Introducción

Queremos predecir un valor real $t_n \in \mathbb{R}$ asociado a una entrada $X_n \in \mathbb{R}^p$ usando un modelo lineal definido como

$$t_n = \phi(X_n) W^T + \eta_n$$

Nota: $\phi(x_n) \in \mathbb{R}^{1 \times Q}$ (vector fila)
$w \in \mathbb{R}^{1 \times Q} \rightarrow \vec{w} \in \mathbb{R}^{Q \times 1}$ (vector columna)
Luego
$\phi(x_n) \cdot w^T \in \mathbb{R}^{1 \times Q} \cdot \mathbb{R}^{Q \times 1} = \mathbb{R}^{1 \times 1}$

donde

- $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^Q$ es una función base
- $w \in \mathbb{R}^Q$ son los parámetros del modelo
- $\eta_n \sim \mathcal{N}(0, \sigma_n^2)$ representa ruido gaussiano.

Asumimos que los datos $\{(X_n, t_n)\}_{n=1}^{N}$ son independientes e identicamente distribuidos.

Mínimos cuadrados

Buscamos estimar el vector de parámetros w que minimiza el error cuadrático entre las predicciones del modelo y los valores observados t_n.

Función objetivo.

La función a minimizar es el error cuadrático medio total

$$J(w) = \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \Phi(x_n) w^T\right)^2$$

Definimos

$$\Phi \in \mathbb{R}^{N \times Q} = \begin{bmatrix} \Phi(x_1) \\ \Phi(x_2) \\ \vdots \\ \Phi(x_n) \end{bmatrix}$$

$$t \in \mathbb{R}^N = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

Re escribiendo

$$J(w) = \frac{1}{2} \| t - \Phi w^T \|_2^2$$

Luego queremos encontrar

$$w^* = \arg\min_w J(w) = \arg\min_w \frac{1}{2} \| t - \Phi w^T \|_2^2$$

Derivando e igualando a cero obtenemos

$$J(w) = \frac{1}{2} \| t - \Phi w^T \|_2^2$$

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

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

Igualando a cero

$$\frac{\partial J}{\partial w} = -\Phi^T (t - \Phi w^T) = 0$$

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

Despejando W (si $\Phi^T \Phi \in \mathbb{R}^{Q \times Q}$ es invertible)

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

Está solución proporciona el estimador lineal de mínimos cuadrados, que minimiza el error cuadrático sobre los datos observados. Es la base de modelos más avanzados como la máxima verosimilitud , la regresión regularizada o el enfoque bayesiano. Su simplicidad permite obtener una solución cerrada, pero es sensible al sobreajuste y a problemas de acondicionamiento si $\Phi^T \Phi$ no es invertible o está mal condicionada.

Mínimos cuadrados regularizados (Regresión Ridge)

Función objetivo

Se busca minimizar el error cuadrático más una penalización sobre la norma de los parámetros.

$$\frac{1}{2} \sum_{n=1}^{N}$$

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

En forma matricial:

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

donde:

- $\Phi \in \mathbb{R}^{N \times Q}$ es la matriz de diseño con $\phi(x_n)^T$ como filas.
- $t \in \mathbb{R}^N$ es el vector de salidas.
- $\lambda \geq 0$ es el hiperparámetro de regularización.

Luego el estimador de los parámetros w es

$$w^* = \operatorname{argmin}_w \left(\frac{1}{2} \|t - \Phi w\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 \right)$$

derivando e igualando a cero

$$\frac{\partial L}{\partial w} = -\Phi^T (t - \Phi w) + \lambda w = 0$$

$$\rightarrow \quad (\Phi^T \Phi + \lambda I) = \Phi^T t$$

de donde

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

La regresión Ridge agrega una penalización cuadrática sobre los parámetros al error de mínimos cuadrados. Esto mejora su estabilidad numérica cuando $\Phi^T \Phi$ está mal condicionada o no invertible, y reduce el riesgo de sobreajuste. Cuando $\lambda = 0$ se recupera la solución de mínimos cuadrados.

