

## Notación:

Sea  $\Phi \in \mathbb{R}^{N \times p}$  la matriz de diseño cuyas filas son  $\phi(x_n)$ . Denotamos  $t = [t_1, t_2, \dots, t_N]^T$  y  $\|v\|^2 = v^T v$ .

## 1. Mínimos Cuadrados (LS)

Buscamos encontrar los pesos  $w$ , tales que estos minimicen el error cuadrático residual:

$$\mathcal{L}(w) = RSS(w) = \sum_n [t_n - \phi(x_n)^T w]^2 = \|t - \Phi^T w\|^2$$

Luego, tenemos que:

$$w_{LS} = \underset{w}{\operatorname{arg\min}} RSS(w) = \underset{w}{\operatorname{arg\min}} \|t - \Phi^T w\|^2$$

Así, para minimizar RSS, escribimos en forma matricial y derivamos:

$$\mathcal{L}(w) = (t - \Phi^T w)^T (t - \Phi^T w) = t^T t - 2 w^T \Phi^T t + w^T \Phi^T \Phi w$$

Colemos el gradiente respecto a  $w$  e igualamos a cero:

$$\nabla_w \mathcal{L}(w) = -2 \Phi^T t + 2 \Phi^T \Phi w = 0$$

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

Suponemos que  $\Phi^T \Phi$  es invertible, la solución analítica es:

$$w_{LS} = (\Phi^T \Phi)^{-1} \Phi^T t$$

Nota: Se intenta minimizar la distancia euclídea entre los datos observados y el punto predicho.

## 2. Mínimos Cuadrados Regularizados (RLS)

Buscamos garantizar la invertibilidad de la matriz  $\Phi^T \Phi$  y controlar la complejidad del modelo de manera que no responda el ruido del mismo, para ello, penalizamos el costo con la norma L2 de los pesos del modelo:

$$L_\lambda(\omega) = \sum_n [t_n - \phi(x_n)^T \omega]^2 + \lambda \|\omega\|_2^2$$

$$L_\lambda(\omega) = \|t - \Phi \omega\|_2^2 + \lambda \omega^T \omega, \quad \lambda > 0$$

Donde  $\lambda$  y  $\|\omega\|_2^2$  son el parámetro o coeficiente de regularización y la norma L2 de los pesos del modelo, respectivamente. Así, calculando el gradiente respecto a  $\omega$  e igualando a cero:

$$\nabla_{\omega} L_\lambda(\omega) = -2\Phi^T t + 2\Phi^T \Phi \omega + 2\lambda \omega = 0$$

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

Como  $\Phi^T \Phi + \lambda I$  es invertible para  $\lambda > 0$ , dado esto asegura que los autovalores de la matriz sean mayores a cero, entonces la solución analítica es:

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

## 3. Máxima Verosimilitud (MLE) bajo ruido Gaussiano

Buscamos encontrar los pesos  $\omega$  tales que permitan maximizar la probabilidad de daterse el conjunto de datos dados estos mismos pesos y el parámetro o varianza del ruido  $\sigma_n^2$ :

$$\underset{\omega, \sigma_n^2}{\operatorname{argmax}} p(t | \Phi, \omega, \sigma_n^2) = \underset{\omega, \sigma_n^2}{\operatorname{argmax}} \prod_n N(t_n | \phi(x_n)^T \omega, \sigma_n^2 I)$$

Considerando datos i.i.d, podemos reconfigurar el problema para maximizar el siguiente costo (Log-Likelihood):

$$\log p(t | \Phi, \omega, \sigma_n^2) = \sum_n [\log N(t_n | \phi(x_n)^T \omega, \sigma_n^2 I)]$$

$$\mathcal{L}(w, \sigma_n^2) = -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} \|t - \Phi w\|_2^2$$

Vemos que maximizar el log-likelihood es equivalente a minimizar  $\|t - \Phi w\|_2^2$ , puesto que el primer sumando es constante respecto a  $w$  y la norma es negativa. Por tanto, el estimador MLE cumple:

$$w_{MLE} = \underset{w}{\operatorname{argmax}} \log p(t | \Phi w, \sigma_n^2) = \underset{w}{\operatorname{argmin}} \|t - \Phi w\|_2^2$$

De donde se recupera la solución de mínimos cuadrados:

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

Nota: La regresión MLE es equivalente a mínimos cuadrados matemáticamente; sin embargo, su planteamiento no se enfoca en minimizar el error sino en establecer el modelo tal que permita los datos observados sean lo más probables posible, aún bajo la condición de nula.

