

① Planteamiento Problema:

Modelo: $T_n = \phi(X_n) w^T + \eta_n$

$T_n \in \mathbb{R}$ Predicción de un escalar

$$X_n \in \mathbb{R}^P$$

$$w \in \mathbb{R}^Q$$

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

$$Q \geq P$$

(Asumimos $\Phi = \phi(X)$)

ESTAMOS llevando a X_n a un espacio superior mediante la transformación $w^T \Phi$

- Mínimos Cuadrados:

Problema de optimización:

$$\min \|T - \Phi(X)w^T\|^2$$

$$= \|T - \Phi w^T\|^2$$

$$\frac{d}{dw} (\cdot) = 2\Phi^T(T - \Phi w^T) = 2\Phi^T T - 2\Phi^T \Phi w^T$$

Para encontrar el mínimo igualamos a cero:

$$-2\Phi^T T + 2\Phi^T \Phi w^T = 0 \Rightarrow \cancel{2}\Phi^T T = \cancel{2}\Phi^T \Phi w^T$$

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

Mínimos cuadrados regulados

Problema de optimización

$$\min \|T - \Phi w^T\|^2 + \lambda \|w\|^2$$

$$\frac{d}{dT} (\cdot) = -2\Phi^T(T - \Phi w^T) + 2\lambda w^T = 0$$

$$\Phi^T \Phi w^T + 2\lambda w^T = \Phi^T T$$

$$(\Phi^T \Phi + 2\lambda I) w^T = \Phi^T T$$

$$w^T = (\Phi^T \Phi + 2\lambda I)^{-1} \Phi^T T$$

Máxima Verosimilitud:

En este caso estimaremos un w que maximice la verosimilitud de los datos

con: $\eta_n \sim N(0, \sigma_n^2)$

$T_n \sim N(\phi(x_n) W^\top, \sigma_n^2)$

ya que son iid, la verosimilitud es

$$P(T|w) = \prod_{n=1}^N N(T_n | \phi(x_n) W^\top, \sigma_n^2) \quad \{\sigma_n = \text{CTe}\}$$

Aplicando log a cada lado:

$$\log(P(T|w)) = \log\left(\prod_{n=1}^N N(\dots)\right)$$

ya que $N = \frac{1}{2\pi\sigma_n^2} \exp\left(-\frac{1}{2\sigma_n^2} \|T_n - \Phi_n w\|^2\right)$

$$\log(P(T|w)) = \log\left(\underbrace{\prod_{n=1}^N}_{\substack{\text{Suma} \\ \text{Pasar} \\ \text{de log}}}, \underbrace{\frac{1}{2\pi\sigma_n^2}}_{\text{CTe}}, \underbrace{\exp\left(-\frac{1}{2\sigma_n^2} \|T_n - \Phi_n w\|^2\right)}_{\text{cancela}}\right)$$

$$= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} \sum_{n=1}^N (T_n - \phi(x_n) W^\top)^2$$

ya que es negativo, maximizar essa função equivale a minimizar sua versão positiva

$$\min_w \cdot = -\frac{N}{2} \log(\cdot) + \frac{1}{2\sigma_n^2} \sum_{n=1}^N (T_n - \phi(x_n) W^\top)^2 = \eta_n^2$$

$$= \min_w \underbrace{\frac{N}{2} \log(2\pi\sigma_n^2)}_{\text{CTe}} + \frac{1}{2\sigma_n^2} \|T - \Phi w\|^2$$

Por lo que la solución es la misma que para minimos cuadrados:

$$(\Phi^\top \Phi)^{-1} \Phi^\top T$$

Máximo a Posteriori

En ese caso, en lugar de tomar w como un valorijo, se asume a w como una distribución probabilística