Máxima verosimilitud

Bajo el supuesto de ruido gaussiano, la salida t_n condicionada a la entrada x_n tiene distribución:

$$p(t_n | x_n, w) = \mathcal{N}(t_n | \phi(x_n)^T w, \sigma^2)$$

La verosimilitud conjunta de los datos

es

$$p(t \mid \Phi, w) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \phi(x_n)^T w, \sigma^2)$$

Luego, el estimador de máxima verosimilitud es

$$w^* = \underset{w}{\operatorname{argmax}} \; p(t \mid \Phi, w)$$

Maximizar la verosimilitud es equivalente a minimizar la log-verosimilitud negativa. Tomando logaritmo y descartando términos constantes:

$$L(w) = \frac{1}{2\sigma^2} \| t - \Phi w \|_2^2$$

Como σ^2 es constante, minimizar L equivale a minimizar el error cuadrático ordinario:

$$w^* = \underset{w}{\operatorname{argmin}} \; \frac{1}{2} \| t - \Phi w \|_2^2$$

Derivando e igualando a cero

$$\Phi^T \Phi w = \Phi t$$

$$\rightarrow \quad \boxed{w^* = (\Phi^T \Phi)^{-1} \Phi^T t}$$

La estimación por máxima verosimilitud produce el mismo estimador que mínimos cuadrados ordinarios, pero con una motivación probabilística: se busca el valor w que maximiza la probabilidad de observar los datos bajo el supuesto de ruido gaussiano.

Maximo a posteriori

Asumimos el mismo modelo lineal con ruido gaussiano:

$$t_n = \phi(x_n)^T w + \eta_n, \quad \eta_n \sim \mathcal{N}(0, \sigma^2)$$

Además, incorporamos una **distribución a priori** sobre los parámetros

$$p(w) = \mathcal{N}(w \mid 0, \tau^2 I)$$

Función objetivo

Se busca el valor de w que maximiza la distribución posterior

$$w^* = \arg\max_w \; p(w \mid t, \Phi)$$

donde

$$p(w \mid t, \Phi) = \frac{p(t \mid \Phi, w) \cdot p(w)}{p(t \mid \Phi)}$$

Como el denominador no depende de w, maximizar la posterior es equivalente a maximizar el producto de la verosimilitud y el prior:

$$w^* = \arg\max_w \; [\, p(t \mid \Phi, w) \cdot p(w) \,]$$

Tomando logaritmo y descartando constantes, obtenemos:

$$L(w) = \frac{1}{2\sigma^2} \| t - \Phi w \|_2^2 + \frac{1}{2\tau^2} \| w \|_2^2$$

Luego

$$w^* = \arg\min_w \left(\frac{1}{2} \|t - \Phi w\|_2^2 + \frac{\sigma^2}{2\tau^2} \|w\|_2^2 \right)$$

Haciendo $\lambda = \frac{\sigma^2}{\tau^2}$:

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

La estimación MAP incorpora conocimiento previo sobre los parámetros, asumiendo que w sigue una distribución gaussiana centrada en cero. Esto conduce al mismo estimador que regresión Ridge, pero con una interpretación Bayesiana: la regularización aparece como consecuencia natural del prior. En el caso límite $\tau^2 \to \infty$ se recupera la estimación de máxima verosimilitud.

Modelo Bayesiano Completo (Regresión lineal Bayesiana con Prior Gaussiana)

Función objetivo

A diferencia de MAP o ML, aquí no buscamos un único valo w, sino que

inferimos la **distribución posterior completa:**

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

Tanto el prior como la verosimilitud son distribuciones normales, por lo que la posterior también es gaussiana. Su forma es

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

con

$$S_N = \left(\frac{1}{\sigma^2}\Phi^T\Phi + \frac{1}{r^2}I\right)^{-1} \quad \text{y}$$

$$m_N = \frac{1}{\sigma^2} S_N \Phi^T t$$

Predicción:

Dada una nueva entrada x_*, se transforma como $\phi_* = \phi(x_*)$. La salida t_* también es una variable aleatoria, cuya distribución predictiva es gaussiana:

$$p(t_* \mid x_*, t, \Phi) = \mathcal{N}(t_* \mid \phi_*^T M_N, \phi_*^T \Sigma_N \phi_* \Sigma_N \phi_* + \sigma^2)$$

Esto nos da

- Una media predictiva $\mathbb{E}[t_*] = \phi_*^T M_N$
- Una varianza predictiva $\mathbb{V}[t_*] = \phi_*^T \Sigma_N \phi_* + \sigma^2$

Este enfoque no entrega un único vector w, sino una distribución posterior completa que refleja la incertidumbre sobre los parámetros dado el conjunto de datos observado. La predicción sobre nuevos puntos también es una distribución, no un valor fijo, lo que permite cuantificar la incertidumbre del modelo.
El modelo Bayesiano completo puede verse como una generalización natural de MAP, donde en lugar de maximizar la posterior se la integra completamente. Es también el antecedente directo de los procesos gaussianos, en los cuales se coloca una distribución directamente sobre funciones en vez de sobre parámetros w.