#### 4. Máxima A Posteriori (MAP)

Se introduce el concepto de **prior**, que representa el conocimiento previo o bien las hipótesis que se tienen respecto a la forma de la distribución de los datos a predecir. Considerando un prior gaussiano sobre  $w$ :

$$p(w, \lambda) = N(w | \phi, \lambda^{-1} I),$$

donde  $\lambda$  es la precisión y codifica la varianza o dispersión del prior. Con la verosimilitud gaussiana del apartado anterior, la posterior es:

$$p(w | t, \phi, \sigma_n^2, \lambda) \propto p(t | w, \phi, \sigma_n^2) p(w, \lambda)$$

Tomando log-posterior:

$$\log p(w | t, \phi, \sigma_n^2, \lambda) = -\frac{1}{2\sigma_n^2} \|t - \Phi w\|_2^2 - \frac{\lambda}{2} \|w\|_2^2 + \text{const}$$

Maximizar la log-posterior (MAP) equivale a minimizar la función de costo:

$$J_{\text{MAP}}(\omega) = \frac{1}{2\sigma_n^2} \|t - \Phi\omega\|_2^2 + \frac{\lambda}{2} \|\omega\|_2^2$$

Multiplicando por  $2\sigma_n^2$ , obtenemos exactamente el problema RLS con  $\lambda' = \lambda\sigma_n^2$ . Derivando y igualando a cero:

$$(\Phi^T\Phi + \lambda\sigma_n^2 I)\omega = \Phi^T t$$

Si reparametrizamos  $\lambda' = \lambda\sigma_n^2$ , se recupera la forma anterior; luego la solución analítica es:

$$\omega_{\text{MAP}} = (\Phi^T\Phi + \lambda' I)^{-1}\Phi^T t,$$

Notando que el estimador MAP con prior gaussiana es matemáticamente equivalente a la solución RLS (Ridge).

## 5. Regresión Bayesiana con Modelos Lineal Gaussiano

Definimos la precisión del ruido como  $\beta = 1/\sigma_n^2$ .

Prior y Verosimilitud:

Elegimos un prior gaussiano conjugado sobre los pesos:

$$p(\omega, \lambda) = N(\omega | \omega_0, \lambda^{-1} I), \lambda > 0$$

La verosimilitud de los datos dados  $\omega$  es:

$$p(t | \omega) = N(t | \Phi\omega, \beta^{-1} I)$$

Posterior  $p(\omega | t)$ :

Por conjugación, la posterior también es gaussiana. Calculamos su precisión y media:

- Desarrollando log-posterior:

$$\log p(w|t) \propto \log(p(t|w) + \log p(w)) = -\frac{\beta}{2} \|t - \Phi w\|_2^2 - \frac{\lambda}{2} \|w\|_2^2$$

- Expanding and reordering the quadratic terms:

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

Los términos cuadráticos en  $w$  definen la precisión posterior:

$$S_N^{-1} = \beta \Phi^T \Phi + \lambda I$$

Ahora, la media posterior satisface:

$$S_N^{-1} m_N = \beta \Phi^T t$$

$$\rightarrow m_N = S_N \Phi^T t$$

Observemos que si se toma el límite  $\lambda \rightarrow 0$  y  $\beta$  fijo,  $m_N$  tiende a la solución de mínimos cuadrados si  $\Phi^T \Phi$  es invertible.

Predictiva para un nuevo  $x_{*k}$ :

Sea  $\phi_* = \phi(x_{*k})$ . La distribución predictiva marginal de  $t_{*k}$  integrando sobre  $w$  es:

$$p(t_{*k}|t, x_{*k}) = \int p(t_{*k}|w, x_{*k}) p(w|t) dw$$

con  $p(t_{*k}|w, x_{*k}) = N(\phi_*^T w, \beta^{-1})$  y  $p(w|t) = N(m_N, S_N)$ , se obtiene:

$$p(t_{*k}|t, x_{*k}) = N\left(\underbrace{\phi_*^T m_N}_{\mu}, \underbrace{\beta^{-1} + \phi_*^T S_N \phi_*}_{\sigma^2}\right)$$

## 6. Regresión Kernel Rígida (Primal y Dual)

Sea la solución primal:

$$\omega = (\phi^T \phi + \lambda I_D)^{-1} \phi^T y \quad (1)$$

y la siguiente identidad matricial:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P^T (B P^T + R)^{-1} \quad (2)$$

Se hará uso de esta para transformar la inversa de dimensión  $D \times D$  en una inversa de  $N \times N$ .

