

Parcial H1

Sea el modelo de regresión $t_n = \phi(x_n)w^T + n_n$, 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 $n_n \sim N(n_n | 0, \sigma_n^2)$

- mínimos cuadrados

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

$w \rightarrow$ parámetros a estimar

$\phi(x_n) \rightarrow$ transformación de las variables de entrada

$n_n \rightarrow$ ruido gaussiano medido o varianza constante

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

$$t = [t_1, t_2, \dots, t_N]^T \in \mathbb{R}^N \quad \text{vector de errores}$$

$$t - \Phi w$$

$$\Phi = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T \end{bmatrix} \in \mathbb{R}^{N \times q}$$

función de costo

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

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

Derivamos función de costo

$$J(w) = (t - \Phi w)^T (t - \Phi w) = t^T t - 2t^T \Phi w + w^T \Phi^T \Phi w$$

se deriva con respecto a w

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

$$-2\Phi^T t + 2\Phi^T \Phi w = 0$$

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

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

$\Phi^T \Phi$: matriz de correlaciones entre las variables de entrada

$\Phi^T t$: correlación entre variable de entrada y objetivo

$(\Phi^T \Phi)^{-1} \Phi^T t$: vector de peso que minimiza la suma de los errores al cuadrado

- Mínimos cuadrados regularizados

se agrega un término más que penaliza la magnitud de w para evitar sobreajuste

Función de costo con Regularización

$$J(w) = \sum_{n=1}^N (t_n - \Phi(x_n)^T w)^2 + \lambda \|w\|^2$$

pasamos a matriz

t : vector columna de todas las salidas t_n

$$t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

Φ : matriz de diseño con todas las transformaciones $\Phi(x_n)^T$

$$\Phi = \begin{bmatrix} \Phi(x_1)^T \\ \Phi(x_2)^T \\ \vdots \\ \Phi(x_N)^T \end{bmatrix}$$

w : vector de peso

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_Q \end{bmatrix}$$

$\hat{t} = \phi w$ valor predicho para todos los datos

$$\Phi w = \phi(x_n)^T w$$

$$\|a\|^2 = \sum_{i=1}^N a_i^2$$

entonces

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

$$\|w\|^2 = w^T w \propto \lambda \rightarrow \lambda \|w\|^2 = \lambda w^T w$$

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

$\Phi \in \mathbb{R}^{N \times q}$ matriz de diseño

$t \in \mathbb{R}^N$ vector objetivo

$\lambda > 0$ parámetro regularización controla el trade-off entre complejidad del modelo y ajuste

expandimos la función de costo

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

$$\text{primer} \quad = t^T t - 2t^T \phi w + w^T \phi^T \phi w \\ \text{termino}$$

$$\text{segundo} \quad = + \lambda w^T w \\ \text{termino}$$

función completa

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

Derivamos respecto a w

derivada de t^T respecto a w es 0
derivada de $-2t^T \phi w$ es $-2\phi^T t$
derivada de $w^T \phi^T \phi w$ es $2\phi^T \phi w$

derivada $\lambda w^T w$ es $2xw$

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

nos daria

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

despejamos

$$-2\Phi^T t + 2(\Phi^T \Phi + \lambda I)w = 0$$

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

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

$\Phi^T \Phi$ matriz de correlacion entre las variables de entrada

λI termino de regularizacion, controla la magnitud del coeficiente de w

$\lambda >$ mayor la penalizacion \rightarrow Se reduce las magnitudes de w

- maxima verosimilitud

funcion de verosimilitud

$$p(t|w, \sigma^2) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (t_n - \Phi(x_n)^T w)^2\right)$$

pasamos a log

$$\log p(t|w, \sigma^2) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \Phi(x_n)^T w)^2$$

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

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

① matriz de características $N \times Q$
t vector $N \times 1$
w vector $Q \times r$

Expandimos el error cuadrático medio

$$\|t - \Phi w\|^2 = (t - \Phi w)^T (t - \Phi w)$$
$$= t^T t - 2t^T \Phi w + w^T \Phi^T \Phi w$$

Derivamos respecto a w

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

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

$$\frac{\partial}{\partial w} (w^T \Phi^T \Phi w) = 2\Phi^T \Phi w$$

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

despejamos

$$-2\Phi^T t + 2\Phi^T \Phi w = 0$$

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

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

- máximo a-posteriori

$$P(t|w, \sigma^2) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (t_n - \Phi(x_n)^T w)^2\right)$$

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

$$P(w) = \frac{1}{(2\pi\alpha^{-1})^{Q/2}} \exp\left(-\frac{\alpha}{2} w^T w\right)$$

α = parámetro de precisión

Posterior Sin Constante

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

$$\log(p(w|t)) \propto \log(p(t|w)) + \log(p(w))$$

Verosimilitud

$$\log(p(t|w)) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t - \Phi w\|^2$$

prior

$$\log(p(w)) = -\frac{\alpha}{2} \log(2\pi\alpha^{-1}) - \frac{\alpha}{2} w^\top w$$

$$\log(p(w|t)) \propto -\frac{1}{2\sigma^2} \|t - \Phi w\|^2 - \frac{\alpha}{2} w^\top w$$

