

PARCIAL I TAM.

1. Modelo base de Regresión.

$$t_n = \phi(x_n)w^T + \eta_n, \text{ con } \{\eta_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N, w \in \mathbb{R}^Q,$$

$\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q, Q \geq P$ y

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

El problema de optimización tiene como objetivo encontrar el vector de pesos w que minimiza la suma de los errores al cuadrado entre salidas reales t_n y las predicciones del modelo $\phi(x_n)^T w$

- Función de costo.

$$J(w) = \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2$$

Llevando a la forma matricial:

- $t \in \mathbb{R}^N$: Vector con salidas reales $[t_1, t_2, \dots, t_N]$
- $\phi \in \mathbb{R}^{N \times Q}$: matriz de diseño
- $w \in \mathbb{R}^Q$: Vector de parámetros, $x_n \in \mathbb{R}^P$

- En forma matricial:

$$J(w) = \|t - \phi w\|^2 = (t - \phi w)^T (t - \phi w)$$

t : Vector de salidas

ϕ : matriz de diseño

w : Vector de pesos

- Gradiente descendiente

Para encontrar el mínimo se deriva $J(w)$ con respecto a w , se iguala a cero y se despeja w

Expandiendo de forma matricial tenemos:

$$J(\omega) = (t^T \cdot \omega^T \phi^T) (t - \phi \omega) = t^T t - t^T \phi \omega - \underbrace{\omega^T \phi^T t}_{\text{Escalar}} + \omega^T \phi^T \phi \omega$$

Dado que $\omega^T \phi^T t = \text{Escalar} ; (\omega^T \phi^T t)^T = t^T \phi \omega$

Simplificación

$$J(\omega) = t^T t - 2 \omega^T \phi^T t + \omega^T \phi^T \phi \omega$$

• Se deriva la expresión respecto a ω : $\nabla_{\omega} J(\omega)$

$$\frac{d}{d\omega} = \nabla_{\omega} (t^T t - 2 \omega^T \phi^T t + \omega^T \phi^T \phi \omega) \quad \nabla_{\omega} J(\omega) = 0 - 2 \phi^T t + 2 \phi^T \phi \omega$$

En forma compacta: Por propiedad matricial

$$\nabla_{\omega} J(\omega) = -2 \phi^T (t - \phi \omega)$$

$$-2 \phi^T (t - \phi \omega) = \underline{\underline{0}}$$

$$\phi^T (t - \phi \omega) = 0$$

Se resuelve la ecuación para ω

1. Prop: de distribución

$$\phi^T t - \phi^T \phi \omega = 0$$

2. despejar ω

$$\phi^T \phi \omega = \phi^T t$$

3. multiplicar a ambos lados por $(\phi^T \phi)^{-1}$, solo cuando la matriz es invertible

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

de forma simplificada

$$\omega_{LS} = (\phi^T \phi)^{-1} \phi^T t$$

2. Mínimo Cuadrados Regularizados

- Función de costo regularizada: En términos de L2

$$J(w) = \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2 + \lambda \sum_{j=1}^q w_j^2$$

hipoparametro

- En forma matricial: $\phi^T w = \|w\|^2$

$$J(w) = \|t - \phi w\|^2 + \lambda w^T w$$

↓
Penalización: A mayor λ los pesos

deben ser mas pequeños.

- Derivada de las sumas de las 2 componentes

$$\nabla_w (\|t - \phi w\|^2) = -2 \phi^T (t - \phi w)$$

- Derivada de la regularización

$$\nabla_w (\lambda w^T w) = 2 \lambda w$$

- Gradientes Completo

$$\nabla_w J(w) = -2 \phi^T (t - \phi w) + 2 \lambda w$$

- Se iguala a cero y despejar el vector de pesos w

$$\frac{-2 \phi^T (t - \phi w)}{2} + \frac{2 \lambda w}{2} = 0$$

$$-\phi^T (t - \phi w) + \lambda w = 0$$

- distribuir $-\phi^T$

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

$$\phi^T \phi w + \lambda w = \phi^T t$$

- Se factoriza w ; λ = escalar, para que la suma matricial sea valida se multiplica por la matriz identidad de tamaño ($Q \times Q$)

$$(\phi^T \phi + \lambda I) w = \phi^T t$$

- multiplicando en ambos lados por la inversa de $(\phi^T \phi + \lambda I)$

$$w_{\text{ridge}} = (\phi^T \phi + \lambda I)^{-1} \phi^T t$$

λI asegura que la matriz $(\phi^T \phi + \lambda I)$ sea invertible siempre que $\lambda > 0$, hace solución numéricamente más estable que minimos cuadrados.