Escojemos:

$$B = \Phi, \quad R = I_N, \quad P^{-1} = \lambda I_D \rightarrow P = \lambda^{-1} I_D$$

Sustituyendo en (2):

• Lado Izquierdo:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = (\Phi^T \Phi + \lambda I_D)^{-1} \Phi^T$$

• Lado Derecho:

$$P^T (B P^T + R)^{-1} = \Phi^T (\Phi \Phi^T + \lambda I_N)^{-1}$$

De esta manera, resulta la igualdad:

$$\omega = (\Phi^T \Phi + \lambda I_D)^{-1} \Phi^T t = \Phi^T (\Phi \Phi^T + \lambda I_N)^{-1} t$$

Definimos entonces:

[Solución Dual]

•  $\alpha = (K + \lambda I_N)^{-1} t$ , donde  $K = \Phi \Phi^T$  y  $\omega = \Phi^T \alpha$  con  $K$ : Matriz Kernel.

De esta manera, se demuestra que la solución primal (ridge) es equivalente a la representación dual en términos del Kernel, donde la identidad provee el paso clave para mover la inversión entre el espacio de características y de muestras.

Resumen

(Truco del Kernel). Este tipo de regresión puede aproximar relaciones no lineales por medio del mapeado del espacio nativo a un espacio de mayores dimensiones, donde el objeto o conjunto de datos a predecir.

## 7. Procesos Gaussianos

Un proceso gaussiano prior sobre funciones es:

$$\tilde{f}(x) \sim GP(0, K(x, x))$$

Sea  $\tilde{\Phi} = [\tilde{\Phi}(x_1), \tilde{\Phi}(x_2), \dots, \tilde{\Phi}(x_N)]^T$ . Por definición:

$$\tilde{\Phi} \sim N(0, K),$$

donde  $K \in \mathbb{R}^{N \times N}$  con  $K_{ij} = K(x_i, x_j)$ . Regresando a la distribución de ruido gaussiano, tenemos que:

$$t = \phi + \eta, \text{ con } \eta \sim N(0, \sigma_n^2 I)$$

Entonces la marginal sobre las observaciones es:

$$p(t | X) = N(0, K + \sigma_n^2 I)$$

Predicción en un punto  $x_*$ :

Consideremos la distribución conjunta de  $[t; \phi_*]$ :

$$\begin{pmatrix} t \\ \phi_* \end{pmatrix} \sim N\left(0, \begin{bmatrix} K + \sigma_n^2 I & K_* \\ K^T & K_{**} \end{bmatrix}\right),$$

donde  $K_* = [K(x_*, x_1), \dots, K(x_*, x_N)]^T$  y  $K_{**} = K(x_*, x_*)$

Así, la condición gaussiana nos da la distribución posterior predictiva del valor latente  $\phi_*$ :

$$p(\phi_*, t) = N(m_*, v_*) \text{, con}$$

$$m_* = K_*^T (K + \sigma_n^2 I)^{-1} t, \quad v_* = K_{**} - K_*^T (K + \sigma_n^2 I)^{-1} K_*$$

Si deseamos la distribución de los observaciones independientes, entonces:

$$p(t_{\text{test}}) = \mathcal{N}(m_k^T K^{-1} \sigma_n^2)$$

Equivocion entre Kernel Ridge y GP (Media Posterior):

Comparando la media posterior de cada modelo:

$$m_{\text{GP}} = K^T (K + \sigma_n^2 I)^{-1} t \quad \text{y} \quad m_{\text{KR}} = K^T (K + \lambda I)^{-1} t$$

Por tanto, si  $\lambda = \sigma_n^2$ , entonces GP y KR son equivalentes en la estimación de la media.

Cada modelo provee una vista a un espacio de representación mutuamente complementarios: GP trata funciones completas y su espacio de representación muestra que KR se encarga de estimaciones puntuales en el espacio de pesos (Consideración de medias predictivas) con un criterio de regularización.