# Parcial H1

1 Sea el modelo de regresion $t_n = \phi(x_n) w^T + n_n$, con $\{t_n \in R, x_n \in R^p\}_{n=1}^N$, $w \in R^Q$, $\phi: R^p \to R^Q$, $Q \geq P$, y $n_n \sim N(n_n | 0, \sigma_n^2)$

- minimos cuadrados

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

$w \to$ parametros a estimar

$\phi(x_n) \to$ transformacion de las variables de entrada

$n_n \to$ ruido gaussiano media 0 varianza constante

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

vector de prediccion

$$\Phi w$$

$t = [t^1, t^2 \dots t^N]^T \in R^N$

vector de errores

$$t - \Phi w$$

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

Funcion de costo

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

$w \in R^Q$

Derivamos Funcion de costo

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

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

$\Phi^T t$: correlacion 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

- Minimos cuadrados regularizados

se agrega un termino mas que penaliza la magnitud de w para evitar sobreajuste

Funcion de costo con Regularizacion

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

pasamos a matriz

t= vector columna de todas las salidas $t_n$

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

$\Phi$= 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

$$w \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_\phi \end{bmatrix}$$

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

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

$$\|q\|^2 = \sum_{i=1}^{N} q_i^2$$

entonces

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

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

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

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

$t \in R^N$ vector objetivo

$\lambda \geq 0$ parametro regularizacion controla el trade-OFF entre complejidad del modelo y ajuste

expandimos la funcion de costo

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

primer termino $= t^T t - 2 t^T \phi w + w^T \phi^T \phi w$

segundo termino $= + \lambda w^T w$

Funcion completa

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

Derivamos respecto a $w$

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

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

nos daría

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

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 correlación entre las variables de entrada

$\lambda I$ término de regularización, controla la magnitud del coeficiente $w$

$\lambda$ ↑ mayor la penalización → se reduce las magnitudes de $w$

- máxima verosimilitud

función 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

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

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

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

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

expandimos el error cuadrático medio

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

Derivamos respecto a $w$

$$\frac{d}{dw}(t^T t) = 0$$

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

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

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

despejamos

$$-2 \Phi^T t + 2 \Phi^T \Phi w = 0$$
$$\Phi^T \Phi w = \Phi^T t$$
$$w = (\Phi^T \Phi)^{-1} \Phi^T t$$

- maximo a-posteriori

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

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

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

$\alpha$ = parametro de precision

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

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

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

Derivamos

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

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

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

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

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

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

Despejamos

$$\frac{1}{\sigma^2}\Phi^T t - \frac{1}{\sigma^2}\Phi^T \Phi w - \alpha w = 0$$

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

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

$\lambda = \sigma^2 \alpha$

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

w se agrega un termino de penalizacion cuadratica

$\lambda >$, mas se penaliza los valores grandes de w

- Bayesiano con modelo lineal gaussiano

Funcion de verosimilitud

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

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

$\Phi$ : matriz $N \times Q$ con Filas $\Phi(x_n)^T$

$t$ = vector $N \times 1$

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

$m_0 = 0$

$S_0 = \alpha^{-1} I$

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

$p(w|t) = N(m_N, S_N)$

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

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

$m_0 = 0$

$S_0 = \alpha^{-1} I$

$S_0^{-1} = \alpha I$

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

$$m_N = S_N \left( \frac{1}{\sigma^2} \Phi^T t \right)$$

para un nuevo punto

$$p(t_* | t) = N(\Phi(x_*)^T m_N, \sigma_*^2)$$

Varianza

$$\sigma_*^2 = \sigma^2 + \Phi(x_*)^T S_N \Phi(x_*)$$

- Regresion ridge kernel

Regresion lineal ordinaria

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

$\Phi$ matriz caracteristica

$t$ vector

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

Penalizacion sobre el tamaño de los pesos

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

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

$$\hat{w} = \Phi^T a$$

$$\hat{w} = \Phi^T a$$

$$\hat{t} = \Phi \hat{w} = \Phi \Phi^T a$$

Resolvemos para $a$

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

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

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

introducimos Kernel

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

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

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

para nuevos puntos

$$f_x = \sum_{n=1}^{N} a_n k(x_x, x_n)$$

en vectorial

$$f_x = k_x a$$

- procesos Gaussianos

$$F(x) \sim GP(m(x), k(x, x'))$$

$m(x)$ = media

$k(x, x')$ - Funcion de covarianza

matriz de covarianza puntos de entrenamiento

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

vector de covarianza

$$k_x = \begin{bmatrix} k(x_x, x_1) \\ \vdots \\ k(x_x, x_N) \end{bmatrix}$$

varianza nuevo punto

$$k_{xx} = k(x_x, x_x)$$

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

media

$$\mu_* = K_*^T C^{-1} t$$

t = vector de observacion

varianza

$$\sigma_*^2 = K_{**} - K_*^T C^{-1} K_*$$

Prediccion $x_*$

$$p(t_* | x_*, x, t) = N(\mu_*, \sigma_*^2)$$

- la regresion lineal por maximo verosimilitud predice valores puntuales sin regularizacion ni incertidumbre
la regresion ridge añade regularizacion L2
la regresion lineal bayesiana incorpora un prior sobre los pesos y permite obtener media y varianza en las predicciones
con Kernel, la regresion ridge Kernel mantiene la regularizacion pero sigue siendo determinista
la regresion bayesiana Kernel añade prior sobre funciones y predice media y varianza
procesos gaussianos modela toda la funcion como una distribuccion gaussiana sobre funciones, predice media y varianza y maneja incertidumbre natural