3. Máxima Verosimilitud - MLE

- Asumiendo que el ruido es Gaussiano, la probabilidad condicional de una observación t_n es:

$$P(t_n | X_n, w, \sigma_n^2) = N(t_n | \phi(X_n)^T w, \sigma_n^2)$$

Dado que los datos son i.i.d la probabilidad conjunta de todo el vector de salidas t es el producto de las probabilidades individuales:

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

- Trabajar con productos es computacionalmente complejo y numéricamente instable, por eso se maximiza con logaritmo

Para convertir productos en sumas:

$$\log P(t|...) = \log \left(\prod_{n=1}^N N(\dots) \right) = \sum_{n=1}^N \log N(t_n | \phi(x_n)^T w, \sigma_n^2)$$

Ahora sustituimos la PDF GAUSSIANA

$$N(z|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

$$z = x_n$$

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

• De forma Simplificada

$$\log P(t|...) = -\frac{N}{2} \underbrace{\log(2\pi\sigma_n^2)}_A - \underbrace{\frac{1}{2\sigma_n^2} \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2}_B$$

A = no depende de w por lo tanto es constante

B = depende de w, al estar precedido por un menor maximizar todo el término, equivale a minimizar la parte positiva.

$$\sum_{n=1}^N (t_n - \phi(x_n)^T w)^2$$

de maximización se transforma a minimización

$$\arg \max_w \log P(t|...) = \arg \min_w (t_n - \phi(x_n)^T w)^2$$

Esto es (SSE) Suma de errores al cuadrado, equivalente a ML suponiendo que el error o ruido sigue una distribución Gaussiana

Dado que el problema de optimización es mínimos cuadrados

MLE:

$$\hat{w}_{MLE} = (\phi^T \phi)^{-1} \phi^T t$$

4) MAXIMO A POSTERIORI

El objetivo es encontrar el vector de pesos w combinando Verosimilitud de los datos con Prior sobre los pesos.

- Se parte del teorema de bayes

$$p(w|t, X) \propto p(t|X, w) \cdot p(w)$$

$\alpha = \text{Precisión}$

Prior: Creciendo sobre w antes de ver los datos, asumiendo prior Gaussiano centrado en cero. Creemos que los pesos pequeños son más probables que los pesos grandes.

$$p(w) = N(w|0, \alpha^{-1} I)$$

- Construir Función Log-Posterior.

$$\text{Log}(w|t, X) \propto \text{Log}(t|X, w) + \text{Log } p(w)$$

1. Log-Likelihood (paso de MLE con $\beta = 1/\sigma_n^2$)

$$\text{Log } p(t|X, w) = -\frac{\beta}{2} \|t - \phi w\|^2 + C$$

2. Log-Prior (Logaritmo de una Gaussiana multivariada):

$$\text{Log } p(w) = -\frac{\alpha}{2} w^T w + \text{const} = -\frac{\alpha}{2} \|w\|^2 + \text{const.}$$

Sumando ambos términos, obtenemos la log-Posterior:

$$\text{Log} p(w|...) \propto -\frac{\beta}{2} \|t - \phi w\|^2 - \frac{\alpha}{2} \|w\|^2$$

$$\underset{w}{\operatorname{argmax}} -\frac{\beta}{2} \|t - \phi w\|^2 - \frac{\alpha}{2} \|w\|^2$$

Esta función es equivalente a minimizar su negativo multiplicando por -1 e ignorando constantes de proporcionalidad. Se obtiene la función de costo a minimizar:

$$J(w) = \frac{\beta}{2} \|t - \phi w\|^2 + \frac{\alpha}{2} \|w\|^2$$

- Relacionar con Min. cuadrados regularizados (Ridge) función de costo MAP

$$J(w) = \frac{\beta}{2} \|t - \phi w\|^2 + \frac{\alpha}{2} \|w\|^2$$

matemáticamente idéntico a la función de costo de la regresión Ridge:

$$J_{\text{Ridge}}(w) = \|t - \phi w\|^2 + \lambda \|w\|^2$$

Multiplicando la función de costo MAP por $\frac{2}{\beta}$

$$J(w) \propto \|t - \phi w\|^2 + \frac{\alpha}{\beta} \|w\|^2$$

λ_{ridge} = relación entre la precisión del prior y ruido de datos (β)

$$\lambda = \frac{\alpha}{\beta} = \alpha \sigma_n^2$$

$$w_{\text{MAP}} = (\phi^\top \phi + \lambda I)^{-1} \phi^\top t$$

Ridge = MAP con prior Gaussiano

• 5 REGRESIÓN BAYESIANA COMPLETA

distribución posterior completa $p(w|t, X)$

Calcular $p(w|t, X)$ usando teorema de Bayes.

