

Diego Alejandro Arboleda Cvero

CC: 1087834896

## Teoría de Aprendizaje de Máquinas

### Parcial 1

#### Punto 1

Dado el modelo de Regresión.

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

donde  $\{t_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N$ ,  $w \in \mathbb{R}^Q$ ,  $\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$ ,

$Q \geq P$ , y  $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$ . Tenemos que

$t_n$ : es el valor observado

$x_n \in \mathbb{R}^P$  Vector de entrada

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

$w \in \mathbb{R}^Q$  Parámetros del modelo

$\eta_n \sim N(\eta_n | 0, \sigma_n^2)$  es el ruido, asumido como una variable aleatoria gaussiana con media 0 y varianza  $\sigma_n^2$ .

Mínimos Cuadrados.

Para este caso no tenemos un prior entonces

$\|f - \Phi(x)w\|_2^2$  y para nuestro caso =

$$J(w) = \|f - \Phi w^T\|_2^2$$

$$\|f - \Phi w^T\|_2^2 = \langle f_n - \Phi w^T, f_n - \Phi w^T \rangle$$

$$\begin{aligned} & (f_n - \Phi w^T)^T \cdot (f_n - \Phi w^T) = (f_n^T - (\Phi w^T)^T) \cdot (f_n - \Phi w^T) \\ & = f_n^T f_n - f_n^T (\Phi w^T) - (\Phi w^T)^T f_n + (w^T \Phi)^T (\Phi w^T) \end{aligned}$$

Dado esto buscaremos minimizar la función de costo, igualando a 0

$$\frac{\partial J(w)}{\partial w} = f_n^T f_n - 2 f^T \Phi w^T + (w^T \Phi)^T (\Phi w^T) = 0$$

$$= f_n^T f_n - 2 f^T \Phi w^T + w \Phi^T \Phi w^T$$

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

$$2 w \Phi^T \Phi = 2 f^T \Phi \text{ entonces tenemos que:}$$

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

Reflexión: Este método es adecuado si no existe colinealidad en las columnas de  $\Phi$  y el ruido es iid con distribución normal.

## 2) Minimos cuadrados regularizados

Para evitar el sobreajuste se agrega penalización a los coeficientes de  $w$ .

Función de costo a utilizar  $\|y - \phi w^T\|_2^2 + I\lambda \|w\|_2^2$

$y = t$  para este caso

$$\mathcal{L}(t - \phi w^T, t - \phi w^T) = (t - \phi w^T)^T \cdot (t - \phi w^T)$$

$$I\lambda \mathcal{L}(w, w) = I\lambda (w^T \cdot w)$$

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

$$t^T t - t^T \phi w^T - t^T \phi^T w + \phi^T w \phi w^T + I\lambda w^T w$$

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

Al igual que el caso anterior, minimizamos la función de costo derivando e igualando a 0

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

~~2w~~

$$-2w \phi^T \phi + 2I\lambda w = 2t^T \phi$$

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

$$w(\phi^T \phi + I\lambda) = t^T \phi \rightarrow \text{entonces}$$

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

Reflexión: Este método introduce un sesgo, pero reduce la varianza, lo que lo hace útil para datos con ruido.

## 8) Máxima Verosimilitud

El objetivo es maximizar la verosimilitud conjunta

$$\Theta = \operatorname{argmax}_{\Theta} L(\Theta; \mathbf{x})$$

Definimos la función de verosimilitud para este caso:

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

Dada la probabilidad conjunta y que presaremos a utilizar el Log-verosimilitud para mayor comodidad.

$$L(w) = \prod_{n=1}^N P(t_n | \phi(x_n)w^T, \sigma^2)$$

$$\log L(w) = \sum_{n=1}^N \log P(t_n | \phi(x_n)w^T, \sigma^2)$$

$$\log L(w) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \phi(x_n)w^T)^2$$

~~$$-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N \|t - \phi w\|_2^2$$~~

Vemos que término es constante y podemos omitirlo y nos quedo

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N \|t - \phi w\|_2^2$$

derivamos respecto a  $w$  e igualamos a 0

$$\frac{\partial L(w)}{\partial w} = 0$$

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

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

$$\frac{1}{\sigma^2} \sum_{n=1}^N (f_n - \phi(x_n)w^T)(\phi(x_n)) = 0$$

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

Reflexión: Maximizar  $\ln L(w)$  es equivalente a minimizar  $J(w)$  de mínimos cuadrados.

#### a) Máxima a Posteriori

MAP combina verosimilitud y un prior sobre  $w$

$$w_{map} = \arg \max_w P(w|t) \rightarrow P(t|w) P(w) = P(w|t) p(t)$$

Siendo el posterior likelihood

$$P(w|t) = \frac{P(t|w) p(w)}{P(t)} - \text{Prior} \rightarrow \text{dado esto}$$

subemos que:

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

entonces

Scribe

D M A

$$w_{\text{map}} = \arg \max_w \log \left( \prod_{n=1}^N N(f_n | \phi(x_n) w^\top, \sigma_n^2) \right) \prod_{q=1}^Q N - \dots \\ \approx \mathcal{L}(w, \phi, \sigma_w^2)$$

$$w_{\text{map}} = \arg \max_w = \frac{1}{2\sigma_n^2} \| t - \phi w^\top \|_2^2 - \frac{1}{2\sigma_w^2} \| w \|_2^2$$

Por ende tenemos que

$$\lambda = \frac{\sigma_n^2}{\sigma_w^2} \quad \text{an} \quad w = f^\top \phi ( \phi^\top \phi + \lambda I )^{-1}$$

## Comparación

### Método

OLS

Ridge

ML

MAP

### Ventajas

Simplicidad, sin  
regularización

Regularización  
mejora la  
estabilidad

Interpretación  
probabilística de  
OLS

Incluye prior para  
robustez

### Desventajas

Inestable en alta  
colinealidad

Introduce sesgo

Requiere asumir  
ruido

Prior mal elegido  
afecta los resultados