Regresión Ridge con Kernel

Partimos del mismo supuesto funcional que en regresión lineal regularizada:

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

Pero en este caso:

- El espacio de características $\phi(\cdot)$, puede ser de muy alta (incluso infinita) dimensión.
- En lugar de trabajar explícitamente con ϕ, se utiliza un Kernel(x, x') definido como:

$$K(x, x') = \phi(x)^T \phi(x')$$

Función objetivo.

Como en Ridge, se busca minimizar

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

En lugar de resolver directamente para w, se utiliza la forma dual, que parte del hecho que la solución w puede expresarse como una combinación lineal de los vectores de entrenamiento:

$$w = \Phi^T \alpha$$

Este cambio permite reformular el problema en términos de un nuevo conjunto de parámetros $\alpha \in \mathbb{R}^N$, eliminando la necesidad de calcular $\phi(x)$ explícitamente.

Todo se puede expresar en función

de productos punto entre vectores transformados, que se sustituyen por evaluaciones del Kernel.

Reemplazando en la función objetivo y derivando respecto a α, se obtiene el sistema dual:

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

donde $K \in \mathbb{R}^{N \times N}$ es la **matriz Kernel** definida como:

$$K_{ij} = K(x_i, x_j)$$

Predicción

Dado un nuevo punto x_*, su predicción se calcula como:

$$t_* = \sum_{n=1}^{N} \alpha_n K(x_n, x_*)$$

Es decir, la predicción se expresa como una combinación ponderada de los

Similitudes entre el nuevo punto y los puntos de entrenamiento.

Kernel Ridge Regression permite trabajar con espacios de características complejos o de dimensión infinita sin calcular explícitamente la transformación $\Phi(x)$ gracias al truco del kernel. Este modelo sigue siendo lineal en un espacio de características (implícito), pero es capaz de capturar relacionales altamente no lineales en el espacio original. Su desventaja principal es que requiere almacenar y operar sobre la matriz kernel $K \in \mathbb{R}^{N \times N}$ lo que limita su escalabilidad. Cuando se utiliza un kernel lineal $K(x,x') = x^T x'$ se recupera el modelo Ridge Clásico.

Procesos Gaussianos

Un proceso gaussiano (GP) es una distribución sobre funciones

$$f(\cdot) \sim GP(m(\cdot), K(\cdot,\cdot))$$

donde

- $m(x) = \mathbb{E}[f(x)]$ es la función media (usualmente se toma como cero)

- $K(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$

  es una función de covarianza o Kernel.

En regresión, se asume que los datos observados provienen de esta función con ruido gaussiano aditivo.

$$t_n = f(x_n) + \eta_n, \quad \eta_n \sim \mathcal{N}(0, \theta^2)$$

Supuestos

- El vector de salidas t está gobernado por una función $f \sim \mathcal{GP}(0, K)$
- Ruido gaussiano con varianza θ^2,
- Kernel $K(x, x')$, simétrico y positivo.

Inferencia y predicción

Dado el conjunto de entrenamiento

$D = \{(x_n, t_n)\}_{n=1}^{N}$, se define la matriz Kernel:

$$K \in \mathbb{R}^{N \times N}, \quad K_{ij} = K(x_i, x_j)$$

Para un nuevo punto x_*, se define

- $k_* = [K(x_1, x_*), \ldots, K(x_N, x_*)]^T$
- $k_{**} = K(x_*, x_*)$

La distribución predictiva para t_* es gaussiana:

$$p(t_*, x_*, t) = \mathcal{N}(\mu_*, \sigma_*^2)$$

con

$$\mu_* = k_*^T (K + \sigma^2 I)^{-1} t \quad \text{y}$$

$$\sigma_* = k_{**} - k_*^T (K + \sigma^2 I)^{-1} k_*.$$

Los procesos gaussianos generalizan por completo el enfoque bayesiano. En lugar de asumir una forma parametrica para la función (como una combinación de $\Phi(x)^T w$) colocan una distribución directamente sobre funciones, permitiendo capturar

relaciones altamente no lineales con control explícito de la incertidumbre.

Este modelo no calcula un vector de parámetros w, ni depende de una transformación explícita $\Phi(w)$ todo se hace a través del kernel. De hecho, si se usa un kernel tipo $K(x,x') = \Phi(x)^T \Phi(x')$ se recuperan los mismos resultados predictivos que el modelo bayesiano lineal con prior gaussiano sobre w.

La principal limitación de los GPs es computacional: su cost es cúbico en el número de datos $O(N^3)$ lo cual restringe su uso a conjuntos relativamente pequeños.