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

según Bayes

verosimilitud

$$\underbrace{P(w|T)}_{\text{Posterior}} \propto \underbrace{P(T|w)}_{\text{Prior}} P(w)$$

igualamos $P(T)$ ya que es constante en la optimización

sacando logaritmo a cada lado

$$\log(P(w|T)) = \log(P(T|w)) + \log(P(w))$$

$$\text{maximizando } \log(P(w|T)) \propto PCWIT$$

$$\max \log \left(\prod_{n=1}^N \frac{1}{2\pi\sigma_n^2} \right) \left\| T - \Phi w^\top \right\|^2 + \log \left(\frac{1}{(2\pi)^{N/2} |\alpha^{-1} I|^{1/2}} \exp \left(-\frac{\alpha}{2} \|w\|^2 \right) \right)$$

máxima verosimilitud

$$= \max \frac{N}{2} \log(2\pi\sigma_n^2) + \frac{-1}{2\sigma_n^2} \|T - \Phi w^\top\|^2 - \frac{\alpha}{2} \log(2\pi\alpha^{-1}) - \frac{\alpha}{2} \|w\|^2$$

multiplicando por (-1)

(equivale a minimizar posteriormente)

$$\min \text{CTE} + \frac{1}{2\sigma_n^2} \|T - \Phi w^\top\|^2 + \frac{\lambda}{2} \|w\|^2$$

es equivalente a regularizarla donde $\lambda = \frac{\alpha}{2\sigma_n^2}$

Por lo que la solución es:

$$w^\top = (\Phi^\top \Phi + \lambda^{-1} I)^{-1}$$

Bayesianos con modelo lineal:
se rompe el prior de los pesos
 $p(w) = N(w|0, \alpha^{-1} I)$

Verosimilitud:

$$P(T|w) = N(T^T \Phi w, \sigma_n^2 I)$$

Por bayes

$$P(w|T) = \frac{P(T|w) P(w)}{P(T)}$$

Ya que el prior, la verosimilitud y evidencia son gaussianos, el posterior también

$$P(w|T) \sim N(w|\mu_N, \Sigma_N)$$

Sacando \ln a cada lado:

$$\ln(P(w|T)) = \frac{-1}{2\sigma_n^2} \|T - \Phi w\|^2 + \frac{\alpha}{2} \|w\|^2 + Cte$$

$$= -(T^T T - 2T^T \Phi w + w^T \Phi^T \Phi w) \frac{1}{2\sigma_n^2} + \frac{\alpha}{2} w^T w + Cte$$

Agrupando w^T :

$$= \frac{1}{2} w^T \left(\frac{1}{\sigma_n^2} \Phi^T \Phi + \alpha I \right) w + \left(\frac{1}{\sigma_n^2} T^T \Phi \right) w + \frac{1}{2\sigma_n^2} T^T T + Cte$$

Recordando la forma de una distribución gaussiana multivariada

$$= \left(\frac{1}{2\pi} \right)^{p/2} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

Expandiendo el exponente:

$$-\frac{1}{2} \left[x^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu + Cte \right]$$

Si reemplazamos $x = w$, encontramos que

$$\Sigma_N^{-1} = \frac{1}{\sigma_n^2} \Phi^\top \Phi + \alpha I ; \quad \Sigma = \left(\frac{1}{\sigma_n^2} \Phi^\top \Phi + \alpha I \right)^{-1}$$

$$\mu_N = \frac{1}{\sigma_n^2} \sum \bar{\Phi}^\top T$$

Con estos datos, ya se puede establecer la PDF del posterior, para realizar una predicción de T_* dados los datos X_*

$$P(T_* | X_*, T) = \int P(T_* | X_*, W) P(W | T) dW$$

Posterior

El observación de este modelo no es un concreto sino una PDF para la predicción T_* dados X_*

$$P(T_* | X_*, W) = N(T_* | \bar{\Phi}^\top W, \sigma_n)$$

$$P(W | T) = N(W | \mu_N, S_N)$$

La integral corresponde a la ecuación 3.57 de "Pattern Recognition and Machine Learning" de Bishop

La integral resulta en otra desembocación normal

$$P(T_* | X_*, W) = N(T_* | \mu_*, \sigma_*)$$

Transformación lineal

Donde dado a que es la convolución de dos Gaussianas se pude hacer uso de la propiedad

$$E[T_*] = E[\bar{\Phi}^\top W + \eta_*] = \bar{\Phi}^\top \mu_N = \mu_*$$

$$\text{Var}(T_*) = \text{Var}(\bar{\Phi}^\top W) + \text{Var}(\eta_*) = \bar{\Phi}^\top \Sigma_N \bar{\Phi} + \sigma_*^2 = \sigma_*^2$$