Configuración

- Likelihood: $p(t|X, w) = N(t|\phi w, \beta^{-1}I)$ - verosimilitud
- Prior: $p(w) = N(w|0, \alpha^{-1}I)$ - creencia previa

Solución para la posterior:

$$p(w|t, X) = N(w|M_N, S_N)$$

M_N : Media posterior \rightarrow centro distribución de los pesos

$$M_N = \beta S_N \phi^T t$$

Covarianza (S_N): Mide la incertidumbre de los pesos

$$S_N = (\alpha I + \beta \phi^T \phi)^{-1}$$

Integral predictiva

$$p(t_*|x_*, t, X) = \int p(t_*|x_*, w) \cdot p(w|t, X) dw = N(t_*|\phi^T M_N, \dots \beta^{-1} + \phi_*^T \sum_N \phi_*)$$

- Calculo de la distribución predictiva:

$$p(t_*|x_*, t, X) = N(t_*|M(x_*) \Gamma^2(x_*))$$

$M(x_*)$: Media predictiva: Se calcula usando la media de los pesos posteriores.

$$\mathbb{E}[t_*] = M(x_*) = \phi(x_*) M_N$$

$\Gamma^2(x_*)$: Varianza predictiva: Mide la incertidumbre total de la predicción $\text{Var}(t_*) = \beta^{-1} + \phi_*^T \sum_N \phi_*$

$$\Gamma^2(x_r) = \frac{1}{B} + \phi(x_r)^T S_N \phi(x_r)$$

- $\frac{1}{B}$ = variancia de los datos
- $\phi(x_r)^T S_N \phi(x_r)$: incertidumbre en los parámetros del modelo.

6. Regresión Rígida Kernel (Kernel Ridge)

- Problema de optimización

$$w = \phi^\alpha$$

$$J(\alpha) = \|t - \phi\phi^\alpha\|^2 + \lambda \alpha^\alpha \phi^\alpha \phi$$

α = vector de coeficientes duros

- Matriz de Kernel $K = \phi\phi^\alpha$, donde $K_{ij} = \phi(x_i)^\alpha \phi(x_j) = K(x_i, x_j)$
- el problema se convierte en:

$$J(\alpha) = \|t - K\alpha\|^2 + \lambda \alpha^\alpha K \alpha$$

- Se deriva la solución para los coeficientes

$$\nabla_\alpha J(\alpha) = 2K^\alpha(t - K\alpha) + 2\lambda K\alpha = 0$$

- Se resuelve para α (sabiendo de K es simétrica, $K^\alpha = K$)

$$-Kt + K\alpha K + \lambda K\alpha = 0$$

$$K(K + \lambda I)\alpha = Kt$$

$$(K + \lambda I)\alpha = t$$

- La solución final para los coeficientes duros es:

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

Interpretación:

- Se realizan predicciones usando el Kernel.

Fórmula de predicción

$$y(x_*) = \phi(x_*)^T w.$$

- Se sustituye $w = \phi^T a$

$$y(x_*) = \phi(x_*)^T (\phi^T a) = (\phi(x_*)^T \phi^T) a$$

$\phi(x_*)^T \phi^T$ = Vector fila \rightarrow elementos son el producto punto de $\phi(x_*)$ con cada $\phi(x_n)$ del conjunto de entrenamiento

- En función de Kernel

$$K(x_*) = [K(x_1, x_*), K(x_2, x_*), \dots, K(x_N, x_*)]$$

- La predicción final es una suma ponderada de las evaluaciones del Kernel:

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

- Todo lo que se necesita es una función Kernel $K(x_i, x_j)$ que calcule el producto punto.

7. Regresión Con Procesos Gaussianos (GP)

- Problema de optimización

Se define un prior de proceso Gaussiano (GP) sobre $f(x)$:

$$f(x) \sim GP(0, K(x, x'))$$

- Distribución Conjunta de los datos.

La propiedad de un GP es que cualquier conjunto finito de puntos extraídos de él sigue una distribución Gaussiana multivariada.

- Se quiere predecir la función f_{*} en unos nuevos puntos de test X_{*} , basado en observaciones ruidosas t en los puntos de entrenamiento X .
- Bajo el prior GP, las salidas (sin ruido) en los puntos de entrenamiento (f) y en los puntos de test (f_{*}) siguen una distribución Gaussiana conjunta:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix}\right)$$

Con ruido:

$$t = f + n. \text{ Este ruido, } n \sim N(0, \sigma^2 I)$$

La covarianza de las observaciones t es:

$$\text{COV}(t) = K(X, X) + \sigma^2 I$$

- Ahora el objetivo es encontrar la distribución de f_{*} es decir $p(f_{*} | X_{*}, X, t)$.

• Solución predictiva.

• distribución predictiva posterior para las salidas para los puntos de test es una Gaussiana.

$$p(f_{*} | X_{*}, X, t) = N(\bar{f}_{*}, \text{COV}(f_{*}))$$

\bar{f}_{*} = Media posterior, es la predicción puntual

$$\bar{f}_{*} = K(X_{*}, X)[K(X, X) + \sigma^2 I]^{-1} t$$

$\text{COV}(\bar{f}_{*})$: Covarianza posterior, esta matriz da la incertidumbre. La diagonal contiene la varianza de cada predicción individual.

$$\text{Cov}(f_n) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma^2_n I]^{-1} K(X, X_*)$$