial w^*}{\partial w} = \left(\frac{1}{\sigma^2} \phi^T \phi + \frac{1}{\sigma^2} I \right) w - \frac{1}{\sigma^2} \phi^T t$$

$$\text{Igualando a cero: } \left(\frac{1}{\sigma^2} \Phi^T \Phi + \frac{1}{\sigma_w^2} I \right) w - \frac{1}{\sigma^2} \Phi^T t = 0 \Rightarrow \left(\frac{1}{\sigma^2} \Phi^T \Phi + \frac{1}{\sigma_w^2} I \right) w = \frac{1}{\sigma^2} \Phi^T t$$

$$\text{Multiplicando a ambos lados por } \sigma^2 \Rightarrow (\Phi^T \Phi + \frac{\sigma^2}{\sigma_w^2} I) w = \Phi^T t$$

Se despeja w y la solución bayesiana con modelo lineal gaussiano corresponde a:

$$w^* = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t \quad \text{donde } \lambda = \frac{\sigma^2}{\sigma_w^2}$$

- Solución por regresión rígida kernel

La regresión rígida kernel es un método de aprendizaje supervisado que combina dos técnicas:

- Regularización (Ridge): Penaliza las soluciones complejas.
- Kernel: Permite trabajar en el espacio de características implícitos de muy alta dimensión.

$$\text{Se asume un espacio de características } \hat{f}(x) = \phi(x_n)^T w$$

La función objetivo debe medir qué tan bien el modelo se ajusta a los datos observados. Para ello se utiliza el error cuadrático medio así:

$$\text{Error} = \frac{1}{2} \|t - \Phi w\|_2^2$$

Adicionalmente, se añade un término de regularización que penaliza soluciones complejas:

$$\text{Regularización } L_2 = \frac{\lambda}{2} \|w\|_2^2 \quad (\text{Ridge})$$

Entonces, el problema de regresión rígida en el espacio original (primal) se formula como

$$w^* = \underset{w}{\operatorname{argmin}} \underbrace{\frac{1}{2} \|t - \Phi w\|_2^2}_{\text{Sesgo}} + \underbrace{\frac{\lambda}{2} \|w\|_2^2}_{\text{Varianza}}$$

- Sesgo: Precisión en los datos de entrenamiento.

- Varianza: Simplicidad para garantizar buena generalización.

$$\text{Derivando respecto a } w \Rightarrow \frac{\partial w^*}{\partial w} = -\Phi^T t + \Phi^T \Phi w + \lambda w$$

$$\begin{aligned} \text{Igualando a } 0 \text{ y despejando } w \Rightarrow & -\Phi^T t + \Phi^T \Phi w + \lambda w = 0 \\ & \Phi^T \Phi w + \lambda w = \Phi^T t \\ & (\Phi^T \Phi + \lambda I) w = \Phi^T t \end{aligned}$$

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

Para eludir problemas de alta dimensionalidad, se utiliza una representación, estableciendo la solución óptima como:

$$w^* = \Phi^T X^* \text{ donde } X \in \mathbb{R}^N \text{ es un nuevo vector de coeficientes}$$

Sustituyendo $w^* = \phi\alpha^*$ en la solución primal $\Rightarrow \phi^T\alpha^* = (\phi^T\phi + \lambda I)^{-1}\phi^T t$

Multiplicando ambos lados por $\phi \Rightarrow \phi\phi^T\alpha^* = \phi(\phi^T\phi + \lambda I)^{-1}\phi^T t$ Ec. 1.