Por lo tanto

$$P(T_* | X_*) = N(T_* | \mu_*, \bar{\Phi}^\top \Sigma_N \bar{\Phi} + \sigma_*^2)$$

Por lo que cada predicción nos da una estimación puntual en: μ_*

y una incertidumbre en desembocación gaussiana con $\text{Var} = \sigma_*^2$

Regresión Ridge Kernel

En este caso el objetivo de optimización es el mismo que en mínimos cuadrados regularizados

$$\min \|T - \phi(x)^T w\|^2 + \lambda \|w\|^2$$

Igualando a 0 y/o derivando y despejando w

$$w = \frac{1}{\lambda} \|T - \phi(x)^T w\|^2 \phi(x) = \frac{1}{\lambda} \sum \{\phi(x_n)^T w - T_n\} \phi(x_n)$$

Si ahora tomamos

$$a_n = \frac{1}{\lambda} \{w^T \phi(x_n) - T_n\}$$

$$w = \sum_{n=1}^N a_n \phi(x_n) = \Phi^+ a$$

$$\text{donde } \Phi^+ = \phi(x_n)^T \quad \text{y } a = (a_1, a_2, \dots, a_N)^T$$

Ahora, en lugar de trabajar con w, se reformula el algoritmo de mínimos cuadrados en términos del

vector a, si ahora se reemplaza $w = \Phi^+ a$ en la función de costo:

$$L\{\Phi^+ a\} = \|T - \Phi \Phi^+ a\|^2 + \lambda \|\Phi^+ a\|^2$$

$$= T^T T - 2a^T \Phi \Phi^+ T + a^T \Phi \Phi^+ \Phi \Phi^+ a + \frac{\lambda}{2} a^T \Phi \Phi^+ a$$

$$\text{Si } K = \Phi \Phi^+ = K^T \quad \text{y } K_{nm} = \phi(x_n)^T \phi(x_m) = K(x_n, x_m)$$

$$L\{\Phi^+ a\} = a^T K K a - 2a^T K T + T^T T + \lambda a^T K a$$

desarrollando respecto a "a":

$$2K^T K a + \lambda K T + 2K a = 0; (K K + \lambda I) a = K T$$

$$a = (K K + \lambda I)^{-1} T$$

Para predecir un nuevo punto x^* dado x^*

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

Procedo y supongo

en este caso se define una función f :

$$T_n = f(x_n) + \eta_n \quad \eta \sim N(0, \sigma^2)$$

señalo de nuevo una distribución sobre dichas funciones:

$$f(x) \sim GP\left(m(x), K(x, x')\right) \quad \text{tomamos } m(x) = 0$$

medea covarianza
con kernel K

la regresión es el dato entre los parámetros por

$$P(T|f(x)) = N(T|f(x), \beta^{-1} I_N)$$

Donde β es la precisión del modelo

Regresando a la definición del GP:

$$P(f(x)) = N(f(x)|0, K)$$

Por lo que $E[f(x)] = f(x)$

$$P(T) = \int P(T|y) P(y) dy = N(T|0, C)$$

Donde $C = K + \beta^{-1} I$

Para hacer una predicción de T^* dado x^*

Necesitamos evaluar la distribución predicción

$$P(T_A|t)$$

para ello evaluamos la conjugada $P(T_{N+1})$, donde

$$T_{N+1} = (T_1, \dots, T_N, T_A)$$

$$P(T_{N+1}) = N(T_{N+1}|0, C_{N+1})$$

Donde:

$$C_{N+1} = \begin{pmatrix} C_N & K \\ K^T & C \end{pmatrix} \quad \text{y} \quad C = K(X_{N+1}, X_{N+1}) + B^{-1}$$

Usando las mismas propiedades de las transformaciones lineales de Gauss-Jordan se encuentra que:

$$\mu(X_{N+1}) = K^T C_N^{-1} T$$

$$\sigma_\eta^2(X_{N+1}) = C - K^T C_N^{-1} K$$

Para ajustar los hiperparámetros se usa el max log de la verosimilitud, siendo estos hiperparámetros el Kernel y la varianza de ser desconocida.