Ocurremos

$$\|t - \Phi w\|^2 = (t - \Phi w)^\top (t - \Phi w) = (w^\top \Phi^\top t - \Phi^\top t \cdot t + 2t^\top \Phi w + w^\top \Phi^\top \Phi w)$$

$$\frac{1}{2\sigma^2} (t^\top t) = 0$$

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

$$-\frac{1}{2\sigma^2} w^\top \Phi^\top \Phi w = -\frac{1}{2} \Phi^\top \Phi w$$

$$-\frac{\alpha}{2} w^\top w = -\alpha w$$

Despejamos

$$\frac{1}{\sigma^2} \Phi^\top t \rightarrow \frac{1}{\sigma^2} \Phi^\top \Phi w + \alpha w = 0$$

$$\left(\frac{1}{\sigma^2} \Phi^\top \Phi + \alpha I \right) w = \frac{1}{\sigma^2} \Phi^\top t$$

$$(\Phi^\top \Phi + \sigma^2 \alpha I) w = \Phi^\top t$$

$$x = \sigma^2 \alpha$$

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

ω se agrega un término de penalización cuadrática

$\lambda >$, mas se penalizan los valores grandes de ω

- Bayesiano con modelo lineal gaussiano

función de verosimilitud

$$p(t|w) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_n - \Phi(x_n)^\top w)^2}{2\sigma^2}\right)$$

$$p(t|w) \propto N(\Phi w, \sigma^2 I)$$

Φ : matriz $N \times Q$ con filas $\Phi(x_n)^\top$

t : vector $N \times 1$

$$p(w) = N(m_0, S_0)$$

$$m_0 = 0$$

$$S_0 = \sigma^{-2} I$$

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

$$p(w|t) \propto N(m_N, S_N)$$

$$S_N^{-1} = S_0^{-1} + \frac{1}{\sigma^2} \Phi^\top \Phi$$

$$m_N = S_N \left(S_0^{-1} m_0 + \frac{1}{\sigma^2} \Phi^\top t \right)$$

$$m_0 = 0$$

$$S_0 = \sigma^{-2} I$$

$$S_0^{-1} = \sigma^2 I$$

$$S_N^{-1} = \alpha I + \frac{1}{\sigma^2} \Phi \Gamma \Phi$$

$$\mathbf{m}_N = S_N \left(\frac{1}{\sigma^2} \Phi \Gamma t \right)$$

para un nuevo punto

$$p(E_{\text{new}} | t) = N(\Phi \Gamma x_{\text{new}})^T m_N, \sigma_x^2)$$

Varianza

$$\sigma_x^2 = \sigma^2 + \Phi(x_{\text{new}})^T S_N \Phi(x_{\text{new}})$$

- Regresión Ridge Kernel

Regresión lineal ordinaria

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \|t - \Phi \mathbf{w}\|^2$$

Φ matriz característica

t vector

$$\hat{\mathbf{w}} = (\Phi^\Gamma \Phi)^{-1} \Phi^\Gamma t$$

penalización sobre el tamaño de los pesos

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} [\|t - \Phi \mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2]$$

$$\hat{\mathbf{w}} = (\Phi^\Gamma \Phi + \lambda I)^{-1} \Phi^\Gamma t$$

$$\hat{\mathbf{w}} = \Phi^\Gamma a$$

$$\hat{\mathbf{w}} = \Phi^\Gamma a$$

$$\hat{t} = \Phi \hat{\mathbf{w}} = \Phi \Phi^\Gamma a$$

Resolvemos para a

$$\|t - \Phi \Phi^\Gamma a\|^2 + \lambda a^\Gamma \Phi \Phi^\Gamma a$$

$$(\Phi \Phi^\Gamma + \lambda I)a = t$$

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

introduciendo Kernel

$$K(x_i, x_j) = \Phi(x_i) \Phi(x_j)$$

$$K = \Phi \Phi^T$$

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

para nuevos puntos

$$\hat{f}_x = \sum_{n=1}^N \alpha_n K(x_n, x_n)$$

en vectorial

$$\hat{f}_x = K_x \alpha$$

- procesos Gaussiános

$$F(x) \sim \mathcal{G}(m(x), K(x, x'))$$

$m(x)$ = media

$K(x, x')$ = función de covarianza

matriz de covarianza: puntos de entrenamiento

$$K = \begin{bmatrix} K(x_1, x_1) & \dots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \dots & K(x_N, x_N) \end{bmatrix}$$

vector de covarianzas

$$K_x = \begin{bmatrix} K(x_n, x_1) \\ \vdots \\ K(x_n, x_N) \end{bmatrix}$$

Varianza nuevo punto

$$K_{xx} \subset K(x_n, x_n)$$

$$C = K + \sigma_n^2 I$$

media

$$\mu_x = K_x^T C^{-1} t$$

t = vector de observación

varianza

$$\sigma_x^2 = K_{xx} - K_x^T C^{-1} K_x$$

Predicción x_p

$$P(t_p | x_v, x, t) = N(\mu_{x_p}, \sigma_{x_p}^2)$$