Introduciendo la matriz kernel $K = \phi\phi^T$ ($N \times N$) $\Rightarrow K_{ij} = \phi(x_i)^T \phi(x_j) = k(x_i, x_j)$

Donde $k(x_i, x_j)$ es la función kernel que calcula el producto punto de dos ejemplos en el espacio de características,

Con la definición de $K = \phi\phi^T$ la expresión de Ec. 1 se convierte en:

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

Usando la identidad de matrices: $\phi(\phi^T\phi + \lambda I)^{-1}\phi^T = (K + \lambda I)^{-1}K$. Se obtiene:

$$K\alpha^* = (K + \lambda I)^{-1}Kt$$

Multiplicando ambos lados por $(K + \lambda I)$, entonces $\Rightarrow (K + \lambda I)K\alpha^* = Kt$

$$K^2\alpha^* + \lambda K\alpha^* = Kt$$

Factorizando $K \Rightarrow K(K\alpha^* + \lambda\alpha^*) = Kt$

Asumiendo que K es invertible $\Rightarrow K\alpha^* + \lambda\alpha^* = t$

Despejando α^* de $(K + \lambda I)\alpha^* = t$, se obtiene que la solución al problema de optimización es:

$$\boxed{\alpha^* = (K + \lambda I)^{-1}t}$$

• Solución mediante (proceso) gaussiano

Los procesos gaussianos especifican directamente un prior gaussiano sobre la función, entonces:

$$f(x) \sim N_p(m(x), K(x, x')) \quad \text{Donde } m(x) \rightarrow \text{función media a priori (tipicamente }= 0)$$

$K(x, x')$ → función kernel (función de covarianza)

Ya que cualquier conjunto finito de evaluación, sigue una distribución gaussiana multivariada

Los datos observados tienen ruido gaussiano aditivo: $t_n = f(x_n) + \eta_n, \eta_n \sim N(0, \sigma^2)$

En forma de verosimilitud, la distribución posterior $p(f|t) \propto p(t|f)p(f)$, es:

$$p(t|f) = N(t|f, \sigma^2 I), \text{ Sabiendo que el prior es: } p(f) = N(f|0, K)$$

La log-verosimilitud marginal corresponde a:

$$p(t|f) \propto \exp \left[-\frac{1}{2} (t - f)^T (\sigma^2 I) (t - f) - \frac{1}{2} f^T K^{-1} f \right]$$

Entonces, el problema de optimización se define como:

$$f^* = \min_f q(f) = \frac{1}{2\sigma^2} \|t - f\|_2^2 + \frac{1}{2} f^T K^{-1} f$$

Derivando respecto a $f \Rightarrow \frac{\partial f^*}{\partial f} = -\frac{1}{\sigma^2}(t - f) + K^{-1}f = 0$

Reorganizando $\Rightarrow \frac{1}{\sigma^2}t = \left(\frac{1}{\sigma^2}I + K^{-1}\right)f$

Multiplicando por $K \Rightarrow K\left(\frac{1}{\sigma^2}I + K^{-1}\right)f = K \cdot \frac{1}{\sigma^2}t \Rightarrow (K + \sigma^2 I)f = kt$

Despejando f se obtiene la solución al problema de optimización:

$$f^* = K(K + \sigma^2 I)^{-1}t$$

• Diferencias y similitudes entre modelos

Todos los modelos parten del mismo modelo lineal con ruido gaussiano $t_n = \phi(x_n)w^T + \eta_n$, y ajustados i.i.d. Buscando estimar los parámetros que mejor ajusten las observaciones.

El modelo de mínimos cuadrados minimiza el error cuadrático y equivale al método de máxima verosimilitud cuando el ruido es gaussiano. Ninguno de los dos incluye regularización.

El modelo de mínimos cuadrados regulados incorpora un término de penalización sobre w , reduciendo el sobreajuste, siendo así, equivalente al modelo máximo a posteriori (MAP) si se asume un prior gaussiano sobre los parámetros.

El método bayesiano generaliza MAP al considerar toda la distribución posterior $p(w|t)$, permitiendo cuantificar la incertidumbre en las predicciones. La regresión rígida kernel extiende el modelo lineal al espacio no lineal mediante $k(x_i, x_j)$ y los procesos gaussianos representan su versión completamente bayesiana, donde se define una distribución sobre funciones en lugar de sobre parámetros.

Todos los métodos comparten la misma base estadística, diferenciándose principalmente en el tratamiento de la regularización y el grado de incertidumbre que modelan.